James O'Reilly

London, N1

(+44) 7478892291 | jamesdanieloreilly@gmail.com | jamesdanieloreilly | jamesdanieloreilly

Experienced data scientist and software developer with an exceptional background in mathematics, machine learning and software development. 3+ years experience applying novel and off-the-shelf data science solutions to complex real-world datasets in an ad-hoc or product-focused manner. Deep knowledge of user-focused product development in multidisciplinary environments. I currently develop an AI platform for drug discovery. Seeking data science and machine learning roles.

Technical Skillset

Data Science

- 2+ years experience implementing end-to-end data science solutions, including data acquisition, data cleaning, feature engineering, model development, model testing, model deployment and assessment.
- In-depth understanding of a broad range of methodologies across the ML landscape, including standard regression and classification approaches, deep learning, NLP, ensemble methods, reinforcement learning, network-based methods, graph embeddings, Bayesian graphical models and factor models.
- Languages and packages: Python, R, Matlab, NumPy, SciPy, Pandas, PyTorch, HuggingFace, SHAP, sklearn, BayesOpt, networkx and PySpark

Software Development and Engineering

- 2+ years professional test-driven development experience with internal and external facing products.
- Languages: Python, R, Matlab, Bash, SQL, Cypher, GraphQL. Limited experience with Julia, C, C#, JS, haskell.
- DevOps: Git, Docker, AWS ECR, CI, pytest, Kubernetes, Kubeflow, Jupyter, Jira

Data Engineering

- Extensive knowledge of cloud-based data solutions, including AWS S3, AWS Redshift, Databricks and Neo4j.
- Experience querying SQL, NoSQL and graph databases with SQL, Cypher and graphQL.

Work Experience

BenevolentAI

London, United Kingdom

Data Scientist

Oct 2021 - Present

- Technical Lead for an active target discovery program (osteoarthritis). Implemented factor models, large language models, and network diffusion models for drug target identification. Built a comprehensive dataset collection used by the internal ML-models for target prediction.
- Represented BenevolentAI as a spokesperson and event host at internal and external functions, including as an MC for company wide events alongside senior leadership.
- Developed transfer learning pipelines for projection of latent factors between genomics data. This allowed for robust signals from large datasets to be projected onto sparse metadata-rich datasets, enabling efficient patient stratification.
- Built a genomics-based biomarker discovery pipeline using ensemble methods. Achieved above state of the art prediction accuracy, allowing for the discovery of novel biomarkers of NASH fibrosis.
- Developed a data engineering pipeline for automated scoping of internal and external data sources across different data modalities, allowing data scientists to quickly assess data landscapes for diseases of interest, with a major impact on company strategy.
- Designed internal metrics to evaluate usage and performance of target identification software applications and products. Presented results quarterly to C-suite and executive leadership to inform future tech strategy. Advised on the integration of these metrics into software products to better facilitate internal reporting.
- Developed a network diffusion model for identification of targets using BenevolentAI's integrated knowledge graph.
- Performed bespoke statistical analyses to support drug discovery scientists in the context of multiple diseases. Presented results to technical and non-technical audiences to advise on target progression into portfolio.
- Built a StreamLit app to visualise single-cell gene expression, clustering and cell-type differentiation data to drug discoverers during target triage.
- Implemented and maintained DevOps infrastructure for bioinformaticians and data scientists, including CI, semantic release, Docker image optimisation and image storage using AWS ECR.

VIB

Leuven, Belgium

Data Scientist

Sep 2020 - Jun 2021

- Bioinformatics data scientist position VIB's lab for functional epigenetics.
- Developed data science pipelines for analysis of tumour heterogeneity in aggressive lung cancers, identifying sources of heterogeneity and linking these to patient outcome,
- Used latent variable models, Bayesian group factor analysis, multi-omic factor analysis, and trajectory inference to disentangle sources of cellular heterogeneity in Lung cancers.
- Used models of heterogeneity and inequality from economics and ecology to define novel metrics for heterogeneity in tumour cell populations.

DataCamp

Leuven, Belgium

Software Engineer Intern

Jun 2020 - Aug 2020

- SWE internship with DataCamp's Experience Engineering team. Implemented automated testing of code correctness for DataCamp's online courses.
- Built and optimised data science Docker images for automated deployment across DataCamp's learning platform.
- Maintained Datacamp's Python backend and automated code feedback system.

University of Bristol - Dept. of Computer Science

Bristol, United Kingdom

NLP Research Intern

Jun 2019 - Aug 2019

- Investigated the application of neural networks in determining the statistical structure of language.
- Studied and implemented the mathematics underlying neural networks, vectorisation of language, and associated NLP concepts.

Education

Katholieke Universiteit Leuven

Leuven, Belgium

M.Sc. in Bioinformatics

Sept 2019 - Present

- Graduated cum laude (75%). Specialised in the application of machine learning methods to high-dimensional genomics datasets.
- Thesis: *"Controlling intra-tumoural heterogeneity using the epigenome"*. Disentangled sources of variability in high-dimensional single-cell lung cancer datasets using latent variable modelling, Bayesian group factor analysis and transfer learning. In collaboration with the VIB lab for Functional Epigenetics.

University of Bristol

Bristol, England

B.Sc. in Mathematics and Computer Science (Joint Honours)

Sept 2016 - June 2019

- Final grade: 68%. Specialised in discrete mathematics, graph theory and information theory; focusing on codes, communication, and cryptographic schemes. Took additional modules in machine learning and computational neuroscience.
- Thesis: Designed and evaluated an interactive VR learning environment for calculus education, facilitating distance learning and learning for disabled students. Available on GitHub.