

# 盛胡丹筠

邮箱: [hudanyun.sheng@outlook.com](mailto: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

电子与计算机工程硕士

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)

美国, 佛罗里达

2018.01 - 2019.12

### 佛罗里达大学

工业系统工程硕士

美国, 佛罗里达

2016.01 - 2017.12

### 同济大学

物理学学士

毕业论文: X射线多层膜KB成像的强度均匀性校正

中国, 上海

2011.09 - 2015.06

## 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

- 基于历史药品销售与市场推广数据, 开发新药早期接受者识别模型, 帮助市场部在新药上市前精准定位目标HCP群体, 提升推广效率
- 负责多源数据整合与数据流程设计, 基于Kedro框架构建可复用数据管道, 保障数据一致性和模型输入质量
- 使用ChatGPT等大语言模型构建RAG流程, 识别日文FAQ与本地药品标签之间的潜在不一致性, 辅助审查流程
- 参与项目从Kedro迁移到Databricks平台, 协助重构特征存储架构, 提升模型特征的可管理性和扩展性

### Zenni Optical

数据科学家

中国, 北京

2023.06 - 2024.01

- 作为Rx眼镜处方识别API的主要开发者, 负责在现有OCR系统基础上持续迭代与优化, 构建FastAPI服务, 实现处方字段自动提取与结构化
- 针对日本市场定制处方解析流程, 处理语言差异及处方样式变体, 服务上线后覆盖率及解析准确性显著提升
- 构建模糊图像检测模块, 采用多任务学习方法识别低质量处方图像, 显著提升系统整体解析准确率
- 在GCP平台部署多版本服务, 结合CloudSQL实现多地区语言支持, 并通过BigQuery分析用户行为, 驱动功能迭代
- 使用Streamlit开发交互式可视化工具, 辅助跨团队沟通并支持产品评审

### 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

- 独立开发CyTOF图像分析工具包, 集成空间信息与单细胞表达数据, 通过并行化处理将效率提升10倍, 并构建Flask+Docker界面供生物研究人员使用
- 基于PyTorch实现Mask R-CNN模型, 用于H&E病理图像中细胞实例分割与六分类任务, 检测率82.5%, 分类准确率82.0%
- 针对不完善标注数据, 自定义模型损失函数, 成功挽救约20%的训练数据, 提升模型泛化能力
- 设计分层分类策略用于多种肿瘤病理图像融合, 助力泛癌种微环境建模与AI辅助诊断研究

### 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

- 主导多模态图像 (RGB、热成像、高光谱) 自动化处理流程建设, 开发图像预处理、统计分析和可视化模块, 大幅提高植物表型数据处理效率
- 基于Mask R-CNN实现植物叶片实例分割, 并构建追踪算法跟踪叶片生长周期, 提升生物实验数据质量
- 深度参与开源项目PlantCV开发, 新增图像分割与特征提取模块, 优化目标检测算法, 编写单元测试与用户文档, 参与教学推广
- 跨团队支持科研人员完成数据清洗与图表生成任务, 提升数据使用效率与科研成果可视化水平
- 担任科学培训与指导委员会联合主席, 组织技术分享与跨领域交流

### 佛罗里达大学医疗中心 精密医学智能合作部

美国, 佛罗里达州, 盖恩斯维尔

数据科学实习生		
2019.05-2019.08		
<ul style="list-style-type: none"><li>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</li><li>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</li><li>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</li></ul>		
个人项目		
医药内容推荐与摘要系统		
<ul style="list-style-type: none"><li>构建基于关键字检索和LLM摘要生成的推荐系统，整合本地结构化FAQ、药品标签段与PubMed文献内容</li><li>支持中英文药品名识别、结构化知识融合展示、Whoosh索引与OpenAI接口调用</li><li>提供可交互的Streamlit界面并已部署上线，用于探索医疗信息自动聚合与问答系统设计</li></ul>		
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
<ul style="list-style-type: none"><li>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</li><li>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</li><li>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</li><li>利用基于纯净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</li></ul>		
出版文章		
<ul style="list-style-type: none"><li>Sheng, H., Wang S., et al. “MTIA: An open-source python package for systematic multiplexed tissue image analysis” (in preparation)</li><li>Sheng, H., Gutierrez, J., Schuhl, H., Murphy, K. M., Acosta-Gamboa, L., Gehan, M., &amp; Fahlgren, N. (2023). Increasing the Throughput of Annotation Tasks Across Scales of Plant Phenotyping Experiments. Authorea Preprints.</li><li>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.</li><li>Panda, K., Mohanasundaram, B., Gutierrez, J., McLain, L., Castillo, S. E., Sheng, H., ... &amp; Slotkin, R. K. (2023). The plant response to high CO2 levels is heritable and orchestrated by DNA methylation. New Phytologist, 238(6), 2427-2439.</li><li>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.</li></ul>		
综合能力		
<ul style="list-style-type: none"><li>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</li><li>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</li><li>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</li></ul>		
奖项与证书		
<ul style="list-style-type: none"><li>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)</li><li>谷歌数据分析证书</li></ul>		