# Introduction

- Current ML systems are **brittle** and fail even on small shifts such as imperceptible changes to images and typos in text
  - Not ready for real-world deployment
  - Do not really perform the underlying task
- I work on improving reliability with broadly the following themes
  - Certified robustness (evaluation)
  - Unlabeled data to improve robustness (training)
  - Understanding deep learning via robustness

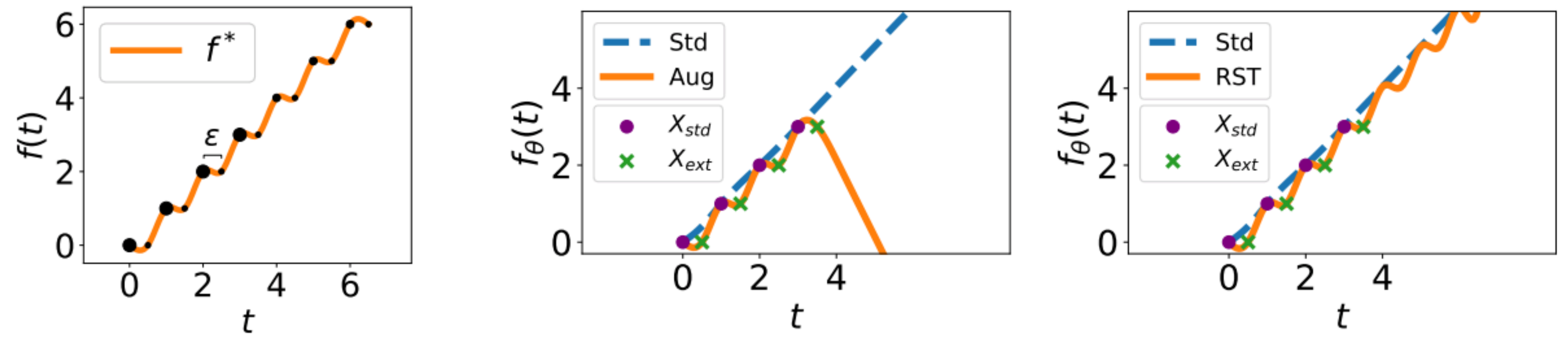
# Certified Robustness


- The predominant paradigm of evaluation in ML is empirical
- Insufficient in the face of malicious users or attackers
- Consider adversarial examples
  - Highlight is the arms race—defenses get broken by stronger attacks
- We train certifiably robust networks: with guarantees against any attack
  - Convex relaxations to reason about the activations of a neural network
  - Semidefinite programming leads to tight verification in general
  - Develop scalable first-order methods suitable for neural networks
- Discrete bottlenecks can be used to reduce attack space in NLP
  - Obtain SOTA heuristic and certified robustness against typos


**Email:** aditir@stanford.edu **Website:** https://stanford.edu/~aditir/


# Unlabeled data for robustness

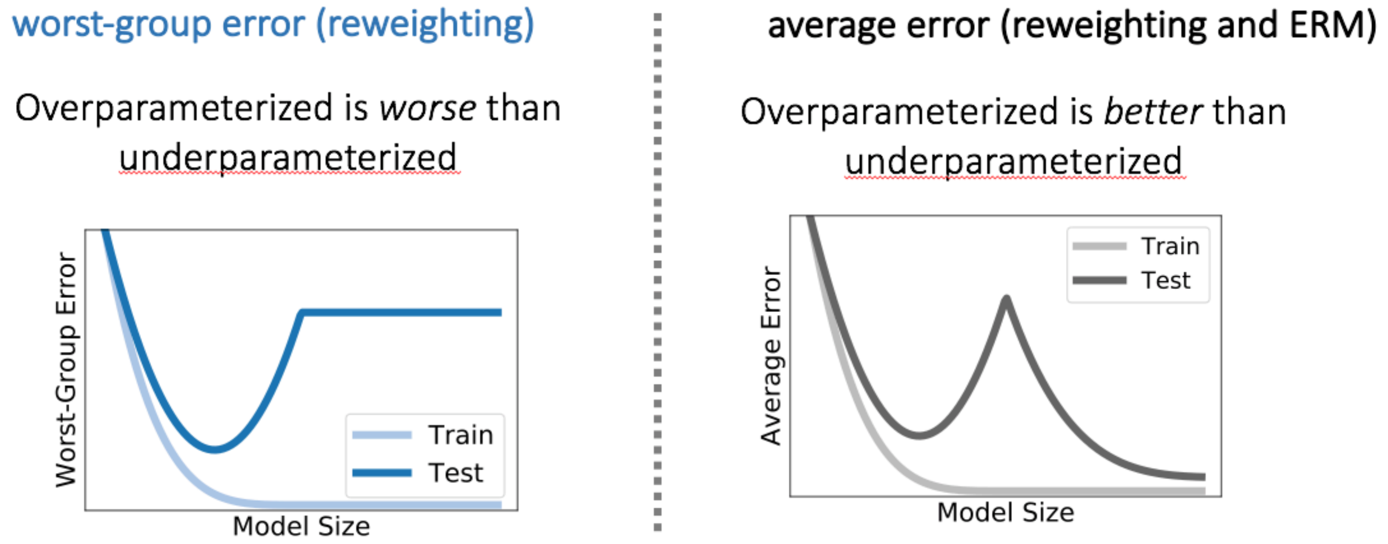
- Robust training does not obtain very high robust accuracy against adversarial examples
  - The labeled data that we have could be insufficient
- We show theoretically and empirically, we only need unlabeled data
  - Method: Robust self-training (RST)
    - Step one: Train a classifier on labeled data (accurate)
    - Step two: Generate pseudolabels on unlabeled data
    - Step three: Robust training on labeled + pseudolabeled data
  - Obtains state-of-the-art robustness
- Robust training typically decreases standard accuracy




- We show that RST with unlabeled data mitigates the tradeoff

# Surprises in robustness

- Overparameterization exacerbates spurious correlations and magnifies disparities in groups




- Neural networks are highly biased towards simple features at the expense of very small margin
  - Contrary to the max-margin insight from linear analysis
- Data augmentation with correct labels can worsen generalization