Steve Herrin

CONTACT INFORMATION

jobs@steveherrin.com
650-814-8865
San Jose, CA

www.github.com/steveherrin
www.linkedin.com/in/herrinsteve

SUMMARY

Engineering leader with over a decade of experience building and leading teams that use software, data, and machine learning to solve novel problems. Scientific (physics PhD) background with demonstrated adaptability to other fields like biotech.

EXPERIENCE

Pathos AI, Chicago, IL (Remote) June 2022 – present

Vice President of Engineering

- Grew agile data-focused engineering team from zero to four engineers
- Built a federated data catalog on GCP to search, utilize, understand, and audit access to thousands of biological datasets and analysis results
- Prototyped LLM embedding pipeline to catalog scientific literature using PyTorch
- Built data analysis and processing environment from scratch supporting Python and R using GCP, Terraform, Nextflow, and Docker
- Developed Slack bots with Typescript & Deno for engineering and operational tasks

D2G Oncology, Mountain View, CA April 2021 – May 2022

Staff Software Engineer

- Created *in silico* simulations of PCR and DNA sequencing, used for oligo design, QC, and automated processing of multiplexed sequencing runs
- Built a pipeline using pydantic to automate ETL and validation of Benchling LIMS data from an HTTP API to a PostgreSQL warehouse
- Developed a Python library exposing GraphQL and REST APIs for accessing and traversing a large knowledge graph of lab, sequencing, public, and analysis data

23andMe, Sunnyvale, CA January 2014 – March 2021

Engineering Management (∼2 years; title: engineering manager)

- Created and grew 3 machine learning and data -focused engineering teams totaling 16 engineers, including a mix of leads and individual contributors
- Led committee of engineering leads to standardize interviewing guidelines and open source release processes, subsequently adopted organization-wide

Engineering Individual Contributor (∼5 years; final title: sr. tech lead engineer)

- Built Django and HBase platform for internal and external researchers to dynamically query k-anonymized data for >10 million customers
- Developed containerized (Docker, AWS ECS) machine learning systems to ensure quality and reproducibility of models in a regulated medical device setting
- Implemented bioinformatics algorithms from literature in C++, Spark, and Rust
- Designed and implemented a data catalog with a unified API for accessing data/metadata across application-specific data stores, eventually used for all customer content
- Created web portal for external researchers to recruit for genomic studies and receive data back, increasing sales by 2% and producing strategic data-sharing agreements

- Migrated a 20 kLOC Django web application to the AWS cloud, upgrading the back-end from Python 2 to 3 and standardizing the front-end using React and Typescript
- Built and scaled 3 generations of distributed data pipelines with Celery, Luigi, and AWS to run Python, C++, and R algorithms operating on petabytes of genetic data
- Automated genotype calling for SNPs and genes using a combination of unsupervised and supervised ML techniques in SKLearn
- Developed and performed a maximum likelihood analysis combining private and public datasets to replace thousands of ineffective genotyping probes

SLAC National Accelerator Lab, Menlo Park, CA May 2008 – August 2013

Research Associate

- Developed batch data pipelines on scientific cluster using Python, C++, and shell scripts to routinely characterize detector by processing TB of calibration data
- Applied machine learning & statistics to improve detector energy resolution by 25%
- Implemented computer vision algorithms in C++ to repurpose detector for 3D cosmic ray reconstruction, reducing cosmogenic background uncertainty by 10x
- Built, programmed, & networked PLC control systems with 600+ channels of heterogeneous sensor data at remote underground site with unreliable connectivity
- Created a PHP logbook webapp with a MySQL backend for tracking lab work
- Coordinated hardware, analysis, and control software development with remote teams distributed around the world
- Mentored junior researchers on lab, coding, & statistical technique

Skills

Languages: Python, Rust, C, C++, SQL, Shell Scripting, Elm, JavaScript/TypeScript, R

Tools: AWS, GCP, NumPy, JAX, SciPy, Scikit-Learn, PyTorch, Mypy, Pydantic, FastAPI, Flask, Django, React, MySQL, PostgreSQL, Git, HBase, Spark, LaTeX, Blender, Unity

Other: Machine Learning, Data Analysis, Linear Algebra, Bayesian/Frequentist Statistics, Simulation, CI/CD, Sensors, Analog & Digital Electronics, Scientific Computing, Neutrino & Particle Physics, Radio (Amateur Extra License), Experienced Underground Miner

Education

Insight Data Science, Mountain View, CA
- Postdoctoral Fellowship

Stanford University, Stanford, CA
- Ph.D. (Physics)

Rice University, Houston, TX
- B.S. (Physics)