CAREER: Trustworthy Machine Learning from Untrusted Models

Algorithmic Learning Privacy Security Lab

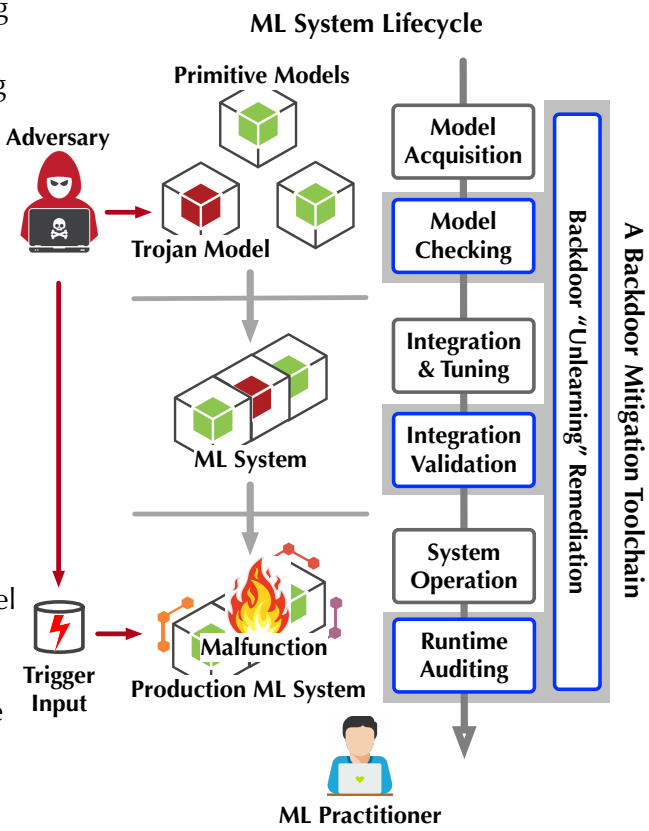
PennState

Challensge:

- Ever-increasing complexity and training cost of machine learning systems
- Paradigm shift from building everything from scratch to reusing existing primitive models as building blocks
- Third-party models as attack vectors to carry malicious functions ("backdoors")
- Obscure to reason about a model's malicious behavior due to high dimensionality and non-linearity

Solution:

- "Lifelong security" throughout the lifecycles of machine learning systems: offline checking + runtime monitoring
- A backdoor mitigation toolchain: model checking, integration validation, system auditing, backdoor fixing
- Ensemble defense strategies against the adversary's adaptive countermeasures



Scientific Impact:

- Explore backdoor threats in new domains (graphs, natural languages)
- Develop new theories to characterize data-model dynamics in backdoor attacks
- Develop mitigation solutions with provable guarantees
- Open-source a comprehensive backdoor testbed[1] to facilitate future research
- Disseminate outcomes through publications, invited talks, and media coverage (thus far 19 papers in top-tier venues, 2 best paper awards)

Broader Impact and Participation:

- Train grad and undergrad students in the research frontier of AI security
- Communicate findings to industry
- Integrate research results into newly developed courses "Adversarial Machine Learning" and "Machine Learning Security"
- Develop STEM education modules in Lehigh' and Penn State's outreach programs

[1]TrojanZoo: https://github.com/ain-soph/trojanzoo

Award #1953893
PI: Ting Wang (ting@psu.edu)
Pennsylvania State University