Anna Dai

dai-anna.github.io
anna.dai@alumni.duke.edu | +41 77 811 41 18

EDUCATION

DUKE UNIVERSITY
Master in Interdisciplinary Data Science
Grad. May 2023 | Durham, US
Cum. GPA: 4.0 / 4.0

- Duke Datathon 2021 Winner
- Duke Cloud Club Co-Founder & Ex-President

UNIVERSITY OF CALIFORNIA, LOS ANGELES (UCLA)
B.A. in Business Economics
Grad. Sept 2016 | Los Angeles, US

LINKS

Github:// **dai-anna**
LinkedIn:// **dai-anna**
Twitter:// **@annauppp**

COURSEWORK

- Machine Learning
- Deep Learning (Research Assistant)
- Cloud Computing (Teaching Assistant)
- Natural Language Processing
- Mathematical & Inferential Stats
- Causal Inference

SKILLS

PROGRAMMING

- Proficient in Python, R & SQL
- Basic Java & Rust

TOOLS & UTILITIES

- Pandas, Dask
- PyTorch, TensorFlow, Scikit-Learn
- OpenAI, Hugging Face
- AWS, GCP, GitHub Codespaces
- Git, Docker, FastAPI

LANGUAGES

- English (Native)
- Chinese (Fluent)
- French (Proficient)

CERTIFICATIONS

- AWS Certified Solutions Architect (Associate)
- Certified Public Accountant (CPA) in the State of California, USA

WORK EXPERIENCE

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE (EPFL)
Master's Valorisation Intern
September 2023 – Present | Lausanne, CH

- Engineered prototype leveraging LLMs and pre-trained word embeddings to intelligently extract and match skills from job postings, course descriptions, and resumes to an evolving taxonomy
- Custom built interactive, multistep annotation tool for non-technical annotator
- Supported interdisciplinary collaborators on Ruby implementation by providing NLP intuition and building FastAPI microservices
- Adapted pipeline as the foundation of multiple published research papers on skill extraction, upskilling recommendations, and synthetic data generation

DUKE UNIVERSITY
Research Assistant
January 2023 – August 2023 | Durham, US

- Reviewed literature and implemented state-of-the-art defenses against adversarial attacks as benchmarks against proposed methods for two papers in the federated learning and model security spaces
- Carefully planned and conducted over 400 experiments over limited period of time to meet conference deadlines

Teaching Assistant
August 2022 - May 2023 | Durham, US

- Held office hours and graded assignments for Data Engineering Systems, Data Analysis at Scale in the Cloud, and Fraud Analytics courses

Race and the Professions Fellow
September 2021 - May 2022 | Durham, US

- Represented the Asian (Canadian) perspective to discuss racial justice issues in professions within and beyond Duke University departments
- Organized data visualization contest to spotlight racial inequality and racism within various Duke University departments

STITCH FIX | Data Science Intern
June 2022 – August 2022 | San Francisco, US

- Explored unknown problem space of incorporating external fashion trends into currently historical-data-driven algorithms
- Developed steel-thread attempt at mapping trending terms to inventory items by customizing internal search pipeline and leveraging pre-trained word embeddings and latent item embeddings to enable trend-level modeling
- Discovered new data source from existing vendor and automated data cleaning and term extraction pipeline from PDF reports to build a database of high-quality trending terms by date and category (i.e. Womens, Mens, Children)

ERNST & YOUNG LLP | Senior Tax Consultant, Tax Staff
January 2017 - April 2021 | San Francisco, US

- Managed compliance and consulting teams of 5-10 staffs and juggled up to 14 engagements at any given time for clients who managed up to $33 billion AUM
- Conducted research tax credit studies, processed raw personnel data to quantify credit values, interviewed clients' technical personnel, researched state-of-the-art software topics, and drafted technical memoranda
- Wrote data validation logic to perform trading security analysis, prepare tax allocations, and reduce human error

PUBLICATIONS

JOBSKAPE: A FRAMEWORK FOR GENERATING SYNTHETIC JOB POSTINGS TO ENHANCE SKILL MATCHING

Authors: **Anna Dai***, Antoine Magron*, Mike Zhang, Syrielle Montariol, Antoine Bosselut [paper]
The 1st Workshop on Natural Language Processing for Human Resources (NLP4HR), EACL 2024 [code]

- Developed a framework to generate synthetic job postings using LLMs to address the lack of annotated job posting data that support skill extraction and matching tasks
- Released a dataset of around 5,000 synthetic job postings to the public

COURSE RECOMMENDER SYSTEMS NEED TO CONSIDER THE JOB MARKET

Authors: Jibril Frej, **Anna Dai**, Syrielle Montariol, Antoine Bosselut, Tanja Käser
The 47th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2024) Perspectives Paper [preprint]

- Leveraged LLMs in NLP pipeline to extract skill and match them to a known taxonomy from job postings, course descriptions, and resumes
- Compared greedy heuristics-based and reinforcement learning-based recommendation systems to recommend courses to job seekers based on their resume and job postings

MODELGUARD: INFORMATION-THEORETIC DEFENSE AGAINST MODEL EXTRACTION ATTACKS

Authors: Minxue Tang, **Anna Dai**, Louis DiValentin, Aolin Ding, Amin Hass, Neil Zhenqiang Gong, Yiran Chen
33rd USENIX Security Symposium (USENIX Security 2024) [paper]

- Proposed novel defense against adaptive model extraction attacks through prediction perturbation by leveraging information theory
- Reviewed and implemented state-of-the-art model extraction defenses and attacks including Adaptive Misinformation, which requires an additional outlier exposure model, as baselines
- Ran experiments on four datasets (MNIST, CIFAR10, CIFAR100, and ImageNet) to benchmark performance of MODELGUARD against other defenses and attacks

PLUGVFL: ROBUST AND IP-PROTECTING VERTICAL FEDERATED LEARNING AGAINST UNEXPECTED QUITTING OF PARTIES

Authors: Jingwei Sun, Zhixu Du, **Anna Dai**, Saleh Baghersalimi, Alireza Amirshahi, David Atienza, Yiran Chen
The Twelfth International Conference on Learning Representations (ICLR 2024) Submission [preprint]

- Assisted with research on robustness of vertical federated learning convolutional models against performance and IP leakage risks when nactive parties unexpectedly quit during deployment
- Completed literature review on possible defenses against various adversarial attacks, implemented the defenses to benchmark results, and reviewed code for different adversarial attacks

PROJECT EXPERIENCES

DRUG DIVERSION DETECTION AND INTERVENTION

Master's Capstone Project
Duke MIDS Keynote Presentation 2023

- Worked with Duke Anesthesiology to tackle the issue of drug diversion by anesthesiologists by defining the problem as a data problem that can be handled with machine learning
- Developed a complete modeling approach to detect potential diversion behavior from surgical data
- Specifically tested unsupervised deep learning approaches including auto-encoder models to detect anomalies

AWS CLOUD-NATIVE AUTOMATIC TWEET GENERATOR

Team Data Engineering Project [code]

- Built using Python an Amazon Web Service cloud-native end-to-end data-scraping and processing pipeline that automatically generates Tweets for currently trending hashtags with machine learning model
- Orchestrated pipeline to run daily, push to web application through Google Cloud Platform as well as post from Twitter bot through Twitter API
- Leveraged Python CDK (IaC), AWS Lambda, EventBridge, Batch, EC2 Spot Instance, ECR and S3 services as well as Docker containers
- Set up continuous integration and delivery pipeline through GitHub Actions and logging and monitoring pipeline in AWS