MIA SHI

minmiashi@gmail.com ◇ Personal Portfolio ◇ https://www.linkedin.com/in/min-mia-shi/

WORK EXPERIENCES

Data Scientist Jan. 2025 - Present
Reframe Data Services *North Bethesda, MD / Remote*
Full-stack data professional building ML models, analytics dashboards, ETL pipelines, and search-based data modeling solutions to transform raw data into feature-rich outputs.

- Developed and deployed cloud-hosted interactive dashboards, embedding them into front-end applications for real-time data visualization. *Tools: Python, Streamlit, Dash, Heroku, AWS, React, JavaScript*
- Led development of ML/DL models (SVM, CNN, RNN, LSTM) for speaker identification, achieving 80-90% accuracy. *Tools: sklearn, PyTorch, TensorFlow, ML, DL, Transfer Learning*
- Built and optimized real-time and batch data pipelines, integrating OpenSearch and Elasticsearch for scalable data modeling and retrieval. *Tools: OpenSearch, Elasticsearch, AWS Glue, Lambda, Spark, Python, SQL*
- Automated and deployed scalable AI-driven data pipelines for content generation, optimizing efficiency and reliability. *Tools: AWS Glue, Lambda, Streamlit, LLM, Python*
- Worked with engineers and researchers to enhance AI-driven analytics, data infrastructure, and machine learning pipelines.

Data Engineer Jun. 2024 - Oct. 2024
The Sunwater Institute *North Bethesda, MD / Remote*
Designed and optimized scalable data pipelines, big data workflows, leveraging Spark, Hive, AWS for high-performance data processing.

- Built and optimized scalable ETL pipelines for large-scale data ingestion and processing, enabling real-time and batch workflows. *Tools: AWS S3, Glue, Lambda, Spark, Python (Boto3)*
- Developed high-performance data workflows using Apache Spark and Hive on AWS EMR for efficient querying and transformation of big data. *Tools: Spark, Hive, AWS EMR, SQL, Pandas*
- Automated speech-to-text and speaker identification pipelines, improving processing efficiency and achieving 90%+ accuracy. *Tools: PyTorch, CNN, Transfer Learning, AWS Transcribe, Textract*
- Engineered and maintained scalable data infrastructure, integrating CI/CD and monitoring for reliability. *Tools: Airflow, Streamlit, Python, GitHub, AWS EC2, Jenkins*

Data Analyst May 2020 - May 2024
The University of Texas at Dallas *Richardson, TX*
Took responsibility for data analysis for 10+ global health/policy projects using advanced statistical models.

- Managed data collection in diverse methods such as Qualtrics surveys and web scraping using R and Python.
- Developed 20+ robust statistical models (multi-variable and fixed-effect regression, difference-in-difference, time-series), utilized ML models and NLP skills to support correlation and causal inference in research.
- Led a team of five research assistants, ensuring timely project completion and publication.

PROJECTS

Twitter Clone: High-throughput Social Media Backend — *Python, Django* May 2024 - Present
Developing the backend for a social media platform using Django (Python) with HBase, MySQL, and Redis.

- Maximized query efficiency by storing objects with HBase, MySQL & Amazon S3 based on query complexity.
- Addressed N+1 slow query issues by implementing Redis caching and denormalization.
- Integrated Celery and RabbitMQ to establish asynchronous workers with varying priority levels.
- Implemented a push model for distributing news feeds to followers efficiently.
- Optimized memory and resource allocation using recursive small batches of asynchronous tasks.

AI-Powered Payment Service Chatbot for Enhanced Customer Engagement Aug. 2023 - Dec. 2023
Developed and deployed an AI-driven chatbot using Python and MySQL, leveraging advanced NLP techniques to enhance customer engagement for Onyx CenterSource.

- Leveraged NLP and MySQL for analyzing and querying an extensive database containing over 10 million entries.
- Improved response efficiency by 25% and achieved 99% accuracy using the XGBoost model.
- Enhanced user engagement, boosting customer satisfaction and strengthening the company's brand image.

Kaggle Plant Pathology Competition: Leveraging Deep Learning CNNs Nov. 2023 - Dec. 2023
Implemented deep learning models using Python and PyTorch to enhance disease identification accuracy in crops.

- Applied transfer learning on CNNs with 13,042 images across 12 categories, significantly improving accuracy.
- Performed image augmentation techniques (rotation, flipping, zooming, noise injection) to enhance data.
- Fine-tuned ConvNext DL models, achieving 86.8% accuracy and securing a Top 3 ranking in the competition.

Optimizing Big Data Risk Analysis for a Company with Hadoop and Tableau Aug. 2022 - Dec. 2022
Engineered data visualization dashboards using Tableau, linked to Hadoop, for business risk analysis.

- Processed and analyzed geospatial data with Hadoop, Hive, and Spark, reducing significant processing time.
- Developed Tableau visualizations linked to Hadoop and built interactive dashboards for business analysis.
- Conducted linear regression and multivariate analysis, contributing to predictive accuracy by 15%.

EDUCATION

Ph.D. in Political Science (Quantitative Methods & Data Science) Dec. 2024 (***GPA: 3.95/4***)
The University of Texas at Dallas
M.S. in Business Analytics (Data Science & Engineering Track) May 2024 (***GPA: 3.95/4***)
The University of Texas at Dallas

SKILLS

Programming & Tools: Python, SQL, Java, Javascript, AWS Web Services, Streamlit, Dash, R
AI & ML Modeling: Statistical Modeling, Deep Learning, Machine Learning, NLP, LLM, PyTorch, Tensorflow
Database & Big Data: SQL Server, MySQL, PostgreSQL, AWS RDS, Hadoop, Spark, Hive, Impala, Sqoop, Pig
Certificates: Certificate in Applied Machine Learning, AWS Certified Cloud Practitioner, Google Analytics