盛胡丹筠

邮箱:hudanyun.sheng@outlook.com | 电话 13683132915 | 个人网站 https://danniesheng.github.io/

教育背景

佛罗里达大学 美国，佛罗里达
电子与计算机工程硕士 2018.01 - 2019.12
硕士论文:基于高光谱图像的柳枝草基因型分类 (Switchgrass Genotype Classification using Hyperspectral Imagery)

佛罗里达大学 美国，佛罗里达
工业系统工程硕士 2016.01 - 2017.12

同济大学 中国，上海
物理学学士 2011.09 - 2015.06
毕业论文: X射线多层膜KB成像的强度均匀性校正

专业技能

- 编程与开发:Python（NumPy, Pandas, SciPy, Streamlit, Plotly），Flask，HTML/CSS，JavaScript，MATLAB
- 自然语言处理与生成式AI:LangChain, RAG, ChatGPT API, NLP, Weights & Biases
- 机器学习与深度学习:PyTorch, TensorFlow, Scikit-learn, Keras, Hugging Face, OpenCV
- 数据工程与分析平台:SQL, PySpark, BigQuery, Kedro, 数据管道开发, EHR数据分析
- 云平台与DevOps:GCP, AWS, Azure, Docker, Kubernetes, Git
- 工具与协作:JIRA, Notion, Confluence, Markdown, MS Office, GitHub Actions

工作经历

Johnson & Johnson | 数据科学家（外包） 中国北京，2024.04 至今

新药早期接受者预测模型构建
- 构建HCP识别模型以辅助新药上市前的市场策略制定
- 使用Kedro构建可复用数据管道，集成销售与目标名单数据

日文*FAQ*一致性审查系统原型开发
- 使用ChatGPT API + RAG方法，识别本地化FAQ与标签内容之间的不一致
- 为日本市场的合规审核流程提供自动化支持

特征存储系统迁移与优化
- 协助项目从Kedro迁移至Databricks平台，参与特征结构的重构与标准化

医生资料检索与问答系统（*RAG*）
- 设计并实现日文医疗场景 RAG 系统:FAISS + E5 向量检索，历史感知检索链，交叉编码器重排（bge-reranker-small），不确定性门控。
- 研发JP/CN 名称归一化与模糊匹配管线（NFKC、假名转罗马音、简繁转换），提升跨写法检索命中率。
- 实现可视化调参 UI（top-k、min_score、token 预算、reranker 开关）与 evidence 展示，支持多轮对话归档与导出。
- 工程化:模块化 llm / retriever / chain，Docker 化部署，HF 模型本地缓存，参数变化触发按层重建。

Zenni Optical | 数据科学家 中国北京，2023.06 – 2024.01

*Rx*处方识别系统开发
- 使用FastAPI构建Rx眼镜处方结构化提取服务，部署至GCP平台
- 针对日本处方格式进行适配优化，显著提高识别准确率
- 使用BigQuery分析用户行为并推动迭代优化

图像模糊检测模块
- 分析API日志与图像模糊影响，提升图像预处理稳定性

交互式评估工具开发
- 使用Streamlit开发内部可视化界面，提升模型评估效率与人机协作体验

德克萨斯西南医学中心 定量生物医学研究中心 | 数据科学家 美国，德克萨斯州，达拉斯, 2021.09 – 2023.05

*CyTOF*图像分析工具开发

- 构建结合空间位置与单细胞表达信息的图像分析包，10倍提速
- 提供基于Flask + Docker 的可部署界面供生物研究人员使用

*H&E*组织切片分析

- 构建Mask R-CNN模型实现细胞实例分割（检测率82.5%，分类准确率82.0%）
- 针对不完整标签定制损失函数，挽救约20%训练样本

*EHR*文本脱敏与*NLP*输入管道

- 构建电子病历去标识化流程，用于癌症风险模型训练数据的准备

唐纳德丹福思植物科学中心 (Donald Danforth Plant Science Center) | 数据科学研究员

美国，密苏里州，圣路易斯 2020.02 – 2021.09

多模态植物图像分析管道开发

- 主导RGB/热成像/高光谱图像的自动化处理流程，提升表型数据处理效率

植物叶片实例分割与生长追踪

- 构建基于Mask R-CNN的分割模型与叶片追踪算法，用于植物生长分析

*PlantCV*开源项目开发与推广

- 新增分割与特征提取模块，参与文档撰写与教学推广
- 支持跨团队图像分析任务，提升数据可用性与科研可视化质量

佛罗里达大学医疗中心 精密医学智能合作部 | 数据科学实习生 美国，佛罗里达州，盖恩斯维尔 2019.05-2019.08

- 参与基于临床生命体征数据的时间序列分析项目，完成数据清洗、异常检测与队列构建，保障分析基础数据质量
- 提取病人住院前24小时内的关键时序特征，应用并复现插值网络等方法处理数据不规律性
- 比较不同聚类算法在病人分群上的效果，支持后续临床风险建模研究

个人项目

AI 新闻聚合与问答助手系统

构建轻量级 *RAG* 管道，实现 *AI* 新闻的自动抓取、摘要生成与问答支持

- 实现定时抓取与分类机制，集成本地嵌入模型与 OpenAI LLM 生成每日新闻摘要
- 支持用户在本地语料库基础上进行新闻问答，结合混合检索与语义匹配
- 使用 Streamlit 构建交互式界面，部署于 Hugging Face Space
- 实现 Docker 化部署流程，支持嵌入模型热切换与模块化扩展

学术研究经历

佛罗里达大学 机器学习及感知实验室 (Machine Learning and Sensing Lab) 美国，佛罗里达州，盖恩斯维尔

机器学习研究助理 2017.05-2019.12

- 基于多实例学习设计图像分割算法，实现微根管图像中植物根系的自动检测，提升数据标注效率
- 搭建高光谱与热成像图像处理流程，包括预处理、解混合与目标检测，应用于植物表型大数据自动化分析
- 研发基于孪生网络的降维算法，增强植物基因型分类的可分性，模型可迁移用于高维医学影像分类
- 利用基于纯正端元提取的高光谱解混合的算法，实现高光谱图像的目标检测和图像分割

出版文章

- 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.

综合能力

- 中英文双语沟通能力:具备跨文化协作经验，能够用英文撰写技术文档并进行技术/非技术团队之间的沟通
- 独立执行与快速学习:多次独立完成从调研、建模到部署的AI项目，具备快速掌握新技术并落地应用的能力
- 跨职能协作经验:在与产品、商业、合规团队合作中，能准确理解业务需求并用数据与模型推动解决方案落地

奖项与证书

- 西门子2023可持续技术大赛（国际黑客松比赛）“电网数据的人群使用行为”组第一名（团队）
- 谷歌数据分析证书