# Zijun Yi

+1 515-357-3516 | jimzijun@gmail.com | www.linkedin.com/in/zijunyi/

## WORK EXPERIENCE

### Data Engineer
*Bases, NielsenIQ*
Remote, US
03/2022 – Present

- **Led development and deployment of simulation models**, improving performance by **40%** and reducing processing time by **30%** through code and workflow optimizations. **Overcame challenges with legacy systems and complex data integration**, enabling a **30% increase in study size** and shortening calibration workload by **25%**, which accelerated project completion. **Collaborated with cross-functional teams** to ensure seamless implementation.
- **Enhanced system reliability and efficiency** by profiling models and optimizing job scheduling, **addressing frequent job failures and system bottlenecks**. This resulted in an **80% reduction in tickets** for lost processes and jobs after migration. **Communicated effectively with stakeholders** to identify issues, decreasing operational costs and boosting user satisfaction across teams.
- **Migrated infrastructure from Azure Batch to Kubernetes on AKS**, eliminating cold start overhead and improving resource utilization, which increased operational efficiency. **Faced and resolved challenges related to resource contention and deployment complexities**, ensuring a smooth transition with minimal downtime. **Coordinated closely with the DevOps team** to streamline processes.
- **Improved system responsiveness and throughput** for production model training and inference, enabling seamless scaling and better performance. **Collaborated with data scientists and engineers** to optimize algorithms, leading to faster deployment of features and improved customer satisfaction. **Facilitated regular communication between teams** to align goals and deliverables.

### Research Assistant
*Science of Science and Computational Discovery Lab*
Syracuse, NY
02/2021 – 12/2021

- Implemented **sentiment analysis** using a pre-trained model to identify gender biases within word embeddings, enhancing model fairness and accuracy.
- Engineered and deployed scalable big data pipelines using **Hadoop**, **Apache Spark** (PySpark), and Apache Airflow, processing vast volumes of data weekly, a search engine and recommendation system for publications powered by **Elasticsearch**.
- Developed a Python-based data collection and analysis tool to fetch figshare.com data via REST APIs, employing **regression**, **random forest**, **gradient boosting**, and **neural network** models to analyze dataset utilization in academic publications.
- Researched and developed a system to identify race and gender biases in AI-driven speech recognition systems, incorporating technologies such as Amazon Alexa, AWS Lex API, and AWS Transcribe.

## EDUCATION

**Master of Science in Applied Data Science**
Syracuse University — Sept. 2020 - Dec. 2021

**Bachelor of Science in Information Management & Technology**
Syracuse University — Sept. 2016 - May. 2020

## CORE SKILLS

**Programming Languages**: Python, Java, R, SQL
**Frameworks and Libraries**: PyTorch, TensorFlow, Scikit-learn, Pandas, NumPy
**Big Data Technologies**: Apache Spark, Hadoop, Kafka, MapReduce
**Web Frameworks**: Django, Flask
**Cloud Platforms**:
**Tools**: Git, Docker, Kubernetes, Airflow, Elasticsearch
**Other Skills**: Machine Learning, Deep Learning, NLP, Computer Vision, Transfer Learning, Data Engineering

## SCIENTIFIC PUBLICATIONS

- Acuna, D.E., **Yi, Z.**, Liang, L., Zhuang, H., Predicting the usage of scientific datasets based on the article, author, institution, and journal bibliometrics (2022 iConference) - [Mar. 2022]