Arumoy Shome

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

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 Scientist and Applied Researcher (PhD Candidate) Jun 2021 — Jun 2025
Delft University of Technology *Delft, The Netherlands*

- 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.
- Applied NLP techniques (text representation, feature extraction and vectorization) to perform exploratory data analysis (EDA) on 3 million python code statements and generated 26 insights on ML pipeline reliability.
- Designed stratified random sampling strategy using TF-IDF scoring to mitigate selection bias in the large dataset and reduce oversampling trivial outliers.
- Developed an end-to-end automated ML pipeline (data ingestion, feature engineering, model training and evaluation) using Python, Scikit-Learn and Bash. Evaluated 4 ML classification algorithms (Logistic Regression, Decision Trees, Random Forest and Ada Boost) against 2 fairness metrics (Disparate Impact and Statistical Parity Difference) across 5 tabular datasets. Project website: https://github.com/arumoy-shome/shome2022qualitative.
- Proposed a data-centric fairness testing methodology to reduce development time and computational costs of ML pipelines. Designed experiments using data from 1,600+ pipeline executions and used statistical hypothesis testing (student t-test), correlation analysis and regression to validate findings.
- Pioneered "data smells" framework for ML dataset quality assessment. Performed EDA and feature engineering on top 25 ML tabular 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 and open source software artifacts released under the Creative Commons Attribution (CC-BY) license.
- Presented results through six publications at international conferences and scientific journals. Delivered technical talks, poster presentations and guest lectures to academic and industry audience, communicating technical concepts into actionable insights for diverse stakeholders.
- Managed two M.Sc. research projects, and an edX MOOC course on Unix tools with 1000+ active students.

Deep Learning Research Engineer (Internship) Jul 2019 — Mar 2020
Netherlands eScience Center *Amsterdam, The Netherlands*

- Built an end-to-end deep learning pipeline for real-time neutrino detection in the KM3NeT Neutrino Telescope using Python, Pandas and PyTorch. Project website: https://github.com/arumoy-shome/km3net.
- Developed a novel signal processing pipeline using Multi Layer Perceptrons achieving 92% accuracy in data filtration quality and a 12% improvement over the state-of-the-art GPU-based solution.
- Designed and implemented a Graph Convolutional Neural Network for event node classification in particle physics data, achieving 67% accuracy on event detection.
- Collaborated across interdisciplinary teams including particle physicists, GPU engineers, and computer scientists, translating complex AI requirements into practical implementations.

Software Engineer (Internship) Sep 2015 — Sep 2016
Shopify *Ottawa, Canada*

- Collaborated with developers, designers and product managers to implement 25+ 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 by 7%.
- Used Object-Relational Mapping (ORM) to optimize database queries and maintain server response times to under 100ms.

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 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. Project demo: 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.

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. 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.

Education

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

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

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

Languages

English (Native), **Dutch** (Basic - A2), **Hindi** (Native), **Bengali** (Native)