

Du Xiang

xd00099@berkeley.edu | www.linkedin.com/in/du-xiang | du-xiang.com

EDUCATION

University of California, Berkeley 08/2022 – 05/2023

Master of Engineering in Electrical Engineering and Computer Science

Courses: ML Design on Multimedia, UI Development, Deep Reinforcement Learning, Parallel Computing, Computer Vision

University of California, San Diego 09/2018 – 06/2022

Bachelor of Science in Data Science, GPA: 3.99/4.0, Summa Cum Laude

Courses: Data Science, ML Systems, Data Structure & Algorithm, ML/DL/NLP, Web, Relational Database, Optimization

SKILLS

Programming Languages: Java, Python, C++, SQL, R, JavaScript, Bash, MATLAB

Certifications: IBM DevOps and Software Engineering Professional

Web: MERN Stack, HTML/CSS, JavaScript; **Cloud & DevOps:** Kubernetes, Docker, Google, IBM, AWS. **Machine Learning:** Pandas, PyTorch, Scikit-learn; **Big Data Analytics:** Spark, Hadoop, Dask; Agile methodologies with Git.

EXPERIENCE

Research Assistant | UC Berkeley & Zendar 09/2022 – continue

- Researched **autonomous driving** and methods to fuse information from radar and camera
- Exploring **object detection** algorithms including Faster R-CNN, Retina-Net, and YOLOv3.
- Exploring and implementing deep sensor fusion (early, cross, late) architectures to integrate radar camera data.

Machine Learning and NLP Intern ([publication](#)) | San Diego Supercomputer Center 02/2021 – 08/2022

- Collaborated in the machine learning team on developing the Synthetic Biology Knowledge System (NSF funded)
- Implemented various ways to train **BioBERT** using **PyTorch**, achieving **92+**% F1 for extracting **9000+** relations
- Leveraged **LDA topic models** trained on 1000+ academic articles, achieved high topic scores rated by experts
- Constructed a **graph database** using **Neo4j** that united 2 unrelated data clusters
- Containerized tasks with **Docker** to speed up deployment on **Kubernetes** by 40%

Instructional Assistant | UC San Diego 03/2020 – 06/2022

- **Distinguished** IA for teaching *Principles of Data Science, Theories of ML, Practice of Data Science & ML*
- Primary tech support for code debugging, environment setup, and assignment release engineering

Data Pipeline, Python Developer Intern | CNS Laboratory of Memory and Brain 01/2020 – 06/2021

- Served as an agile python developer who works on **3D brain images pipelines, analysis, and task automation**
- Applied machine learning, including random forest and regression, to fMRI images to analyze brain features
- Engineered a **data pipeline** to create automated data visualizations from MRI data for research posters
- Automated tasks on a Linux server using **Bash** and **Python** to improve image segmentation efficiency by 10+x

PROJECTS | **GitHub:** <https://github.com/xd00099>

Customer Accounts Microservice - DevOps | Independent

- DevOps capstone project that implements microservices for customer accounts, deployed on cloud
- Followed **Agile** and **Scrum** methodologies, documented and completed tasks in 3 sprints using Kanban board
- Implemented Read, Update, Delete, List RESTful Flask services using **Test Driven Development** approach
- Deployed the microservice on Kubernetes using an **automated CI/CD** pipeline with **GitHub Actions** and **Tekton**

Social Wall - Full Stack Web | Independent

- Designed and implemented the Instagram-like frontend view of home page and main page using **React.js**
- Implemented modern **authentication** system that supports native login with **MongoDB** and Google **OAuth2.0**
- Used **Redux.js** to manage app states and applied middleware to handle user permissions
- Implemented RESTful APIs with Express.js and Node.js to handle HTTP requests and responses

Faculty Allocation Recommendation System - AI & NLP | Backend Leader

- Built web crawlers to retrieve publication data and google scholar labels for natural language processing
- Trained LDA models that automatically and successfully labels and categories each faculty based on publications.
- Led backend team to build pipelines: **Data ETL, Preprocess, Modeling, and Deployment using Heroku**

Scalable Data Analytics on Amazon Product | Independent

- Data processing, feature extraction, and ML modeling on distributed system w/ Amazon Product data (~25GB).
- Overcame challenges managing large amount of data using **Apache Spark** on **Kubernetes** with a distributed system
- Worked on word2vec embedding, PCA, and Tree-based models w/ **PySpark DataFrame APIs** and **MLlib**.
- Performed optimization on models, query, and data partitions to improve the runtime efficiency by 20%