

Danny Collinson

dannycollinson12@gmail.com | 610-290-3410 | West Chester, PA, USA
linkedin.com/in/danny-collinson | github.com/dannycollinson | dannycollinson.com

Summary

Data Scientist and recent Caltech graduate with five years of experience building data and machine learning solutions in academic and industry settings. Skilled at developing pipelines for processing and analyzing large datasets across biological, geospatial, and astrophysics applications. Experienced in applying advanced machine learning techniques, including state-of-the-art vision and language models. Co-founder of a pre-seed AI startup, I designed and engineered our machine learning models, data pipelines, databases, and underlying infrastructure. Significant experience developing statistical models to extract insights from complex datasets, including teaching graduate students at Caltech. With a diverse set of experiences, a deep skill set, and unmatched determination, I am equipped to tackle any challenge that I face and make a major impact on any team.

Education

California Institute of Technology

September 2019 - June 2024

Bachelor of Science (BS), Computation and Neural Systems

Pasadena, CA

- GPA: 4.0/4.0
- Relevant Coursework: Large Language and Vision Models, Data Analysis and Statistical Inference, Relational Databases, LLMs as Agents, Deep Learning, Machine Learning and Data Mining, Computational Biology and Bioinformatics

Experience

Data Science Intern

June 2023 - September 2023

Recursion Pharmaceuticals

Salt Lake City, UT

- Developed 2 new statistical metrics for analysis of large experimental datasets that were used to improve model performance
- Implemented advanced machine learning techniques in PyTorch to improve data processing efficiency and reduce costs
- Deployed monitoring tools to the data science and QA teams in collaboration with a 20-person cross-functional team, ensuring data integrity for downstream processing and models

Computational Biology Researcher

May 2022 - September 2022

Parker Lab at Caltech

Pasadena, CA

- Built data pipeline leveraging deep learning and statistical modeling to accelerate image processing speed by a factor of 100
- Automated microscopy analysis and increased measurement accuracy by an estimated 10% for an upcoming publication

Computational Astrophysics Researcher

May 2020 - January 2021

Harrison Lab at Caltech

Pasadena, CA

- Implemented statistical analysis methods using common data science tools including Python, NumPy, pandas, SciPy, and Jupyter notebooks on HPC clusters to perform source classification and deliver insights from large datasets
- Initiated development of an end-to-end data pipeline for automated data processing and classification, improving processing speed and efficiency in handling large datasets and decision-making

Teaching Assistant

September 2023 - December 2023

Professor Justin Bois at Caltech

Pasadena, CA

- Instructed 85 students in the graduate-level course Introduction to Data Analysis in the Biological Sciences, developing their skills in statistical modeling, numerical optimization, data visualization, and exploratory data analysis in Python

Skills

Programming Languages: Python, SQL, PyTorch, NumPy, pandas, scikit-learn, Jupyter, Matplotlib, SciPy, Bokeh, seaborn

Tools: Git, GitHub, AWS, Google Cloud, HPC, Docker, Bash, shell, Linux, MCMC, APIs, PostgreSQL, MySQL

Topics: deep learning, statistics, computer vision, LLMs, Bayesian stats, large datasets, prompt engineering, exploratory analysis

Projects

Mavira AI, Co-founder and Machine Learning Engineer

September 2024 - Present

- Co-founder and Machine Learning Engineer for pre-seed AI startup for personalized second-hand fashion recommendations
- Transformed a business idea into reality by designing data pipelines, ML training and inference frameworks, and PostgreSQL databases, all custom-built to develop models using PyTorch on Google Cloud and serve them to the production website

Temperature Prediction from GIS Spectra, Project Lead

September 2023 - December 2023

- Led team of 3 in ML project to predict surface temperatures from spectral data, achieving error of less than 1 C
- Created a new ML dataset from raw NASA data in collaboration with JPL scientists to study of novel research questions
- Built model training and testing frameworks and designed CNN-based model architectures alongside Professor Katie Bouman for use with autoencoder-generated embeddings to deliver accurate predictions