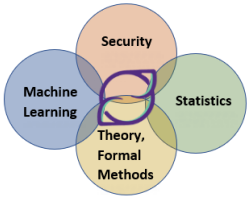
The Center for Trustworthy Machine Learning

SaTC: CORE: Frontier: Collaborative:
End-to-end Trustworthiness of Machine-Learning Systems



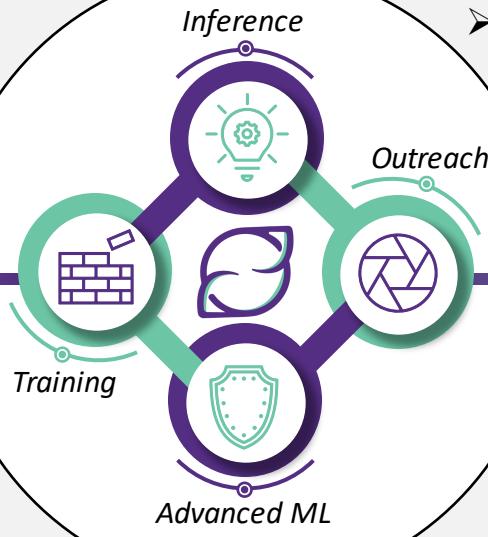
Challenge:

Machine learning, a disruptive force in many domains, is **vulnerable across its lifecycle**:

- Fragmented understanding of threat space
- Attacks cripple security-critical domains
- Adaptive adversaries sidestep defenses
- Synthetic reality threatens society
- Lack of strong provable guarantees

Scientific Impact:

- Develop cross-disciplinary science to provide a basis for trustworthy ML systems
- Design new algorithms & systems that provide provable robustness guarantees
- Foster and grow community of researchers



Solution:

Three parallel & interconnected thrusts:

I. **Inference Time Robustness**: methods to defend models from *adversarial inputs*

II. **Training Time Robustness**: measures of robustness to attacks that *corrupt data* & are robust to *manipulation*

III. **Security Implications of Advanced ML**: explore abuse, detection, & defenses against *generative* ML models

Broader Impacts and Participation:

- Raise awareness of security, privacy, and fairness impacts of ML
- Organize summer camps for students & teachers for early engagement
- Generated policy workshops:
  I. *Cybersecurity and Machine Learning Vision Document:* NSF & DFG, 2021
  II. *Artificial Intelligence and Cybersecurity: Opportunities and Challenges:* NITRD, 2020
- More than 50 publications and 65 keynotes & distinguished lectures

Award# 1805310, Pennsylvania State University, University of California Berkeley, University of California San Diego, Stanford University, University of Wisconsin, and University of Virginia
Contact: Amy Hasan (alh31@psu.edu) Webpage: www.ctml.psu.edu