Arumoy Shome

+31 611641508 | Amsterdam, The Netherlands | contact@arumoy.me | github.com/arumoy-shome | linkedin.com/in/arumoyshome | arumoy.me

Professional Summary

Data Scientist with 5+ years of experience at Delft University of Technology, Netherlands, and prior web development experience at Shopify, Canada. Led four end-to-end projects and pioneered three frameworks to improve the MLOps workflow. Hands-on in Python, large-scale data mining, and machine learning. **Authorized to work without sponsorship**.

Technical Skills

- **Data Science and Machine Learning:** Python, Scikit-Learn, PyTorch, Numpy
- **Data Engineering and Analysis:** Pandas, SQL, PySpark
- **Visualization and Analytics:** Seaborn, Matplotlib, D3.js, Three.js
- **Development Tools:** Git, Docker, Linux, Unix, Nix
- **Programming Languages:** Python, Ruby, JavaScript, Bash, HTML, CSS, 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

Data Science and Applied Research (PhD Candidate) 2021 — 2025

Delft University of Technology *Delft, The Netherlands*

- Conducted large-scale empirical analysis of code quality in ML Jupyter notebooks. 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 classification 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.

Data Science (Research Intern) 2019 — 2020

Netherlands eScience Center *Amsterdam, The Netherlands*

- Built ML pipeline for neutrino detection in the KM3NeT Neutrino Telescope, implementing Multi Layer Perceptions and Graph Convolutional Neural Networks to process high-volume particle physics data. Project website: https://github.com/arumoy-shome/km3net
- Delivered solution that outperformed state-of-the-art 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.

Software Engineering (Web Developer Intern) 2015 — 2016

Shopify *Ottawa, Canada*

- 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, & CSS.
- 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

3D Kadaster 2018

- Developed 3D model of all buildings in the Netherlands using AHN2 point cloud dataset (1.6 TB) and BAG building polygon 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.

ACE: Art, Color and Emotions (https://youtu.be/B1ZM6EQgEvU) 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.
- 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.

Elevate (https://arumoy.me/elevate) 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.
- 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.

Education

Delft University of Technology Delft, The Netherlands
PhD Computer Science, Software Engineering for Artificial Intelligence *2021 — Present*

VU University Amsterdam Amsterdam, The Netherlands
M.Sc. Computer Science, Big Data Engineering *2018 — 2020*

University of Waterloo Waterloo, Canada
B.Sc. Systems Design Engineering, Entrepreneurship Minor *2012 — 2018*

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 Natural Language Based Software Engineering (NLBSE) ‘24 international workshops. Critically evaluated and provided constructive feedback on four 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)