NJIT | INSTITUTE FOR DATA SCIENCE

Data Science Seminar Series

In Collaboration with
The Department of Data Science

Towards Secure and Safe AI-enabled Systems Through Optimizations

Hosted by Hai Pha

Guanhong Tao

Purdue University

Date:	**Monday, March 18, 2024**
Time:	**2:30 PM - 3:30 PM (Coffee served at 2:15 PM)**
Location:	**GITC Building Room 3600 (3rd floor Seminar Room)**
Web Link:	**Zoom Meeting Link**

With the widespread integration of Artificial Intelligence (AI) in various sectors, the security and safety of AI-enabled systems have not yet been fully ensured. Just like conventional systems having software bugs or errors, applications leveraging AI are not free of bugs. In this talk, I will present an optimization-based framework for identifying and mitigating backdoor vulnerabilities in machine learning models. My talk will cover novel optimization techniques that more efficiently and effectively detect backdoors in both white-box and black-box settings, achieving substantial improvement in performance. My work contributed to the Purdue team securing the first place in IARPA TrojAI Trojan Detection Competition (Rounds 1-4). I will share insights on the essence of backdoors and their presence in naturally pre-trained models. I will also introduce the first hardening framework for mitigating backdoor vulnerabilities. Finally, I will conclude with an outlook on securing emerging AI techniques, such as generative AI (GenAI), and the evolving ecosystem enabled by GenAI.

Guanhong Tao is a Ph.D. candidate at Purdue University, advised by Prof. Xiangyu Zhang. His research focuses on security and safety of AI-enabled systems. He pinpoints realistic vulnerabilities in real-world AI systems and builds practical solutions to mitigating identified vulnerabilities and problems using optimizations. He has led the Purdue team to secure the first place in IARPA TrojAI Trojan Detection Competition (Rounds 1-4). He is the recipient of Maurice H. Halstead Memorial Research Award in Purdue, Best Paper Award in ECCV 2022 AROW Workshop, and CSAW 2021 Best Applied Security Paper Award TOP-10 Finalists. His work has been published at top-tier security (S&P, USENIX Security, CCS, NDSS), machine learning (NeurIPS, ICML, ICLR), and software engineering (ICSE, FSE) conferences. https://www.cs.purdue.edu/homes/taog/