Hudanyun Sheng

hudanyun.sheng@outlook.com | (+86) 13683132915 | https://danniesheng.github.io/

Data Scientist with cross-disciplinary experience across healthcare, AI/ML systems, and computer vision. Skilled in building end-to-end machine learning solutions—from data ingestion to deployment—with proven ability to work independently and collaborate across functions. Experienced in both U.S. and China environments, with strong domain adaptation skills and research grounding.

EDUCATION

University of Florida - *M.S. in Electrical and Computer Engineering (GPA: 3.86/4), December 2019*

Master thesis: Switchgrass Genotype Classification using Hyperspectral Imagery

University of Florida - *M.S. in Industrial and Systems Engineering (GPA: 3.87/4), December 2017*

Tongji University - *B.S. in Physics (GPA: 4.45/5), June 2015*

Bachelor thesis: The Correction of the Intensity Unevenness of X-Ray KB Imaging

TECHNICAL SKILLS

- Programming & Development: Python (NumPy, Pandas, SciPy, Streamlit, Plotly), Flask, HTML/CSS/JS, MATLAB
- Machine Learning & Deep Learning: PyTorch, TensorFlow, Scikit-learn, Keras, Hugging Face, OpenCV
- NLP & Generative AI: ChatGPT API, LangChain, RAG, NLP pipeline design, Weights & Biases
- Data Engineering & Platforms: SQL, PySpark, BigQuery, Kedro, Databricks, EHR data processing
- Cloud & DevOps: GCP, AWS, Azure, Docker, Kubernetes, Git
- Tools: JIRA, Notion, Confluence, GitHub Actions, Markdown, MS Office

PROFESSIONAL EXPERIENCE

Johnson & Johnson | **Data Scientist (Contractor)** Center of Excellence   Beijing, China | April 2024-present

Early Adopter Prediction Model for New Drug Launch

- Built ML models for HCP targeting based on historical sales and promotion data
- Designed modular data pipelines with Kedro to ensure input integrity and scalability

RAG-based Off-label Detection for Japanese FAQs

- Developed a prototype retrieval-augmented generation (RAG) workflow using ChatGPT API
- Automated detection of inconsistencies between localized Japanese FAQs and regulatory labels

Feature Store Migration and Optimization

- Participated in platform migration to Databricks; supported redesign of feature storage architecture

Doctor Lookup & Q&A System (Japanese Medical RAG)

- Designed and deployed a Japanese medical RAG system integrating FAISS + E5 embeddings, history-aware retrieval chains, cross-encoder reranking (bge-reranker-small), and uncertainty gating
- Built robust JP/CN name normalization and fuzzy-matching pipeline (NFKC normalization, kana-to-romaji conversion, traditional/simplified Chinese unification) to improve match accuracy across writing systems
- Developed interactive UI for parameter tuning (top-k, min_score, token limits, reranker toggle) and evidence tracing, with multi-turn dialogue logging and export
- Modularized LLM/retriever/chain components with Dockerized deployment, local HF model caching, and automatic layer-level rebuilding triggered by parameter changes

Zenni Optical | **Data Scientist** AI/ML Department   Beijing, China | June 2023-Jan 2024

Prescription (Rx) Recognition API

- Built FastAPI service for extracting structured Rx fields (sphere, cylinder, axis) from images

- Adapted parsing logic for Japanese prescription formats, significantly improving accuracy
- Integrated CloudSQL and BigQuery for multi-region support and usage tracking

Blur Detection and Quality Monitoring

- Analyzed API call logs and user behavior to evaluate and improve document image quality

Interactive Evaluation Interface

- Built Streamlit-based internal tool to visualize OCR results and assist human-in-the-loop validation

University of Texas Southwestern Medical Center | **Data Scientist** Quantitative Biomedical Research Center

Dallas TX USA | Sep 2021-May 2023

CyTOF Image Analysis Toolkit

- Designed and developed a Flask-based image analysis package integrating spatial and single-cell expression data
- Achieved 10× speed-up through parallelized processing and Dockerized deployment for researchers

Cell Segmentation in H&E Pathology Slides

- Implemented Mask R-CNN for nuclei segmentation and 6-class classification (82.5% detection, 82.0% accuracy)
- Customized loss functions to handle incomplete labels, recovering ~20% of training data

EHR De-identification and NLP Input Pipeline

- Built workflows for anonymizing and preprocessing electronic health records to support cancer scoring models

Donald Danforth Plant Science Center | **Data Science Researcher** Data Science Facility

St. Louis MO USA | Feb 2020-Sep 2021

Multimodal Plant Image Analysis Pipeline

- Built automated pipelines for RGB, thermal, and hyperspectral data processing and visualization

