CS 499/599: Machine Learning Security

01.03: Part I: Course Introduction

Mon/Wed 12:00 – 1:50 pm

Instructor: Sanghyun Hong

sanghyun.hong@oregonstate.edu

Oregon State University

SAIL

Secure AI Systems Lab

Notice (No exceptions)

- OSU's COVID-19 Policy
  - You **MUST** use a face covering when in indoor spaces and classrooms
    - You can be asked to leave the class if you don't wear face coverings
  - Acceptable face coverings
    - Cloth garments that cover the noise and mouths
    - Medial-grade disposable masks
    - Clear plastic shields that cover the forehead, below the chin and wraps around the sides
  - You are **ONLY** allowed to remove face coverings
    - If you are presenting and six feed away from the others
    - If you are drinking water during the lecture

Sanghyun Hong

Who am I?

- Assistant Professor of Computer Science at OSU (since Sep. 2021!)
- Ph.D. from the University of Maryland, College Park
- B.S. from Seoul National University, South Korea

What I do?

- **Formal:** I work at the intersection of security, privacy, and machine learning
- **Informal:** I "hack" machine learning, expose security threats, and defeat them

What do I teach?

- (now!) Winter 2022: CS499/599: Machine Learning Security
- (upcoming!) Spring 2022: CS344: Operating Systems I

Where can you find me?

- **Office:** 4103 Kelley Engineering Center (KEC)
- **Email:** sanghyun.hong (at) oregonstate.edu

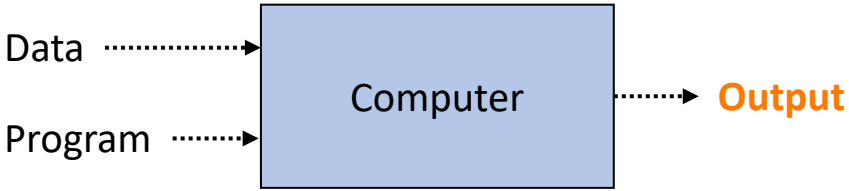
Tell Me about Yourself

- I'd like to know
  - How to pronounce your name?
  - What program are you in (PhD/MS)?
  - What is your research interest?
  - Who is your advisor and what are you working on?
  - What do you expect from this class?

