ROBI BHATTACHARJEE

28 Quenstedtstrasse
Tübingen 72076, Germany

CONTACT
robibhatt@gmail.com
robibhatt.github.io
github.com/robibhatt

EDUCATION

UC San Diego: Ph.D. Computer Science 2023. Advisors: Kamalika Chaudhuri, Sanjoy Dasgupta
MIT: B.S. Mathematics, 2016

EXPERIENCE

Postdoctoral Researcher, Univ. of Tübingen 2023–Present
Explainability, feature selection. 3 first-author publications, 2 preprints (NeurIPS, COLT)

Ph.D. Researcher, UC San Diego 2018–2023
Robustness, distribution shift, online clustering. 7+ first-author papers (ICML, ALT, NeurIPS).

Full Time Assistant Trader, Five Rings Capital 2016–2017
Developed python code to simulate trading strategies over past market data, and applied this to test novel strategies.

Software Engineering Intern, Google Summer 2014
Implemented sampling and imbalance correction for fraud detection.

Research Intern, Jane Street Capital Summers 2013, 2015
Developed and implemented analytical tools for adding corrective terms to correlations in data.

SELECT PUBLICATIONS

Bhattacharjee, Frohnapfel, von Luxburg. *COLT 2025.* "Safely Discarding Features via SHAP."
We analyzed the soundness of a widely used feature-importance heuristic based on the SHAP explainability method. We then leveraged this analysis to develop an extremely simple data-preprocessing step that enables provable guarantees for feature selection.

Bhattacharjee, von Luxburg. *NeurIPS 2024.* "Auditing Local Explanations is Hard."
In sensitive contexts, providers of machine learning algorithms are often required to give explanations for their algorithms' decisions (i.e., explanations for why an applicant was rejected for a bank loan). However, explanations made solely by the model provider could themselves be misleading or fraudulent. Our results show that an auditor with access solely to (a) model predictions and (b) provided explanations for those predictions requires too much data to confidently audit. In particular, collectives of users, for example, coordinated by non-governmental organizations (NGOs), are never in the position to audit explanations, and a third-party auditor with more access (e.g. to model architecture or weights) is required.

Bhattacharjee, Dasgupta, Chaudhuri. *ICML 2023.* "Data-Copying in Generative Models: A Formal Framework"
We provided an algorithm with provable guarantees for detecting data-copying in continuous spaces, where a generative model strongly over-represents instances that are geometrically close to training data. I subsequently supervised a masters thesis that demonstrated that this algorithm could scale to higher dimensional models when combined with random projections.

SKILLS

Machine learning engineering: Implemented pre-training (tokenizer + transformer) from scratch (followed Stanford's CS336 class independently)
Proficient with vLLM, TRL, pytorch: As part of ongoing work, implemented fine-tuning on Qwen2.5-3B to mitigate hallucinations over the TriviaQA dataset.

AWARDS

Honorable Mention, Putnam (2012, 2013)
USAMO (2009–2012, qualified for MOP 2010, and placed in the top 30 in 2012.)