Arnav Ahuja

aa5790@columbia.edu — linkedin.com/in/arnav-ahuja/ — arnavahuja.github.io/ — (716) 994-1303

Education

- **Columbia University** *New York, US*
  Master of Science in Data Science *Aug 2025 - Dec 2026*
  Relevant Coursework: Deep Learning, Reinforcement Learning, Financial Analysis, Algorithms
- **Birla Institute Of Technology and Science** *Pilani, IN*
  Bachelor of Technology Computer Science *Aug 2018 - Jul 2023*
  Master of Science in Mathematics
  Relevant Coursework: Probability, Statistics, Optimization, Operations Research, Linear Algebra, Graphs

Technical Skills

- **Programming Languages** Python, C++, Shell Scripting, SQL, Java, R, MATLAB
- **Cloud Tools** Lambda, S3, IAM, Stepfunctions, EC2, Glue, VPC, DynamoDB, Sagemaker, Cloudwatch, ECR
- **Data Science Libraries** Boto3, PyTorch, TensorFlow, Keras, Pandas, Numpy, OpenCV, Scikit-Learn
- **Machine Learning/Deep Learning** Regression, SVM, PCA, LLMs, CNNs, RNNs, Reinforcement Learning
- **Platforms/Tools** Jenkins, Gitlab, Github, Google Colab, Jupyter Notebook, MATLAB, PowerBI, ELK/Elastic Stack

Work Experience

Barclays | **Data Scientist** *Aug 2023 - Apr 2025*

- Deployed ETL jobs using AWS Glue for consumer credit risk data, supporting 50+ quantitative models and enabling 10+ teams to deliver faster, more accurate risk insights for regulatory and business reporting
- Designed an automated dataset delivery service with Couldwatch, DynamoDB & Lambda, saving time by 80%
- Enhanced a scalable microservices based infrastructure deployed on EC2 in a master slave format to integrate and run 100+ quantitative models, decreasing setup time from 1 week to 1 day
- Secured 1st place at Barclays Global Generative AI Hackathon for a GenAI solution predicting money laundering risk from company financials, enabling faster fraud decision-making

Western Australia Department of Health | **Data Science Researcher** *Jan 2023 - May 2023*

- Identified critical community health drivers from ∼19,000 social attributes across 373 suburbs, enabling data-driven interventions and optimizing 20+ health initiatives
- Used PCA to reduce dimensionality by 95% and fit a Gamma distribution, simplifying analysis of key health drivers
- Clustered 373 suburbs to identify 3 representatives as pilot sites, guiding data-driven health policy for 3M residents
- Authored a thesis identifying 5 critical insights from detailed analyses, driving evidence-based policy decisions

Amazon | **Applied Scientist Intern** *Jun 2022 - Dec 2022*

- Built domain-agnostic models for user action detection in e-commerce sites, enabling scalable product data extraction
- Engineered reinforcement learning based webpage navigation agents to identify user action elements on web pages, automating manual labeling and cutting costs caused by external agency labelling by 50%
- Designed webpage segmentation algo with 84.2% accuracy in user action prediction, boosting crawl completeness
- Leveraged graph machine learning to automate product mapping across 100k+ competitor sites, expanding coverage

Publications

Speculative Actions: A Lossless Framework for Faster Agentic Systems
(*ICLR '2026*) *International Conference on Learning Representations*

- Reduced response latency by 30% for Agentic Systems and accelerated decision making through speculative actions

Use of spatio-temporal features for earthquake forecasting of imbalanced data
(*IEEE*) *International Conference on Intelligent Innovations in Engineering and Technology*

- Developed spatiotemporal ML models that achieved 94% accuracy on imbalanced seismic data across global regions

Disease Identification in Tomato Leaf using pre-trained ResNet and Deformable Inception
(*Springer*) *5th International Conference on Computational Intelligence in Data Science*

- Devised a novel Inception-ResNet model to identify 15 diseases in plant leaves with 98.16% accuracy on 50k images

Research Experience

Agent Safety *May 2025 - Present*

- Engineered safeguards that transformed AI agent decision-making, preventing destructive behaviors while driving safer and more reliable automation
- Created a dataset of failure scenarios using OSWorld to identify agent-induced harms across multiple categories