Alon Rosen

Introduction

We are now no longer the Knights Who Say Ni![*]

The intersection of machine learning and cryptography is uncharted territory. Both fields have had extraordinary practical success, but they arrived there by rather different routes.

Machine learning has changed what computers can do. Much of this progress has been empirical, driven through benchmarks and at scale. We lack a satisfying theory of why a method works, and even a rigorous definition capturing everything we mean by "works." Yet success is difficult to dispute when models exhibit superhuman capability in a wide array of tasks.

Cryptography's success has followed a route in which theory has been indispensable. Definitions specify the adversary and what counts as a break, assumptions state what is believed to be hard, and proofs establish the connection between assumptions and the security of constructions.

Modern AI systems span an extraordinary hierarchy of abstraction, from atoms to hardware, through neural-networks, to behavior and societal effects. Cryptography occupies only a narrow band in this hierarchy, yet its implications reach both up and down.

Putting the two fields side by side leads to fascinating questions. Some connections go back to the foundations of learning theory; others arise only because modern models exhibit behaviors and structure for which there was no earlier theory. The course moves across this spectrum.

A classical connection between the fields. We begin with pseudorandom functions, efficiently computable objects that are hard to learn. In the other direction, hard learning problems such as LPN and LWE permeate advanced cryptography. We then move from LWE to its continuous variants, CLWE and hCLWE, which will reappear later. This connection also gives us a first way to contrast the two fields and see their different methodologies and perspectives at work.

Adversarial examples. We next turn to adversarial robustness. Here we have two dual attacks: sensitivity, where a small perturbation causes a large change in output, and invariance, where a substantial change in input leaves the output essentially unchanged. Understanding how an attack works can point the way toward defenses.

Old cryptography, new applications. We then turn to new applications of familiar cryptography. Backdoors bring them to good old single- and multilayer perceptrons, where hiding must contend with weight distributions produced by training rather than designed for hiding. Hardness of score estimation gives these continuous assumptions a new application and a chance to look more closely at how diffusion models operate. Watermarking gives an opportunity to see cryptographic ideas at work in both language and diffusion models.

Model extraction. Model extraction asks how much of a model can be reconstructed from its input–output behavior, especially when probabilities or logits are given as output.

[*] Monty Python, *Monty Python and the Holy Grail*.

Cryptography for ML and verifiability. We then use cryptography to protect and verify computation with machine-learning models. The challenge is to obtain meaningful guarantees without losing the efficiency and structure that make modern models usable.

Theory from phenomena. We will also study low logit rank and sketching, where empirical structure in modern learning systems suggests new theory and efficient algorithms.

Untrusted models. Finally, we consider a special adversary: the model itself. How should one test a model that can strategically underperform or manipulate the evaluation? Work on sandbagging, noise injection, AI control, and computational no-coincidence principles pushes on a basic question: what can observed behavior reveal when the system producing it may be actively trying to mislead us? This opens questions about computation that barely arise in classical settings.

Recurring themes. Three questions will recur throughout the course:

- What exactly is being sought, and what is being guaranteed?
- What assumptions does the guarantee rest on? Can we prove it rigorously?
- Can strong cryptography be relevant at scale? What meaningful relaxations are possible?

New questions will also emerge from empirical observations, and from trying to get a theoretical grip on this astounding technology, both to protect it and to protect ourselves from it.

Uncharted territory, interesting times.