The Center for Trustworthy Machine Learning

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

www.ctml.psu.edu

Patrick McDaniel, **Amy Hasan** Penn State; **Dan Boneh**, **Percy** Liang, Stanford; **Kamalika Chaudhuri**, UC San Diego; **Somesh Jha**, U of Wisconsin; **David Evans**, U of Virginia; **Jacob Steinhardt**, UC Berkeley

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

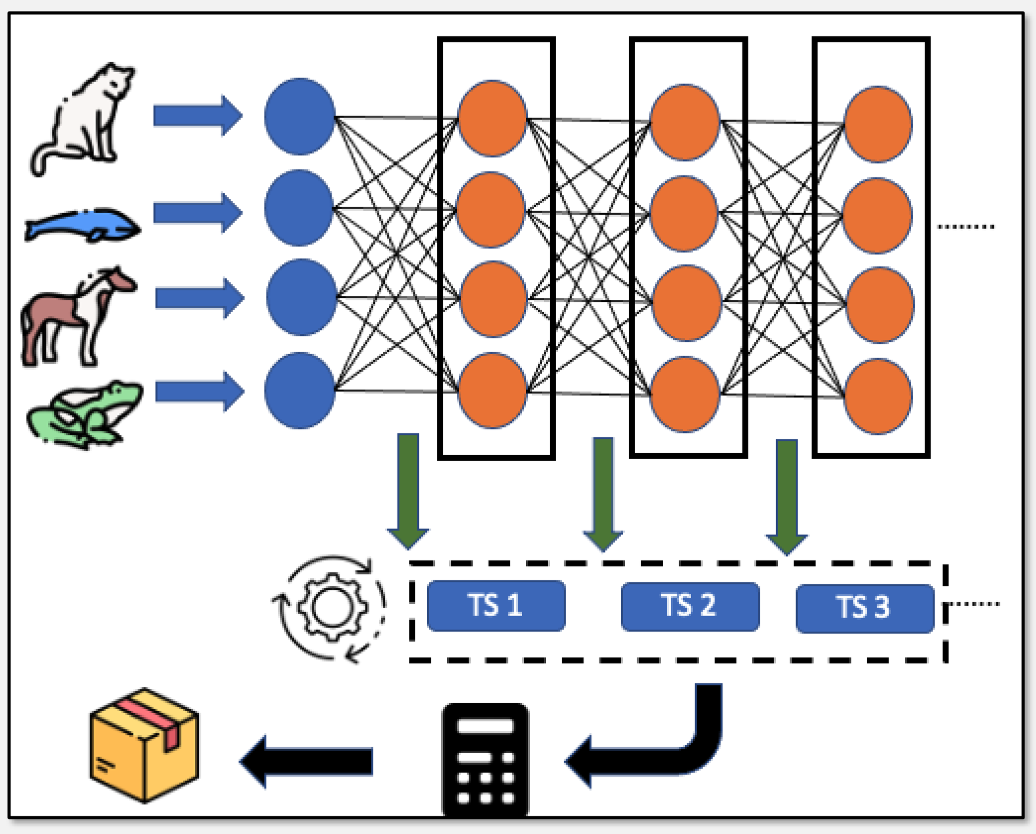
Scientific Impact:

