Centre for Machine Intelligence & Data Science
C-MInDS • IIT Bombay

शैक्षिक प्रसार
Educational Outreach

NPTEL+
Anyone Anywhere Anytime

Reliable Machine Learning

18-19 July 2026 (Sat–Sun)

11:00 AM - 04:00 PM

Online (Synchronous)



Introduction to the Course

Progress in machine learning is often measured under controlled, well-understood conditions. However, safety-critical workflows in realistic settings require ML systems to be reliable even when faced with new and unexpected conditions. Sufficiently adverse conditions may violate the statistical assumptions underlying common ML models, causing undesirable behavior. This undesirable behavior may manifest in different ways such as: i) lack of robustness to unseen and adversarial inputs, ii) leakage of private data or model parameters, and iii) uneven performance across subpopulations. This course will equip students to rigorously reason about the conditions under which unreliable behavior occurs and use these insights to build reliable ML systems.

Course Topics

- Anomaly detection
- Poisoning Attacks
- Poisoning Defenses
- Evasion Attacks
- Evasion Defenses
- Fundamental limits on robust learning
- Privacy Attacks
- Privacy Defenses (Time permitting)

Who May Benefit

- Researchers
- Students
- Faculty members from other institutes
- Industry and corporate professionals

Hands-on component

Jupyter notebook based guided exercises in Adversarial example construction and adversarial training methods (1 hour)

Instructor

Prof. Arjun Bhagoji
IIT Bombay

Eligibility

- Required - Introductory course in machine learning
- Desirable - Working knowledge of probability, optimization and linear algebra

Fees

Students & Postdocs:
₹ 500 +18% GST = ₹ 590

Faculty & Industry professionals:
₹ 1000 +18% GST = ₹ 1180

Scan to register