HANDOUT 8 - WEEK 10 READING

Adversarial Attacks on ML Systems

CIS 350: AI for Cybersecurity - Smart Defense for the Digital Business

Colorado State University - Computer Information Systems

How to use this reading (and an ethics note)

This handout supports Week 10. We study attacks ON machine-learning models - adversarial examples, data poisoning, and privacy attacks - ONLY to detect and defend against them. Every attack is paired with a defense. The figures and code match Lab 8 (a sandboxed, defensive lab on a toy model); the review questions map to Quiz 8.

Estimated reading time: 20-25 minutes. No prior coding background needed.

1. A New Target: the Model Itself

Until now, AI was our tool (Weeks 3-7) or the attacker's tool (Week 8). Now the machine-learning model is the TARGET. Because a model learns its rules from data, it has new weak points: the input it sees at test time, the data it learns from during training, and the private data it might leak. The good news: each attack maps to a defense.

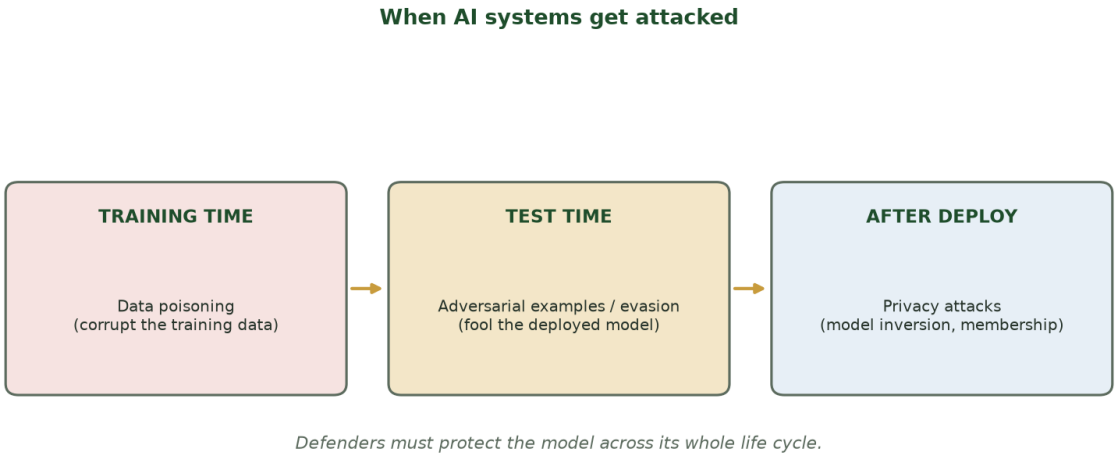

Figure 1. Three phases, three attack types: data poisoning (training), adversarial examples/evasion (test time), and privacy attacks (after deployment).

Why models are fragile

Models learn statistical patterns, not human understanding, so they can be confidently wrong. A tiny input change a human ignores can cross a decision boundary - and a model blindly trusts whatever data it was trained on.

2. Adversarial Examples & Evasion (Test Time)

An adversarial example is an input changed by a small, crafted amount that flips the model's prediction - often invisibly to a human. It pushes the input just across the decision boundary. Used against a deployed model, this is called EVASION (e.g., malware tweaked to dodge a detector, or stickers that fool a self-driving car's vision).

