CYB-4203/6203

Secure and Trustworthy AI

Presentation 12: Recap and Midterm Exam Discussion

Monday, March 9, 2026

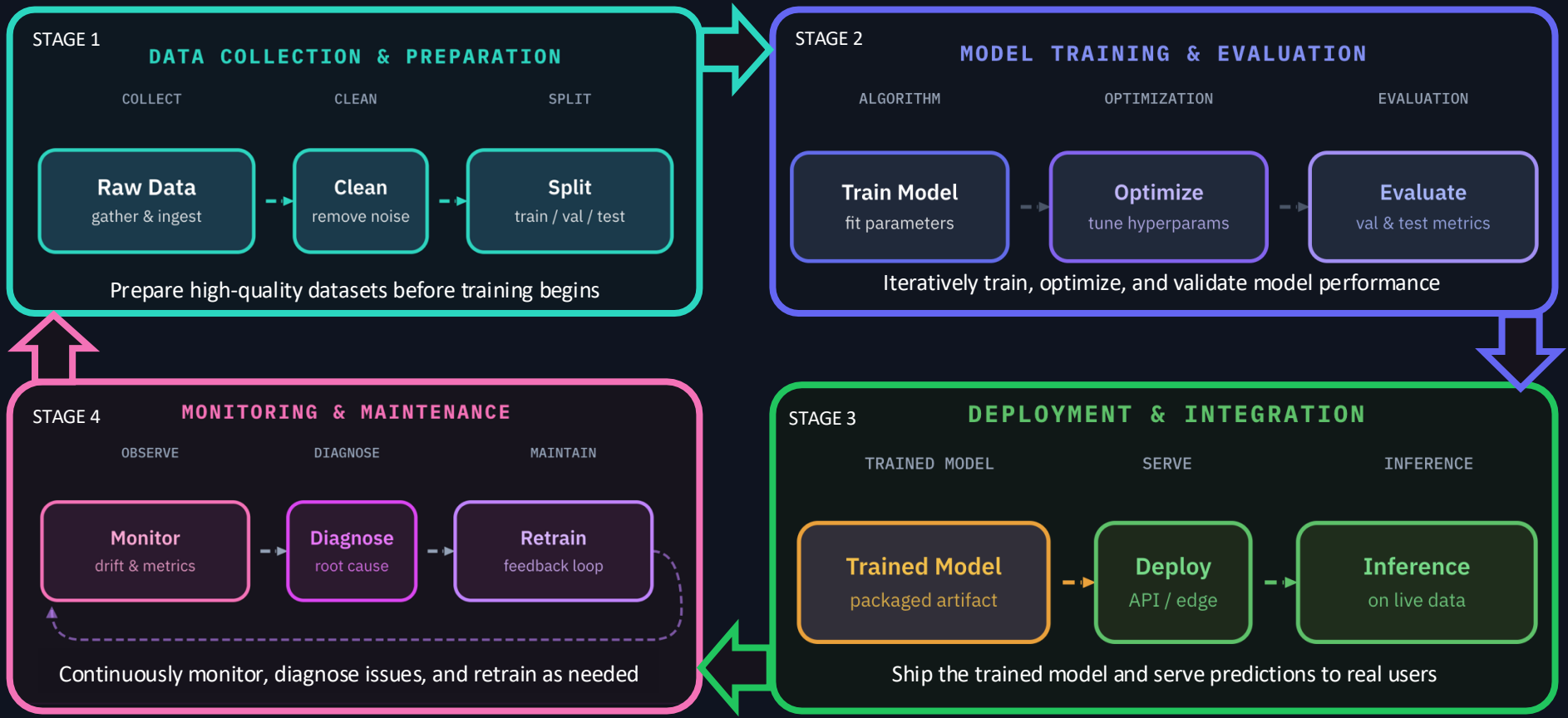
Topics: 1.1 - 6.2

Today's Agenda

- Brief report from Spain
- Interactive recap of last week's material
- Discussion:
    - Exams
    - Learning
    - Reward hacking
    - The purpose of a university education

ICISSP – Marbella, Spain – March 3-5

When (in this lifecycle) is the exam?



Which ML paradigm represents the University education model?

Whether (1) supervised, (2) unsupervised, or (3) reinforcement learning:

- **Training Phase**: The model learns from training data inputs how to make good predictions
- **Deployment phase**: The trained model performs inference (makes predictions) on live data.

What is the attack surface of your mind?

- **Attack surface**: the total sum of all potential entry points or ‘attack vectors’ including digital, physical, and social vulnerabilities that a threat actor can exploit.
- **Threat model**: proactive, structured process used to identify, quantify, and address security vulnerabilities by anticipating potential attacker methods.

Image generated by ChatGPT; slop retained for ironic value.

Natural Learning Lifecycle Stage 2: Model Training & Evaluation

What's supposed to happen in this stage?

- **Architecture selection** and hyperparameter tuning for the task
- **Training runs** on GPUs/TPUs – iteratively adjusting model weights
- **Evaluation** on benchmarks to measure accuracy, fairness, robustness
- **Transfer learning**: fine-tuning pre-trained models for specific tasks

Model weights are learned parameters – often the most valuable IP in the pipeline

Natural Learning Evaluation: Exams

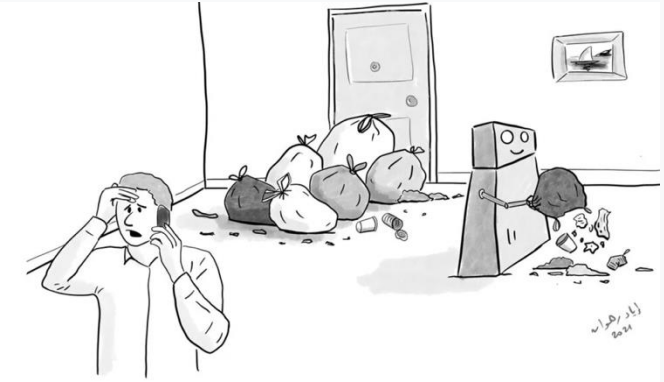
What do exams do?

- Measure our retention
- Part of the learning process
  - Force us to replay learning
  - Force us to PAY ATTENTION

What is the University ‘reward signal’?

Vulnerability: RL Reward Hacking / Specification Gaming

- RL systems find shortcuts that maximize the reward signal without achieving the goal
- Alignment failure, not adversarial attack, but the behavior is exploitable
- Reward hacking can generalize across tasks;
  A model that learns hack one type of rewards is more likely to hack others (AIA, 2024)
- Frontier LLMs increasingly exhibit this behavior (METR, 2025)



“As soon as it’s done cleaning the house, it brings in trash from the street, and starts all over again!”

https://www.evilaicartoons.com/archive/design-good-carrots-and-sticks

Defenses: reward model ensembles, process-based rewards, constrained optimization

Natural Learning Lifecycle Stage 3: Deployment & Integration

What stage of human life does this represent?

- **Serving** trained models via APIs, batch processing, or edge deployment
- **Model optimization**: compression, quantization, distillation for production
- **Integration** with applications, databases, and user-facing systems

The attack surface shifts from the pipeline to the interface.

This is where adversaries interact with models directly.