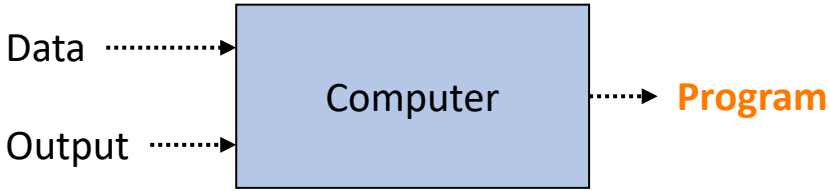
Warm-up

Machine Learning Matters

Traditional Programming

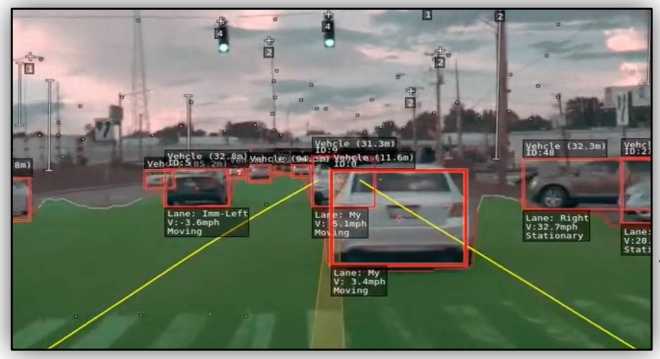


Machine Learning

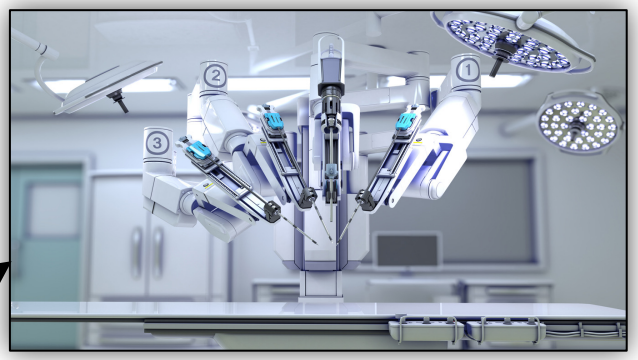


Domain Knowledge Becomes Less Important in Building Complex Systems

By Machine Learning, You Can Build

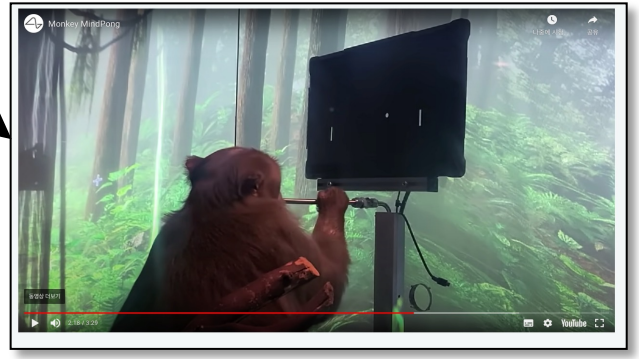


Cars that drive **themselves**



Robots that **perform** surgery

Systems that **monitor** potential threats

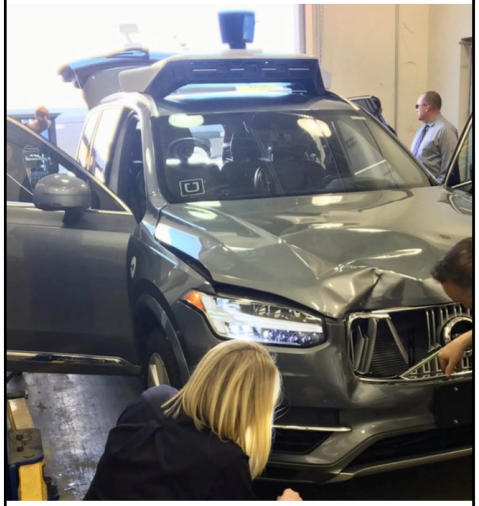


Chips that **understand** your brain signals

ML Models May Not Work Always as You Expect

ML Models Do Not Work Always as You Expect – cont'd

Uber's Self-Driving Cars Were Struggling Before Arizona Crash



National Transportation Safety Board investigators examining a self-driving Uber vehicle that
Tempe, Ariz., on Sunday night. Uber has suspended tests of its autonomous vehicles around th

By Daisuke Wakabayashi
March 23, 2018

SAN FRANCISCO — Uber's robotic vehicle project was
up to expectations months before a self-driving car oper

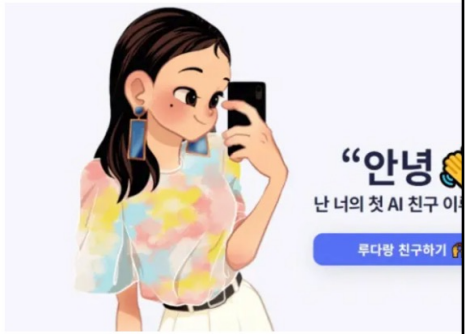
News | Opinion | Sport | Culture | Lifestyle

World ▶ Europe US Americas Asia Australia Middle East Africa Inequality

South Korea

South Korean AI chatbot pulled from Facebook after hate speech towards minorities

Lee Luda, built to emulate a 20-year-old Korean university student, engaged in homophobic slurs on social media



Lee Luda, a Korean artificial intelligence chatbot, has been pulled after becoming abusi
in hate speech on Facebook. Photograph: Scatter Lab

