Title: “Robust ML: Attacks and Defenses”

Objective: Crash course on robustifying machine learning in the presence of adversaries.

Outcome: Become up to date with the rapidly expanding literature on robust ML.

Target audience: Graduate students, postdocs, and faculty working in ML, data analytics, etc.

Prereqs:

- Graduate experience with ML, DL (EE/CPRE529, ME592, or equivalent).
- We will assume everyone in the audience is familiar with the basics.

Tentative topics:

- Overview of adversarial ML (applications, attacks/defense models)
- White-box attacks and black-box attacks
- Robustifying Reinforcement Learning
- Edge cases / certificates of robustness
- Defense algorithms (optimization, learning-based approaches)
- Data poisoning / train-time attacks (+ defenses)

Timetable

Session	Topic	Date	Room	Time
Week 1	Intro to adversarial ML (attack/defense models)	Aug 24 (Fri)	Coover 3043	12-1pm
Week 2	White box vs black box attacks	Aug 31 (Fri)	Black 2004	12-1pm
Week 3	Attacks on RL agents	Sept 7 (Fri)	Black 2004	12-1pm
Week 4	Edge (“minimal distortion”) cases + defense	Sept 14 (Fri)	Black 2004	12-1pm
Week 5	Defense against the dark arts	Sept 26 (Wed)	Black 2004	12-1pm
Week 6	Data poisoning	Oct 9 (Tue)	Black 2004	12-1pm