# William Thistlethwaite

Princeton, NJ williamthistle@gmail.com LinkedIn GitHub 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, productionizing code, and statistical analysis.
- Previous work has been published in high impact scientific journals, including *Cell*, *Cell Systems*, and *Nature*.
- Adept at collaborating with interdisciplinary teams to achieve large-scale objectives.
- Seeking opportunity as a **Data Scientist** or **ML Engineer** to drive impactful data-driven solutions.

## Education

**Ph.D. in Quantitative and Computational Biology** **Completion in Dec 2024 (anticipated)**
*Princeton University, Princeton, NJ*

**M.A. in Computer Science**
*Brandeis University, Waltham, MA*

**B.A. in Mathematics and English**
*University of North Carolina at Chapel Hill, Chapel Hill, NC*

## Technical Skills

**Programming Languages**: Python, R, Ruby, Bash, SQL, JavaScript
**Machine Learning Frameworks**: PyTorch, XGBoost, Keras, scikit-learn
**Data Analysis & Visualization**: pandas, NumPy, Matplotlib, seaborn, tidyverse
**Cloud & DevOps Platforms**: Google Cloud Platform (GCP), MLflow, Docker
**Databases**: PostgreSQL, NoSQL, MongoDB
**Version Control**: Git, GitHub, SVN
**Project Management**: Jira, Asana

## Experience

**Princeton University, Princeton, NJ** **September 2019 – Present**
*Graduate Student, Troyanskaya Lab*

- Developed **R package** for 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**). Wrote back-end code using **Bash** and **R** to support an **interactive web server** and provided design specifications to front-end software engineer.
- Applied Bayesian hierarchical modeling and other analytical techniques to study regulatory processes that persist in the innate immune system after resolution of influenza infection (**bioRxiv**).
- Collaborated with team of **4+** researchers in Dog Aging Project consortium to create **Google Cloud Platform** based pipelines for genomic data ingestion and processing (**Nature**).

**Baylor College of Medicine, Houston, TX (Full Time)** **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 managing **20+ external stakeholders** for the Data Coordination Center of the 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 **5** conference workshops on data submission and statistical analysis of pipeline output.
- Designed and built the **Extracellular RNA Atlas** (**Cell #2**) using **Ruby** and **JavaScript**. 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 from thousands of samples. Metadata were stored in **MongoDB** and were standardized using clinical ontologies.
- Served as project liaison for the **Virtual Biorepository**. Gave oral and poster presentations at **7** conferences, participated in monthly calls with **10+ external stakeholders**, and provided feedback to software engineers for fixing bugs and improving features, resulting in **50% reduction** in user-reported issues.

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

- Developed the extracellular RNA processing toolkit (**exceRpt**), a bioinformatics platform specialized for aligning and normalizing extracellular RNA data, using **Make** and **Ruby** (**Cell Systems #2**). Engineered parallelization that resulted in **20× speedup** for data processing.