Tanay Shah

Boston, MA | (503)-680-6579 | shah.tanay2@northeastern.edu | linkedin.com/in/tanay2/
Availability: August - December 2025

EDUCATION

Northeastern University Boston, MA
Bachelor of Science in Data Science *September 2023 - May 2026*

- GPA: 3.76/4.00
- Relevant Coursework: Algorithms & Data, Database Design, Advanced Programming with Data, Foundations of Data Science, Fundamentals of Computer Science 1 & 2, Discrete Structures

EXPERIENCE

Capital One Dallas, TX
[Incoming] Software Engineer Interm *June 2025 - August 2025*

Harvard Medical School Boston, MA
Student Researcher (Gupta Lab) *July 2024 - Present*

- Performed genomic and causal analysis on data from 30,000+ patients of Mass General Brigham and associated hospitals.
- Identified trends and triggers associated with the onset of rhabdomyolysis, a life-threatening degenerative muscle tissue condition.

Northeastern University Boston, MA
Machine Learning Researcher (Gyori Lab) *May 2024 - August 2024*

- Developed an ML model to query and score new PubMed papers on their relevance to the Bioregistry using TF-IDF vectorization and aggregating classifiers. Achieved an AUC-ROC score of 0.991, indicating highly precise results, significantly expediting the process of discovering and curating new resources.
- Designed a workflow to automate the monthly deployment of the above ML pipeline's results to a GitHub issue.
- Implemented a script to automate the analysis and visualization of changes in the Bioregistry's metadata over time. Utilized to create visuals for presentations that secured lab funding.
- Updated and added 1,500+ lines of metadata in the Bioregistry, fixing degraded resources and curating new entries, enhancing the registry's comprehensiveness and accuracy.

PROJECTS

Paper Ranking Pipeline (GitHub**)** *May 2024 - August 2024*

- ML model used to score new PubMed papers on their relevance to the Bioregistry, using TF-IDF vectorization and aggregating classifiers.

TECHNICAL SKILLS

Programming Languages: Python, SQL, Java, Kotlin, C++, YAML, HTML/CSS/JavaScript
Tools/Libraries: Git, Pandas, Scikit-learn, NumPy, Matplotlib, Microsoft Excel
Concepts: Data Structures and Algorithms, Object-Oriented Programming, Agile Methodology, REST APIs, Web Development, Relational Databases, Machine Learning, NLP, Data Science and Engineering