Introduction

Cryptography	**Machine learning**
Theory first	Empirical
Adversarial	Optimistic
Maximalistic	Pragmatic

Two successful fields, different methodologies

- Machine learning: empirical success, driven by benchmarks and scale.
- Why methods work, what "works" even means?
- Cryptography: grounded in definitions, assumptions, and proofs.
- Specified adversary, what constitutes a break, and hardness assumptions.

Layers of abstraction in neural-network systems

From the physical substrate to societal consequences:

- Atoms and physical materials
- Transistors and electronic devices
- Logic gates, memory, and arithmetic circuits
- Processors, accelerators, and communication hardware
- Operating systems and software runtimes
- Tensor operations and computational graphs
- Neural-network layers and architectures
- Training data, objectives, and optimization
- Learned weights, features, and representations
- Model predictions and generated outputs
- Model behavior and interactions with people
- Products, institutions, and societal effects

Cryptography: a narrow band with implications across the entire hierarchy.

A classical connection: learning and pseudorandomness

- PRFs: efficient to compute, hard to learn.
- LPN and LWE: hard learning problems as cryptographic foundations.
- CLWE and hCLWE: continuous variants, revisited later.

Adversarial examples: two dual attacks

- Sensitivity: small change in input, large change in output.
- Invariance: large change in input, little change in output.
- Understanding attacks as a guide to defenses.

Old cryptography, new applications

- Backdoors even in single- and multilayer perceptrons.
- Hiding within weight distributions produced by training.
- Hardness of score estimation and diffusion models.
- Watermarking in language and diffusion models.

Model extraction

- How much of a model is revealed by input–output behavior?
- What more is revealed by probabilities or logits?

Cryptography for ML and verifiability

- Protecting and verifying ML computations.
- Meaningful guarantees at practical scale.

Theory from empirical phenomena

- Low logit rank and sketching: empirical structure suggests algorithms.
- New theoretical questions from observed phenomena.

Untrusted models

- The model itself as the adversary.
- Sandbagging, noise injection, manipulation of tests.
- AI control and computational no-coincidence principles.
- What can observed behavior reveal from a strategic model?

Three recurring questions

1. What is sought, and what is guaranteed?
2. Which assumptions? Which rigorous guarantees?
3. Cryptography at scale: which relaxations are meaningful?

New phenomena, new questions: understanding AI, protecting it, protecting ourselves.

. . . what to my taste is characteristic for all intelligent thinking . . . is, that one is willing to study in depth an aspect of one's subject matter in isolation for the sake of its own consistency, all the time knowing that one is occupying oneself only with one of the aspects . . .

— Edsger W. Dijkstra, *On the Role of Scientific Thought* (1974)