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 using software, data, and machine learning to solve novel scientific and technological problems. Experience building and managing engineering and data teams. Scientific (physics PhD) background with demonstrated adaptibility to other fields like biotech.

Experience

Pathos AI, Chicago, IL — June 2022 – present

Vice President of Engineering

- Grew the engineering team from zero to four engineers of varying seniority through recruiting and acquisitions
- Developed a data catalog and management system using PostgreSQL and Python to search for, manage, understand, and audit access to thousands of datasets and analysis results
- Prototyped large language model (LLM) pipeline for scientific literature review
- Used Google Cloud, Terraform, and Nextflow to orchestrate and automate distributed scientific computing jobs

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 several multiplexed sequencing runs per month
- 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 API library exposing GraphQL and REST hooks for accessing and traversing the relations of a large knowledge graph of lab, sequencing, and analysis data, as well as external data sets

23andMe, Sunnyvale, CA — January 2014 – March 2021

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

- Built application using HBase and Python to internal and external researchers to dynamically query k-anonymized data for >5 million customers using a SQL interface, an HTTP API, or a web front end
- Migrated a 20 kLOC Django web application to the AWS cloud, upgrading the backend from Python 2 to 3 and standardizing the front-end using React and Typescript
- Designed and implemented a Python library providing a unified API for accessing customer data across MySQL, HBase, and other data stores, eventually used for 100% of customer content
- Created Genotyping Services, a Django webapp on AWS allowing external researchers to easily run genomic studies, increasing sales by over 2% and leading to strategic data-sharing agreements

- Architected and led building of containerized (Docker, AWS ECS) systems to ensure quality, reproducibility, and rapid deployment of machine learning models, used for over 90% of production models
- Built 3 generations of distributed data pipelines with Celery, Luigi, and AWS Simple Workflow to run Python, C++, and R algorithms that impute, transform, and analyze petabytes of genetic data
- Developed and automated a maximum likelihood analysis combining private and public datasets that flagged $\sim 0.5\%$ of genotyping probes as bad for replacement in future platforms
- Automated genotype calling for SNPs and genes using a combination of unsupervised and supervised ML techniques

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

- Created and recruited for 3 machine learning and data -focused engineering teams totaling 16 engineers, including a mix of leads, senior, and junior level individual contributors
- Formed and led team of engineering leads to standardize interviewing guidelines and open source release processes, subsequently adopted by all engineering teams

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

Research Associate

- Applied machine learning algorithms & statistical analysis to improve detector energy resolution by 25%
- Repurposed the detector for 3D cosmic ray muon reconstruction using computer vision algorithms, yielding a 10x reduction in cosmogenic background uncertainty
- Created a PHP logbook webapp with a MySQL backend for tracking work on the EXO-200 experiment
- Built, networked, and programmed PLC control systems using over 600 channels of heterogeneous sensor data at a site with unreliable internet connectivity, successfully protecting $10M of liquid xenon
- Developed batch data pipelines using Python, C++, and shell scripts to routinely measure detector characteristics by processing TB of calibration data.
- Coordinated hardware and analysis software development with remote teams distributed around the world
- Mentored 1–2 junior graduate students (at any given time) on lab, coding, and statistical technique

Stanford University, Stanford, CA April 2008 – December 2011

Teaching Assistant (3 quarters)

- Supervised 8 to 12 undergraduate students in laboratory and classroom settings
- Communicated advanced physics concepts, including computational physics using MATLAB and Python

Rice University, Houston, TX May 2005 – May 2007

and **University of Washington**, Seattle, WA June 2006 – August 2006

Undergraduate Research Assistant

- Implemented (in C++) and evaluated random forest and boosted decision tree algorithms that contributed to the discovery of single top quark production by Fermilab's D0 experiment
- Investigated and benchmarked many different machine learning classification algorithms for their power to discriminate signals of supersymmetry from backgrounds

Skills

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

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

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

Open Source

SpookyOTP: A lightweight Python implementation of TOTP/HOTP authentication

Education

Insight Data Science, Mountain View, CA — December 2013

- Postdoctoral Fellowship

Stanford University, Stanford, CA — June 2013

- Ph.D. (Physics)

Rice University, Houston, TX — May 2007

- B.S. (Physics)

Selected Talks

Panelist, "Machine Learning"
XLDB Conference — 1 May 2018

"Migrating Bioinformatics Pipelines to the Cloud"
Biological Data Science (Biodata) Conference — 27 October 2016