Instance Segmentation and Growth Tracking

- Developed instance-level segmentation and tracking pipeline to analyze leaf growth across timepoints

Open-source Development (PlantCV)

- Contributed new image analysis modules, wrote unit tests, documentation, and conducted community training
- Supported interdisciplinary teams with visualization, data wrangling, and reproducible analysis

University of Florida Academic Health Center | **Data Science Intern**

Precision and Intelligent Systems in Medicine Partnership Lab — Gainesville FL USA | May 2019-Aug 2019

- Processed and cleaned patient vital sign time-series data for early risk modeling
- Extracted 24-hour pre-hospital features and evaluated multiple clustering techniques
- Reproduced interpolation-based models to handle data irregularity in patient cohorts

PERSONAL PROJECTS

AI News Agent – LLM-based News Summarization and QA Assistant

Open-source project: daily AI news aggregation, summarization, and question answering

- Built a lightweight Retrieval-Augmented Generation (RAG) pipeline combining web-scraped news, local embedding models, and OpenAI LLMs
- Implemented scheduled scraping, auto-tagging, and news summarization with category filtering
- Supported user question answering over news corpus using hybrid search + local knowledge base
- Designed interactive front-end via Streamlit, deployed on Hugging Face Space; explored local embedding model switching and Docker-based deployment

PUBLICATION

- Sheng, H., Wang S., et al. “MTIA: An open-source python package for systematic multiplexed tissue image analysis” (in preparation)
- Sheng, H., Gutierrez, J., Schuhl, H., Murphy, K. M., Acosta-Gamboa, L., Gehan, M., & Fahlgren, N. (2023). Increasing the Throughput of Annotation Tasks Across Scales of Plant Phenotyping Experiments. Authorea Preprints.
- Rong, R., Sheng, H., Jin, K.W., Wu, F., Luo, D., Wen, Z., Tang, C., Yang, D.M., Jia, L., Amgad, M. and Cooper, L.A., 2023. A deep learning approach for histology-based nucleus segmentation and tumor microenvironment characterization. Modern Pathology, 36(8), p.100196.
- Panda, K., Mohanasundaram, B., Gutierrez, J., McLain, L., Castillo, S. E., Sheng, H., ... & Slotkin, R. K. (2023). The plant response to high CO2 levels is heritable and orchestrated by DNA methylation. New Phytologist, 238(6), 2427-2439.
- Yu, G., Zare, A., Sheng, H., Matamala, R., Reyes-Cabrera, J., Fritschi, F.B. and Juenger, T.E., 2020. Root identification in minirhizotron imagery with multiple instance learning. Machine Vision and Applications, 31, pp.1-13.

ACADEMIC RESEARCH EXPERIENCE

Machine Learning and Sensing Lab | **Graduate Research Assistant** Gainesville FL USA | Mar 2017-Dec 2019

- Developed machine learning models for root detection in minirhizotron imagery using multi-instance learning.
- Proposed and implemented a Siamese-network-based dimensionality reduction method to classify plant genotypes.
- Designed and maintained hyperspectral/thermal data processing pipelines to support remote sensing research.

Institute of Precision Optical Engineering | **Undergraduate Research Assistant** Shanghai, China | June 2014-June 2015

- Conducted simulations of X-Ray KB (Kirkpatrick-Baez Microscope) imaging through programming, addressing and rectifying irregularities in the imaging process
- Developed expertise in correcting the unevenness of X-Ray KB imaging for improved accuracy and precision.
- Designed and implemented a user-friendly Graphical User Interface using MATLAB, facilitating a seamless and intuitive experience for navigating the simulated imaging process.

PROFESSIONAL STRENGTHS

- **Analytical & Fast Learner:** Quickly grasp new technologies and apply them independently; demonstrated by rapid onboarding and execution across diverse ML and NLP projects.
- **Effective Communicator:** Capable of translating technical concepts for both technical and non-technical audiences; frequently deliver stakeholder-facing demos and documentation.
- **Cross-cultural Collaboration:** Bilingual in English and Chinese with experience working across US and China teams, enabling smooth coordination in international and cross-functional environments.

CERTIFICATES, HONORS, REWARDS AND MISCELLANEOUS

- 1st Place of the “Swarm Behavior on the Grid” track in the Siemens "Tech for Sustainability Campaign 2023” (2023)
- Google Data Analytics Certificate– a rigorous, hands-on program that covers the entire scope of the data analysis process
- Co-Chair of the Committee for Scientific Training and Mentoring at Donald Danforth Plant Science Center