Matthew Sooknah

Software engineer with experience in machine learning, computer vision, data engineering and computational biology

179 Noe Street, San Francisco CA 94114
401-588-2644 — mattsooknah@gmail.com
mattsooknah.github.io — github.com/mattsooknah

Experience

Calico Labs — **South San Francisco, CA**
Senior Machine Learning Engineer — *August 2019 – present*

- Developed state-of-the-art computer vision models for object segmentation, feature extraction, and representation learning on biomedical images (MRI, CT, and cell microscopy).
- Worked directly with biologists to initiate research projects, develop data collection and annotation protocols, and identify opportunities to apply machine learning and computer vision in Calico's aging research and drug development efforts.
- Improved code quality, data curation, documentation and deployment practices by drawing on previous experience in data engineering.
- Played a leadership role in defining project goals, running meetings and delegating responsibilities.
- Mentored a summer intern doing research on deep-learning techniques for CT image reconstruction, denoising, and domain-specific post-processing.
- Used PyTorch, TensorFlow, Keras, and scikit-learn for ML; OpenCV and ITK for image processing.

Data Platform Engineer — *August 2017 – July 2019*

- Built a data platform and web app for internal scientists to manage genomic samples, analyze and visualize results of experiments. Implemented server backend and data pipelines. Worked with UI designer to develop web-based GUI. Worked with scientists to collect requirements. These applications have been used for 3+ years to process thousands of experiments across basic research and drug development programs.
- Wrote pipelines to aid in assembling and publishing the genome sequence of an important cell-line (WI-38) from a variety of data sources. Implemented distributed pipeline for genome assembly, correction, and validation. Developed public-facing website and genome explorer to accompany the paper at wi38.research.calicolabs.com.
- Designed and co-taught a course on Python programming and data analysis for 30 internal wet-lab scientists, with the aim of empowering them to manage and analyze their own data without needing help from computational scientists. Students were able to apply what they learned to real data generated at Calico, increasing scientific productivity across the organization.
- Used Python, R, Google Cloud Platform, Docker, Airflow, MySQL.

10X Genomics — **Pleasanton, CA**
Scientist, Computational Biology — *January 2016 – August 2017*

- Contributed to development of Cell Ranger and Long Ranger software packages, which are the industry standard for processing genomic data generated by the 10x microfluidics platform.
- Developed and refined algorithms for quality control, gene expression quantification, high-dimensional clustering, statistical analysis, and data visualization.
- Profiled and optimized performance of Python and Rust programs for I/O and memory-intensive applications.

The Broad Institute of MIT and Harvard — **Cambridge, MA**
Software Engineer, Data Sciences & Data Engineering — *May 2014 – December 2015*

- Developed and optimized analysis pipelines with Scala and Java for petabyte-scale genomics data from one of the largest DNA sequencing centers in the world.
- Contributed to development and support of multiple open source, industry standard tools for genome analysis (Picard and HTSJDK).
- Worked on backend database and data model for early prototype of cloud genome analysis platform (terra.bio).
- Developed methods for analyzing gene expression and pathway activity from customized RNA sequencing assays, to gain insight into the mechanisms of autoimmune disease.

Nabsys — **Providence, RI**
Associate Scientist, Algorithms — *June 2013 – May 2014*

- Developed tools in Java for signal processing and genome analysis to support development of the Nabsys microfluidics-based DNA mapping technology.

Education

Massachusetts Institute of Technology — **Cambridge, MA**
S.B. Physics, GPA 4.9/5.0 — *2009 – 2013*

Skills

- **Programming Languages:** Python (expert); Java, Scala, Rust, JavaScript, R, SQL (familiar)
- **Tools and Frameworks**: PyTorch, TensorFlow, Keras, sklearn, OpenCV, GCP, Docker, Airflow

Publications

- Ilya Soifer, Nicole Fong, Nelda Yi, Andrea Ireland, Irene Lam, **Matt Sooknah**, et al. Fully Phased Sequence of a Diploid Human Genome Determined de Novo from the DNA of a Single Individual. G3 (Genes, Genomes, Genetics), September 2020. DOI: 10.1534/g3.119.400995
- Daniel O'Connell, Raivo Kolde, **Matt Sooknah**, et al. Simultaneous Pathway Activity Inference and Gene Expression Analysis Using RNA Sequencing. Cell Systems, May 2016. DOI: 10.1016/j.cels.2016.04.011

Presentations

- "Mapping, processing, and duplicate marking with Picard tools." BroadE Workshop on GATK Best Practices. Broad Institute, Cambridge, MA. March 2015.

Poster / Talk Contributions

- Florian Schmid, Georgios Koukos, Yi Liu, **Matt Sooknah** et al. "High-resolution kidney MRI in mice for longitudinal tracking of kidney volume and cyst burden." International Society for Magnetic Resonance in Medicine. Virtual Conference, May 2021. https://index.mirasmart.com/ISMRM2021/PDFfiles/0423.html
- Grace Zheng, Jessica Terry, Paul Ryvkin, **Matt Sooknah**, et al. "Single Cell RNA profiling of a Million Neurons by a Massively Parallel and Scalable Droplet Platform." Advances in Genome Biology and Technology. Hollywood Beach, FL. February 2017.
- Haynes Heaton, Patrick Marks, **Matt Sooknah**, et al. "Alignment and Variant Calling in Segmental Duplications with Linked-Read Data." Genome Informatics. Wellcome Genome Campus, Hinxton, Cambridge, UK. September 2016.