Arumoy Shome

PhD Candidate | Software Engineering for AI

Phone: +31 611641508 **Email:** contact@arumoy.me **Location:** Amsterdam, NL
LinkedIn: arumoyshome **GitHub:** arumoy-shome **Website:** arumoy.me

Professional Summary

5+ years of experience in Data Science and Software Engineering. Led 4 end-to-end projects and pioneered 3 frameworks to improve the MLOps workflow. Hands-on in Python, large-scale data mining, and machine learning. In possession of a valid work permit.

Education

PhD, Software Engineering for Artificial Intelligence *Delft University of Technology* **Netherlands** *2021-Present*
M.Sc. Computer Science, Big Data Engineering *Vrije Universiteit Amsterdam* **Netherlands** *2018-2020*
B.Sc. Honors Systems Design Engineering, Entrepreneurship Minor *University of Waterloo* **Canada** *2012-2018*

Technical Skills

- **Data Science & Machine Learning:** Python, Scikit-Learn, PyTorch, Statistical Analysis, Multi Layer Perceptrons, Graph Convolutional Neural Networks
- **Data Analysis & Engineering:** Pandas, Data Mining, Big Data Engineering, PySpark, Data Pipeline Development, Data Collection
- **Visualization & Analytics:** D3.js, Three.js, Information Visualization, Visual Analytics
- **Development Tools:** Git, Docker, Linux, Unix, Nix
- **Programming Languages:** Python, Ruby, JavaScript, HTML, SCSS, Bash, LaTeX
- **Web Development:** Ruby on Rails, UI/UX Development, Object-Relational Mapping, D3.js
- **Research:** Data Quality Analysis, Fairness Testing, Static Analysis, Quantitative Research

Professional Experience

PhD Candidate | Data Science *Delft University of Technology* **Delft, Netherlands** *2021-2025*

- Developed large-scale data mining pipeline to process 297,800 Jupyter notebooks (283 GB) from GitHub and Kaggle. Extracted 3 million code statements using Python, Pandas, and Bash. Released results as an open-source dataset to enable future research. Project website: https://github.com/arumoy-shome/shome2023notebook
- Proposed data-centric fairness testing methodology for machine learning models. Evaluated 4 ML algorithms against 2 fairness metrics across 5 datasets with 1,600 end-to-end ML pipeline executions using Python and Scikit-Learn. Project website: https://github.com/arumoy-shome/shome2022qualitative
- Pioneered "data smells" framework for ML dataset quality assessment. Analyzed top 25 ML datasets from Kaggle using Python and Pandas. Identified 14 data quality antipatterns and published the findings as an open-source catalog. Project website: https://arumoy.me/data-smells
- Executed the data and ML pipelines on a Linux server and distributed the workloads across 20 CPU cores using Bash and Unix commands. Used Git and Docker to provide reproducible software artifacts.
- Managed 2 M.Sc. research projects, and an edX MOOC course on Unix tools with 1000+ active students.
- Delivered technical talks, poster presentations and guest lectures to academic and industry audiences. Translated technical concepts into actionable insights for diverse stakeholders.

Research Intern | Data Science *Netherlands eScience Center* **Amsterdam, Netherlands** *2019-2020*

- Built ML pipeline for neutrino detection in the KM3NeT Neutrino Telescope, implementing Multi Layer Perceptrons and Graph Convolutional Neural Networks to process high-volume particle physics data. Project website: https://github.com/arumoy-shome/km3net
- Delivered solution that outperformed existing GPU-based systems in filtration quality while meeting strict performance requirements for real-time particle detection.
- Collaborated across interdisciplinary teams including particle physicists, GPU engineers, and computer scientists, translating complex AI requirements into practical implementations.

Web Developer Intern *Shopify* **Ottawa, Canada** *2015-2016*

- Collaborated with developers, designers and product managers to implement UI/UX features such as web components, animations and styling on a mature Ruby on Rails project using Ruby, JavaScript, HTML, & SCSS.
- Applied Object-Oriented Programming (OOP) principles and Test-Driven Development (TDD) to refactor code and improve test coverage.
- Used Object-Relational Mapping (ORM) to optimize database queries and reduce server response time.

Technical Projects

ACE: Art, Color and Emotions *ACM International Conference on Multimedia* *2019*

- Built ACE, a visual sentiment analysis platform by developing custom ML models trained on the large-scale OmniArt dataset (512 GB) to enable data-driven analysis of artistic emotions. Demo video: https://youtu.be/B1ZM6EQgEvU
- Designed and implemented full-stack solution featuring intuitive D3.js interface with optimized interaction patterns and scalable web architecture capable of handling high-volume image processing and real-time sentiment analysis.

3D Kadaster *University of Amsterdam* *2018*

- Developed 3D model of all buildings in The Netherlands using AHN2 point cloud dataset (1.6 TB) and BAG building polygons dataset (177 GB).
- Processed massive geospatial datasets using PySpark distributed computing framework for scalable data processing.
- Executed algorithms on SurfSara supercomputer infrastructure for high-performance geospatial analysis.
- Created interactive 3D visualizations using Three.js for web-based exploration of national building infrastructure.

Elevate *University of Waterloo* *2017*

- Developed an improved and cost-effective alternative to state-of-the-art cognitive assessment tools for Down Syndrome using web technologies, adaptive learning, and human-centric design. Project website: https://arumoy.me/elevate
- Established international research partnership with Waterloo Regional Down Syndrome Society (Canada) and Fundacion Paraiso Down (El Salvador) to explore specialized educational resource needs through comprehensive user surveys and interviews.
- Implemented iterative design methodology with usability testing, user testing, and engagement testing as primary validation protocols.
- Developed comprehensive business plan and presented product at multiple startup incubators and pitch competitions to obtain funding.

Professional Activities

- Collaborated as committee member for International Conference on AI Engineering (CAIN) '25. Demonstrated teamwork, communication, and project management skills to coordinate logistics and enhance engagement for 125 participants.
- Served as program committee member for DeepTest '25 and NL-based Software Engineering (NLBSE) '24 international workshops. Critically evaluated and provided constructive feedback on 4 technical research papers.
- Writes and maintains open-source shell scripts using Python and Bash to automate frequent tasks and improve personal workflows. Publishes technical implementation details as blog posts.

Languages

- **English** [Native], **Dutch** [Basic] - A2, **Hindi** [Native], **Bengali** [Native]