- 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

Advanced ML · *Inference* · *Training* · *Outreach*

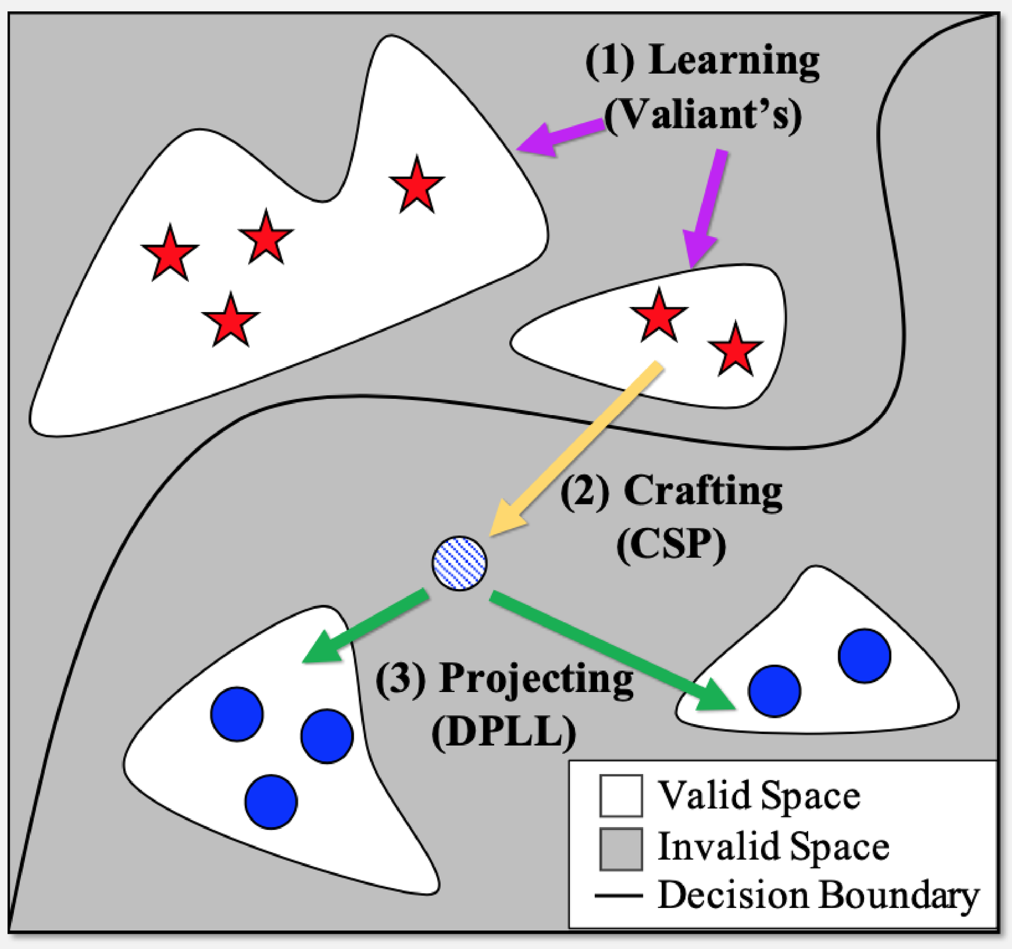
Solutions/Key Innovations:

