Debanjan **Datta**, PhD

Applied Scientist II
Amazon Web Services

Email: ddatta@vt.edu
Web: https://www.linkedin.com/in/debanjan-d
Phone: 408-314-0878

Seattle, WA, US

Summary

Debanjan graduated with PhD in Computer Science from Virginia Tech in 2022, advised by Dr.Naren Ramakrishnan, with a focus on Machine Learning. His research interests include general data mining and machine learning, with subject areas including anomaly detection, time series mining, explainable AI, NLP and LLMs. He has led and worked on research projects with real-world, complex, and large-scale data from different domains and applications, along with cross-functional collaboration. He is currently part of Agentic organization in AWS, with focus on data processing agents. As part of AWS Sagemaker's managed AutoML service team, he designed scalable and effective solutions for a variety of AutoML problems across data types (tabular, text, and image), including building LLM-based (GenAI) agents for AutoML. He also has experience in designing and executing large-scale experiments and analyses to solve ML tasks with a focus on customer-facing solutions. In addition to his research experience, he has professional experience in software engineering that drives his interest in building real-world systems with customer impact.

Education & Training

PhD Computer Science	Virginia Tech , GPA 3.9	Arlington, VA, US	Aug, 2016 - May,2022
Masters in Computer Science	Virginia Tech	Arlington, VA, US	Aug, 2016 - May, 2021
Bachelor of Engineering, Computer Science	Jadavpur University	Kolkata, India	Aug, 2009 - June, 2013

Research & Professional Experience

Amazon Web Services — *Seattle, WA, US*
Applied Scientist II, Team: AutoML and Data Processing Agents — June 2022 - Present

- Designed systems for automated data preparation and transformation pipeline for AutoML, including MCP based systems. Worked on multi-agent system for chat-driven AutoML, on LLM agent, and underlying pipeline synthesis tool. Collaborated with engineering & product to provide short- and long-term solutions to general tabular data mining problems.
- Performed research comparative evaluation and adoption of models for transfer learning for AutoML, in Computer Vision and NLP. Performed large-scale experiments for LLM fientuning with PEFT and accelerated inference for LLMs in resource-constrained environments.
- Performed research on novel research directions in MLOps, creating a testing framework along with synthetic data generation framework for monitoring and quantifying model performance.

Amazon Web Services — *Seattle, WA, US*
Applied Scientist Intern, Team: Lookout For Metrics — May 2021 - Aug, 2021

- Achieved key insights to disambiguate handling time series series types for anomaly detection through literature review, large-scale experimentation and analysis. Worked with multiple stakeholders incluyding engineering and science teams.
- Developed methods to discern specific data types relevant to application specific anomaly detection objectives, and proposed changes to system architecture for improved efficacy.

Yahoo Inc — *Sunnyvale, CA, US*
Software Engineer (Level 2), Team : Yahoo! Sports, back-end/API — Jan, 2015 – July, 2016

- Ownership and delivered key features on: core components for data ingestion and processing for internal REST API, live feed processing and optimized data queries.
- Independently created and improved upon internal frameworks and tools for data feeds, Database and API optimization (flexibility, scalability, performance)
- Effectively collaborated with multiple teams including devOps for end-to-end development and deployment. Delivered time-sensitive and real-time features to address data-ingestion pipeline and API service issues.

Yahoo Inc — *Bangalore, India*
Software Engineer (Level 1), Team : Yahoo! Mobile Web, back-end/API — July, 2013 – Dec, 2014

- Accomplished design and test driven development of core backend modules for serving Yahoo Mobile homepage and multiple media verticals.
- Delivered features to provide dynamic experience to users mobile phones (across platforms, devices,regions and languages), API speed optimization, Internal frameworks & tools, CICD integration & migrations
- Ownership of key platform modules, delivered time-constrained troubleshooting efforts for production, solved multiple end-user facing issues in real-time deployed systems.