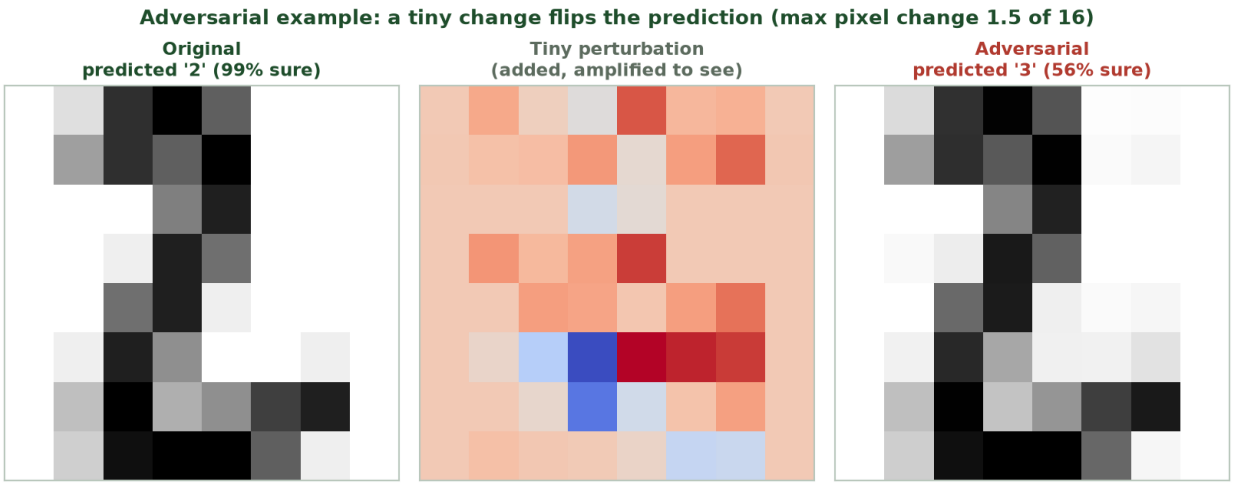

Figure 2. A '2' plus a tiny perturbation (max pixel change 1.5 of 16) is read as a '3.' The images look identical, but confidence fell from 99% to 56%.

```
# Defensive study on a TOY model (Lab 8): craft a tiny perturbation
direction = clf.coef_[target] - clf.coef_[true_class]   # sensitive direction
direction = direction / np.linalg.norm(direction)
x_adv = np.clip(x + 3.5 * direction, 0, 16)             # small nudge
print(clf.predict([x])[0], '->', clf.predict([x_adv])[0])   # e.g., 2 -> 3
```

3. Data Poisoning & Backdoors (Training Time)

Poisoning corrupts the TRAINING data so the model learns the wrong thing - a few mislabeled or malicious examples can shift the decision boundary. It is dangerous because models often train on large, scraped, or user-submitted data. A BACKDOOR is a stealthy poisoning attack that plants a hidden trigger: the model behaves normally until it sees the attacker's secret pattern, so normal testing misses it.

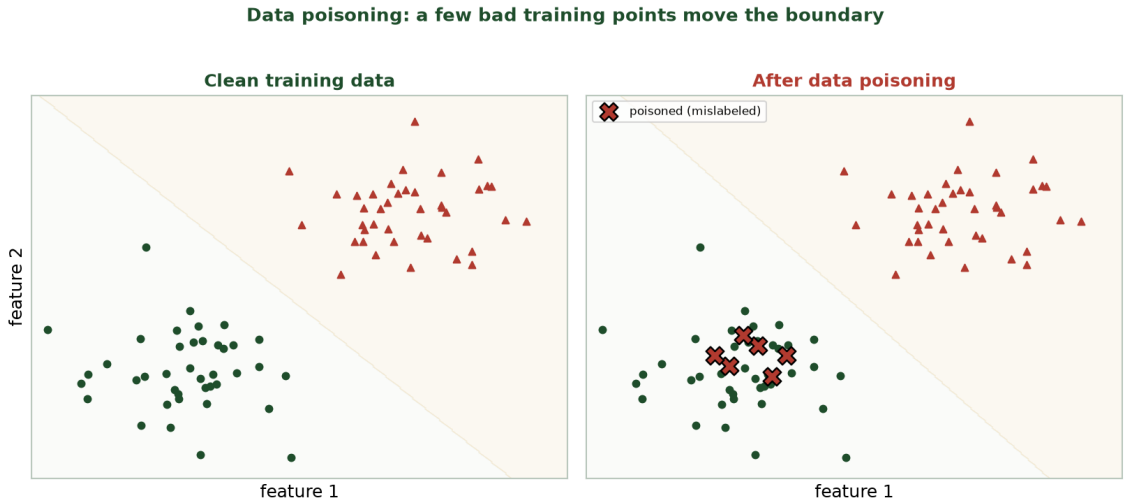

Figure 3. Injecting a handful of mislabeled points shifts the learned boundary - degrading the model while it still looks normal.

Poisoning vs. evasion

Same goal (wrong output), different timing and access. Poisoning happens at TRAINING time and needs access to the data/pipeline; evasion happens at TEST time and only needs to craft an input. Different timing means different defenses.

4. Privacy Attacks (After Deployment)

To learn, models absorb patterns from their training data - and sometimes memorize specific, sensitive examples. Attackers probe a deployed model to pull that information back out. MODEL INVERSION reconstructs training data (e.g., a face); MEMBERSHIP INFERENCE reveals whether a specific record was in the training set; MODEL EXTRACTION clones the model by querying it a lot. 'The model is trained, so the data is safe' is false.