A General Framework for Detecting Anomalous Inputs to DNN Classifiers



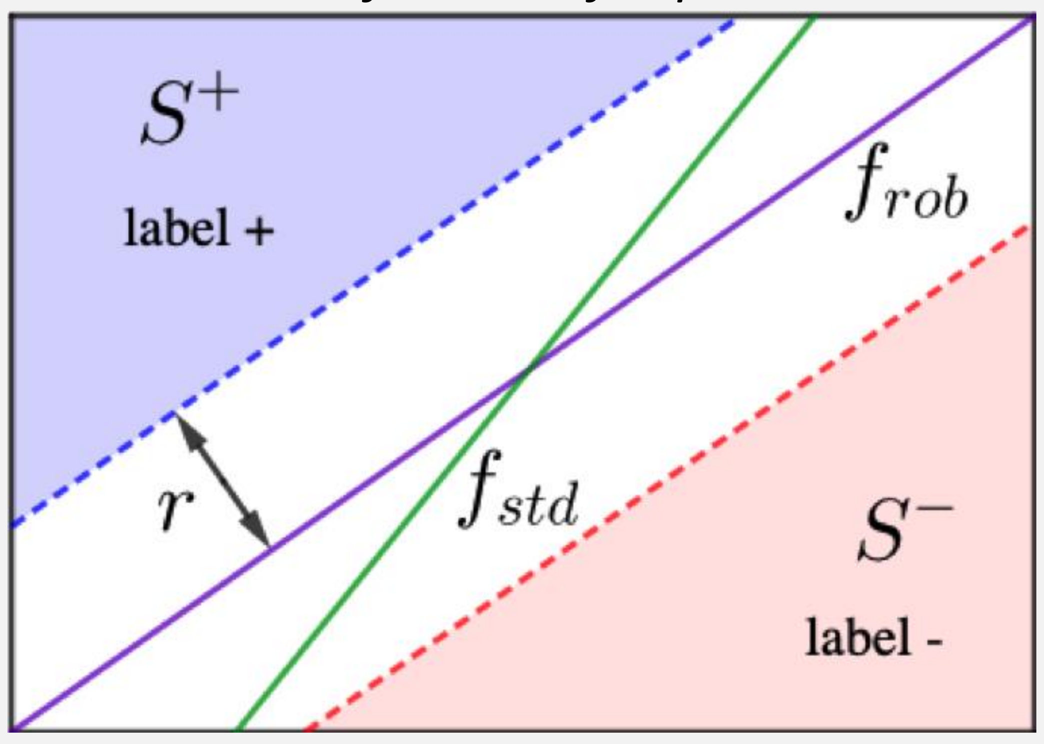
- Anomalous inputs (either out-of-distribution or adversarial examples) can cause abnormal model behavior and should be detected at inference.
- We propose an unsupervised framework that exploits layer information with modular statistical tests.
- Unlike prior approaches, our framework scales, does not require hyperparameter tuning, can compute arbitrary test statistics, among other advantages.
- We need modular anomaly detection frameworks to be quickly adaptable to the domain at hand.

On the Robustness of Domain Constraints



- Network flows and binaries have *domain constraints:* complex feature relationships that must be obeyed for an attack to be representative of the domain.
- Up to 80% of adversarial examples produced by state-of-the-art attacks violated domain constraints.
- Enforcing domain constraints on invalid adversarial examples restored up to 34% of model accuracy.
- We need to incorporate domain constraints for realistic threat modeling of adversaries in ML.

Sample Complexity of Adversarially Robust Linear Classification of Separated Data



- Prior work suggests that additional data may be necessary for robust classifiers (when classes overlap)
- We explore when data is well-separated and a perfectly robust and accurate linear classifier exists.
- When the data is linearly separated with margin *r*, finding an *r*-robust classifier needs $\Omega\left(\frac{1}{n}\right)$ samples, while an accurate classifier needs only $O\left(\frac{1}{n}\right)$.
- However, for separation greater than *r*, then only $O\left(\frac{1}{n}\right)$ is sufficient to find an *r*-robust classifier.
- This shows for well-separated data, finding robust models is not only possible, but also tractable.

Machine Learning · *Theory* · *Security* · *Statistics*

Broader Societal Impacts:

Outreach to Government

- *Cybersecurity and Machine Learning Vision Document* – NSF & DFG, 2021
- *Artificial Intelligence and Cybersecurity: Opportunities and Challenges* – NITRD, 2020

Public Policy Briefings

- *Preparing for the Age of Deepfakes and Disinformation,* 2020

Broader Impact on Education:

REUs

- 5 at PSU (4 female)
- 8 at UVA (2 female)
- 1 female at UCSD & Stanford

Advancing Cybersecurity in K-12 Education

- Week-long training for middle/high school teachers (2019)

2023 IEEE Conference on Secure and Trustworthy Machine Learning

Broadening Participation:

Girls Who Code

- Targets middle/high school girls (2021, 2022)

AI4ALL

- Targets underrepresented highschoolers in San Francisco (2020, 2021)

Summer Camps

- On-going (all universities); targets females and underrepresented kids