Virginia Tech — Arlington, VA, US
Graduate Research Assistant, PhD Research — *Aug,2016 - May,2022*

- *Anomaly Detection and Explainability in Trade Data*:
  - Created an end-to-end framework for automated detection and analysis of suspicious timber transactions in real world shipment data, incorporating domain knowledge at multiple stages through collaboration with domain experts from WWF.
  - Formulated novel approaches for (i) unsupervised anomaly detection for categorical & heterogeneous tabular data (ii) human-in-the-loop approach to anomaly detection (iii) explainability and algorithmic recourse in anomaly detection. Key metrics improved were Average Precision, scalability and avoidance of hyperparameter tuning.
- *AI in Transportation Safety (Virginia Tech Transportation Institute)*:
  - Performed research, developed models for multiple traffic safety related projects with different research focus. (i) Developed temporal LSTM based model for driver gaze detection (Data Mining). (ii) Led system design for crash prevent-ability determination using post-accident reports (data mining). (iii) Transfer learning for semantic segmentation on real traffic data (Computer Vision). (iv) Created a expert driven model for tracking adoption of AI topics in transportation research from open source indicators (NLP).
- *Protein Function Prediction*: Developed a deep learning model (LSTM with attention mechanism) for toxin identification in nucleotide sequences.
- *Next Generation Social Science & Fragile Families Challenge*: Designed a graph based recommender system customized for social science literature based on citation networks, along with data ingestion and front end integration.
- *Influenza Forecasting*: Research on time series models based models influenza forecasting, with exogenous inputs and mechanistic models. Improved upon an existing model and participated in CDC ILI forecasting challenge 2017.
- **Graduate Courses Conducted (Teaching Assistant)** : Advanced Data Structures, ML with Big Data, Social Media Analytics

Publications and Presentations

Papers

[1] **Debanjan Datta**, Feng Chen, and Naren Ramakrishnan. "Framing Algorithmic Recourse for Anomaly Detection". In: *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 2022.

[2] Brian J. Goode and **Debanjan Datta**. "A Geometric Approach to Predicting Bounds of Downstream Model Performance". In: *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. KDD 2020. 2020.

[3] **Debanjan Datta** et al. "TimberSleuth: Visual anomaly detection with human feedback for mitigating the illegal timber trade". In: *Information Visualization*. 2023.

[4] **Debanjan Datta** et al. "Detecting Suspicious Timber Trades". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. IAAI, 2020.

[5] **Debanjan Datta** et al. "Scrutinizing Shipment Records To Thwart Illegal Timber Trade". In: *Proceedings of Outlier Detection Workshop, ACM SIGKDD*. 2021.

[6] Brian J Goode, **Debanjan Datta**, and Naren Ramakrishnan. "Imputing Data for the Fragile Families Challenge: Identifying Similar Survey Questions with Semiautomated Methods". In: SAGE Publications Sage CA: Los Angeles, CA, 2019.

[7] Sathappan Muthiah, **Debanjan Datta**, et al. "ProtTox: Toxin identification from Protein Sequences". In: *14th Machine Learning in Computational Biology (MLCB) Meeting, NeurIPS'19*. 2020.

[8] Matthew J. Salganik et al. "Measuring the predictability of life outcomes with a scientific mass collaboration." In: *Proceedings of the National Academy of Sciences*. 2020.

- ACM Multimedia 2023 Tutorial Talk: *Unleashing the Power of AutoML on Multimedia doi:10.1145/3581783.3613858*

Peer Review & Academic Service

- **Program Committee Member**: (i) IEEE International Conference on Data Mining (ICDM) 2022, 2023, 2024 (ii) SIAM International Conference on Data Mining (SDM) 2022, 2023, 2024 (iii) AutoML Conference 2023, 2024 (iv) ACM/IEEE ASONAM 2023, 2024 (v) Annual AAAI Conference on Artificial Intelligence 2024, 2025 (vi) IEEE VIS 2021, 2022 (vii) PAKDD 2024 (viii) ACM WSDM 2024 (Demo Track) (ix) ACM SigKDD 2025 (Applied Data Science) (x) IEEE AIxSET 2024
- **Journal Review**: (i) ACM Transactions on Intelligent Systems and Technology (ii) IEEE Transactions on Big Data (iii) Information Visualization (iv) Data & Knowledge Engineering
- Guest Editor for MDPI Applied Sciences Special Issue:*AutoML: Advances and Applications*

Graduate Courses Attended

• Probability and Distribution Theory • Statistical Inference • Data Analytics I • Data Analytics II • Deep Learning • Advanced Topics in Data and Information (Deep Learning) I • Advanced Topics in Data and Information (Deep Learning) II • Theory of Algorithms • Urban Computing • Professionalism and Ethics in Data Science • User Interface Design