





George Stein, Ph.D.

Machine Learning Scientist

 george.f.stein@gmail.com  [georgestein.github.io](https://github.com/georgestein)  [georgestein](https://github.com/georgestein)  [george-stein](https://www.linkedin.com/in/george-stein)

SUMMARY

5+ years experience adapting and advancing machine learning techniques to solve applied problems at scale, 9 years in high performance computing and data analysis, with a proven track record of extracting meaningful insights from large datasets.

PROFESSIONAL EXPERIENCE

Lawrence Berkeley National Laboratory

Berkeley, CA

Postdoctoral Scholar, Machine Learning

Sept. 2019–Present

- Compiled a dataset of 75 million galaxy images and trained a series of convolutional neural networks on multi-node GPU-accelerated systems, leading to two peer reviewed publications. | *PyTorch*
- Implemented self-supervised learning to utilize unlabelled data and improve image classification performance by 170%.
- Created and deployed a [web app](#) for data discovery, and released community-adopted ML models. | *Streamlit*
- Principal investigator of data science proposal awarded over \$150,000 USD of GPU resources.

University of California, Berkeley

Berkeley, CA

Postdoctoral Scholar, Machine Learning

Sept. 2019–Present

- Designed a new anomaly detection method and applied it to win a blind anomaly detection challenge by characterizing an anomaly occurring in only 0.08% of 1 million events, outperforming 11 other teams.
- Built a neural network emulator to replace a costly calculation, achieving 10,000x speed-up. | *JAX*
- Developed a deep generative model to extract insights from noisy and incomplete high-dimensional time series data, solved the inverse problem to reduce the error of a key metric by 45% from the standard non-ML method. | *TensorFlow*

Canadian Institute for Theoretical Astrophysics

Toronto, ON

Graduate Researcher, High Performance Computing

Sept. 2014–Aug. 2019

- Constructed a pipeline for generating simulated observations of our universe on high performance computing systems, achieving 100x speed-up from previous methods, adopted by numerous forward modeling pipelines. | *Fortran*
- Implemented CNN segmentation models to predict regions of interest in physical simulations. | *TensorFlow*

SELECTED INDEPENDENT PROJECTS

Segmentation of Satellite Imagery

Nov. 2021–March 2022

- Trained an ensemble of segmentation models to [identify cloud cover](#) in satellite imagery as part of a popular data science competition, with my team of two finishing in the top 3% of 850 participants. | *Pytorch-Lightning*
- Leveraged public APIs to increase public dataset size 10-fold, and designed a custom set of physically motivated augmentations that improved model generalization and nearly eliminated overfitting.

ML-in-cosmology

Dec. 2018–Present

- Curator of [a comprehensive archive](#) of ML applications to the study of cosmology, with over 250 stars on GitHub.

EDUCATION

University of Toronto

Toronto, ON

Ph.D., Astrophysics

Sept. 2014–Aug. 2019

University of British Columbia

Vancouver, BC

B.Sc. (Honours), Physics and Astronomy, with distinction

Sept. 2010–May 2014

SKILLS

Programming Languages: Python (advanced), Fortran (experienced).

Frameworks/Packages: Proficient in PyTorch, TensorFlow, PyTorch-Lightning, Scikit-Learn, NumPy, SciPy, Pandas, Jupyter Notebook, Matplotlib, Git, Bash, and Streamlit. Familiar with JAX, SQL, CSS, HTML, TensorFlow Probability, Pyro, and NumPyro.

Machine Learning Experience: Computer Vision, CNNs, Self-supervised learning, Generative models, Density estimation, MCMC sampling, Dimensionality reduction & Clustering (Autoencoders, TSNE, UMAP, k-NN), Classification, Segmentation, etc.

COMMUNICATION AND TECHNICAL WRITING

Lead author of 7 publications in machine learning and cosmology. Presented research at 50+ venues.

Full publication list and additional projects available at [georgestein.github.io](https://github.com/georgestein)