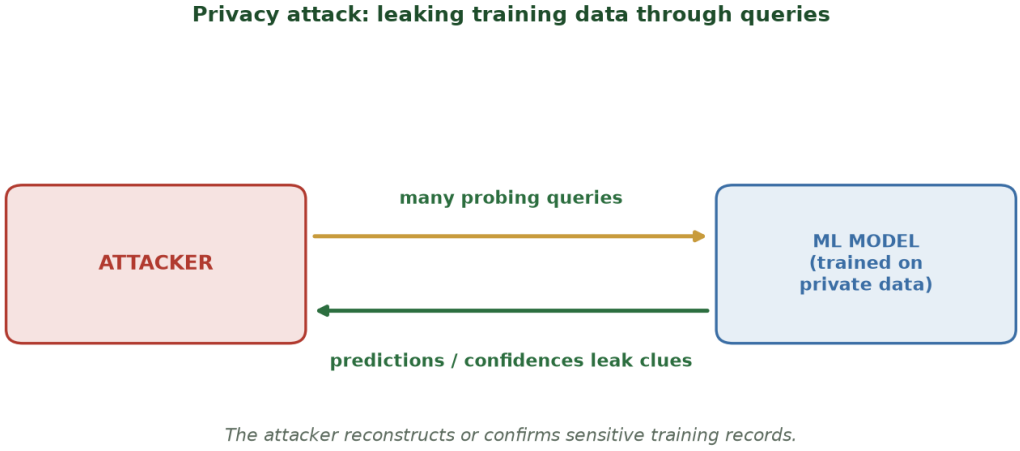

Figure 4. A privacy attack: many probing queries plus the model's confident responses leak clues about its private training data.

5. Defenses: Every Attack Has One

No single trick stops everything - defend in depth, matching the defense to the attack's timing and target. Key defenses:

- Evasion/adversarial -> adversarial training, input checks, and detecting LOW-CONFIDENCE predictions.
- Poisoning -> vet & clean training data, track data provenance, anomaly-check the training set, test for backdoors.
- Privacy -> differential privacy (add calibrated noise), limit exposed detail, rate-limit queries.
- Model theft -> throttle queries, watermark, monitor for abuse.

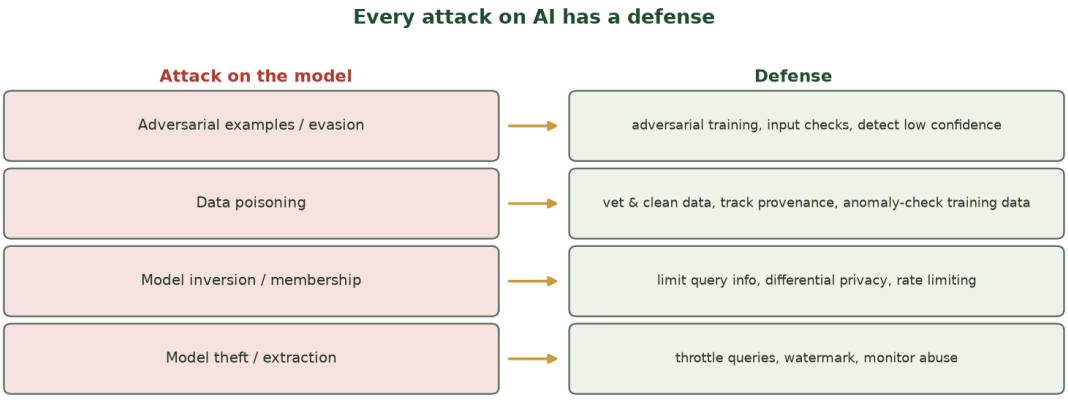

Figure 5. Each attack on AI maps to concrete, layered countermeasures across the model life cycle.

There is no perfect defense

AI security is an arms race, and robustness trades off against accuracy and speed. The realistic goal is to raise the attacker's cost, DETECT attempts, and plan to RECOVER - guided by standards like NIST AI 100-2, the NIST AI RMF, and MITRE ATLAS.

Key Terms

- Adversarial example - a small, crafted input change that flips a prediction.
- Evasion - using adversarial examples against a deployed model (test time).
- Data poisoning - corrupting the training data (training time).
- Backdoor - poisoning that plants a hidden trigger for secret misbehavior.
- Model inversion - reconstructing training data from a model.
- Membership inference - determining whether a record was in the training set.
- Model extraction / theft - cloning a model by querying it.
- Adversarial training - training on adversarial examples to build robustness.
- Differential privacy - adding calibrated noise so no single record stands out.

Review Questions (self-check for Quiz 8 and Lab 8)

- What is an adversarial example, and why does it work?
- How does data poisoning differ from evasion (when does each happen)?
- What is a backdoor attack, and why is it hard to catch?
- Name a privacy attack and what it reveals.
- Give a defense for evasion, one for poisoning, and one for privacy attacks.
- Why is there no perfect defense, and what should defenders do instead?

Remember: AI tools are not permitted on quizzes and labs; Week 10 material is for DEFENSE only.

References

- NIST AI 100-2 (2023): Adversarial Machine Learning - taxonomy and mitigations.
- I. Goodfellow et al., 'Explaining and Harnessing Adversarial Examples,' 2015.
- R. Shokri et al., 'Membership Inference Attacks Against ML Models,' 2017.
- M. Fredrikson et al., 'Model Inversion Attacks,' 2015.
- MITRE ATLAS - adversarial threat landscape for AI systems; OWASP ML Security Top 10.
- CISA/NCSC, Guidelines for Secure AI System Development, 2023.