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, Seaborn, Plotly, Streamlit），MATLAB，Flask，HTML，CSS，JavaScript，Ajax
- 机器学习：PyTorch，Scikit-learn，Weights & Bias，TensorFlow，Keras，深度学习，计算机视觉，OpenCV，LangChain，检索增强生成（RAG），自然语言处理（NLP），HuggingFace
- 云计算与DevOps：AWS，GCP，Azure，Databricks，BigQuery，Docker，Kubernetes，Git
- 数据分析与可视化：PySpark，SQL，BigQuery，数据管道开发，EHR数据分析，MS Office

工作经历

Johnson & Johnson 中国，北京
数据科学家 2024.04 至今
- 整合和处理药品上线前的数据，包括潜在的类似药物数据（内部药物、竞争药物）、销售数据及市场推广数据等
- 通过复杂的特征工程任务，为模型训练提供高质量的输入数据；使用Kedro框架管理项目的数据流，确保数据的准确性和一致性
- 开发和优化了早期接受者分类模型，以预测新药上市后六个月内的市场接受情况，支持上线决策
- 利用ChatGPT API和其他大语言模型进行简单的Retrieval-Augmented Generation（RAG）任务，处理日语药品标签的解析
- 参与将Kedro管理的项目迁移至Databricks，并参与Databricks Feature Store的改造，以提高特征管理和可扩展性

Zenni Optical 中国，北京
数据科学家 2023.06 – 2024.01
- 调研用于自动化解释和解析眼镜处方的先进光学字符识别（OCR）技术，开发、部署及持续优化端到端Rx提取API服务
- 在Google Cloud平台（GCP）上成功推出并维护Rx提取的FastAPI服务的多个迭代；并提供专门针对日本市场语言差异和处方风格变化的定制版本
- 为API引入了能够接收不同国家/地区和语言输入的功能，确保这些信息能够被正确记录到CloudSQL的数据库中，以满足用户的个性化需求
- 为处方提取框架的图像质量评估模块引入处方模糊检测功能，包含数据采集、数据标注和多任务模型训练，以提高处方数据处理的准确性和效率
- 利用BigQuery数据库对处方提取API的使用数据进行深入分析，促进决策和服务的进一步优化
- 维护实时完备的技术文档以及软件发布变更日志
- 设计并制作多个 Streamlit 可视化应用，以辅助跨部门高效沟通

德克萨斯西南医学中心 定量生物医学研究中心 美国，德克萨斯州，达拉斯
数据科学家 2021.09 – 2023.05
- 开发基于Python的面向对象的CyTOF图像处理分析工具包，实现空间信息和单细胞基因信息的数据融合；利用并行运算提升10倍处理速度；独立设计并制作基于Flask框架的图形用户界面；实现Docker容器化
- 搭建基于PyTorch框架的 Mask R-CNN 模型，实现H&E染色的癌症病理组织影像的细胞检测、分类以及掩码分割（检测率：82.5%，6分类准确率：82.0%）
- 通过重新定义Mask R-CNN 模型的损失函数，避免由于非完善数据标签造成的20%数据的损失
- 设计分层分类损失函数，实现多种肿瘤病理组织影像的数据融合，辅助泛癌分析

唐纳德丹福思植物科学中心 (Donald Danforth Plant Science Center) 美国，密苏里州，圣路易斯
数据科学研究员 2020.02 – 2021.09
- 建立并完善包括预处理、处理、后期处理、统计分析及可视化等在内的自动化处理RGB、热成像、以及高光谱图像的算法
- 利用 Mask R-CNN 预训练模型对植物叶子进行实例分割；通过研发追踪算法来理解植物的生命周期
- 参与开源软件包PlantCV (Plant phenotyping using Computer Vision)的开发：优化及添加新的图像分析、图像分割、目标检测、特征提取等算法及相应单元测试；撰写软件文档；版本控制；参与组织PlantCV教学输出以及相关教学材料准备

- 为整个科研团队提供数据处理、可视化以及统计结果分析
- 作为科学培训和指导委员会联合主席，组织学术讲座

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

- 预处理原始数据，包括数据清洗、异常检测、批次合并、可视化等，并确定研究队列
- 从病人的主要生命体征数据中提取时间序列特征；调研处理不规则时序特征的文献，复现包括插值网络在内的算法，并通过对比分析临床数据来比较时序数据聚类结果

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

- 研发基于多实例学习的图像分割的机器学习算法，用于从植物微根管图片中自动检测植物根
- 搭建并完善自动分析处理植物高光谱图片及热成像图片的架构
- 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；研发基于孪生神经网络的分类友好型降维算法，实现相同植物种不同基因型的分类

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

- 流利的英文口语及书面沟通能力
- 较强的独立思考、工作能力，以及较强的学习能力，能够快速理解新知识、掌握新技术
- 较强的沟通交流和团队协作能力，可以同时与技术人员和非技术人员沟通技术工作

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