William Thistlethwaite

williamthistle@gmail.com LinkedIn GitHub Google Scholar Personal Website

Qualifications

- Data scientist with 10 years of experience in machine learning, data engineering, database management, data modeling, version control, API development, data visualization, deep learning, dimensionality reduction, and statistical analysis.
- Previous work has been published in high impact scientific journals, including *Cell*, *Cell Systems*, and *Nature*.
- Proven track record of partnering with scientific labs and engineering teams to achieve large-scale objectives.

Education

Ph.D. in Quantitative and Computational Biology **2019 – 2024 (anticipated)**

Princeton University, Princeton, NJ

Awards: Institute Scholars Award

M.A. in Computer Science **2012 – 2014**

Brandeis University, Waltham, MA

B.A. in Mathematics and English **2005 – 2009**

University of North Carolina at Chapel Hill, Chapel Hill, NC

Awards: UNC Chapel Hill Honors Program, Pi Mu Epsilon (National Mathematics Honor Society)

Experience

Princeton University, Princeton, NJ **September 2019 – Present**

Graduate Student, Troyanskaya Lab

- Developed SPEEDI, a fully automated end-to-end framework for data-driven batch inference, data integration, and cell type labeling for single-cell data (**Cell Systems #1**). Published **R package** on GitHub. Wrote back-end code using Bash and R to support an **interactive web server** and provided front-end design specifications.
- Performed multi-omic integration on single-cell transcriptomic and epigenomic data from blood samples to study regulatory processes that persist in the innate immune system after resolution of influenza infection (**bioRxiv**).
- Collaborated with other researchers in **Dog Aging Project** consortium to create cloud-based pipelines on Terra for genomic data ingestion and processing (**Nature**).

Baylor College of Medicine, Houston, TX **August 2014 – July 2019**

Senior Software Engineer and Data Analyst, April 2019 – July 2019

- Facilitated a smooth transition by creating comprehensive documentation, conducting knowledge transfer sessions, and providing ongoing support to ensure the team could easily continue development on projects after my departure.

Software Engineer and Data Analyst II, July 2016 – April 2019

- Served as technical lead for the Data Coordination Center of the NIH Common Fund's Extracellular RNA Communication Consortium (**Cell #1**). Developed data processing pipelines in Ruby to extract, transform, and load data from thousands of biological samples, efficiently processing terabytes of data on a high performance computing cluster. Taught workshops on data submission and statistical analysis of pipeline output at 5 scientific conferences.
- Designed and built the **Extracellular RNA Atlas** (**Cell #2**). Created RESTful **JSON-LD API** using Ruby on Rails that adhered to FAIR (Findability, Accessibility, Interoperability, and Reusability) principles, enabling seamless programmatic access to Atlas metadata and data. Metadata were stored in MongoDB and were standardized using clinical ontologies including SNOMED CT and LOINC.
- Served as project liaison for the **Virtual Biorepository**. Gave oral and poster presentations at 7 conferences, participated in monthly calls with stakeholders, and provided feedback to software engineers for fixing bugs and improving features.

Software Engineer and Data Analyst I, August 2014 – July 2016

- Helped develop the extracellular RNA processing toolkit (**exceRpt**), a bioinformatics platform specialized for preprocessing, aligning, and normalizing human and mouse extracellular RNA data (**Cell Systems #2**). Engineered parallelization that resulted in 20× speedup for data processing.

Technical Skills

Languages: Python, R, Ruby, Bash, SQL
Frameworks: PyTorch, XGBoost, Keras, scikit-learn, pandas, NumPy, Matplotlib, seaborn, tidyverse
Databases: PostgreSQL, NoSQL (MongoDB)
Platforms: Git, SVN, MLflow, Docker