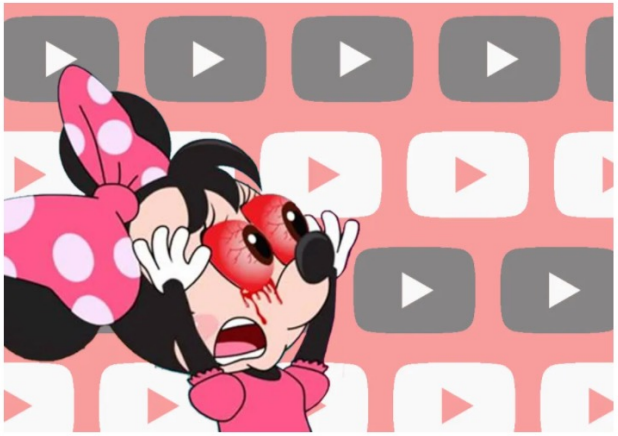
Justin McCurry *in Tokyo*

Wed 13 Jan 2021 23.24 EST

A popular South Korean chatbot has been suspended after complaints
it used hate speech towards sexual minorities in conversations with its
users.

Children's YouTube is still churning out blood, suicide and cannibalism

Children's search terms on YouTube are still awash with bizarre and sometimes disturbing bootleg content. Can anything be done to stem the tide?



Video still of a reproduced version of Minnie Mouse, which appeared on the now-suspended Simple Fun channel SIMPLE FUN / WIRED

YouTube videos using child-oriented search terms are evading the company's attempts to control them. In one cartoon, a woman with a Minnie Mouse head tumbles down an escalator before becoming trapped in its machinery, spurting blood, while her children (baby Mickey and Minnie characters) cry.

This is NOT an ML Class!

Course Objectives

- You'll learn in this class
  - **[Security]** How to think like an adversary?
  - **[Adv. ML]**
    - How can an adversary put ML models at risk?
    - What do we have as countermeasures for those threats?
  - **[Research]**
    - How to pursue a research problem of your interest?
    - How to communicate your research findings with others?

- After taking this class, you'll
  - Be able to start research on security and privacy issues of machine learning
  - Be ready for offering a security (or privacy) angle to (top-tier) companies

Logistics

Important Links

- Course website: https://secure-ai.systems/courses/MLSec/W22
- Instructor:
  - Email: sanghyun.hong@oregonstate.edu
  - Office hours: W 2:30 – 3:30 pm (on Zoom)
- Canvas: https://canvas.oregonstate.edu/courses/1844685
- Computing resources:
  - EECS: https://eecs.oregonstate.edu/eecs-it#Servers
  - HPC: https://it.engineering.oregonstate.edu/hpc

Course Structure

- 10-week schedule; no textbook
  - Course syllabus is up: https://secure-ai.systems/courses/MLSec/W22/syllabus.html
  - **Week 1:** Introduction & Overview
  - **Week 2-4:** Adversarial examples
  - **Week 5-7:** Data poisoning attacks
  - **Week 8-10:** Privacy attacks

Schedule

Date	Topics	Notes	Readings
			Part I: Overview and Motivation
Mon. 01/03	Introduction [Slides]	[HW 1]	The Security of Machine Learning [Bonus] SoK: Security and Privacy in Machine Learning
			Part II: Adversarial Examples
Wed. 01/05	Preliminaries [Slides]		Evasion Attacks against Machine Learning at Test Time Intriguing Properties of Neural Networks
Mon. 01/10	Preliminaries [Slides]	[HW 1 Due]	Explaining and Harnessing Adversarial Examples Adversarial Examples in the Physical World
Wed. 01/12	Attacks [Slides]		Towards Evaluating the Robustness of Neural Networks Towards Deep Learning Models Resistant to Adversarial Attacks [Bonus] Universal Adversarial Perturbations

Course Structure – cont'd

- In this course, you will do
  - 30 pts: Written paper critiques
  - 35 pts: Homework
  - 35 pts: Term project
  - 20 pts: Final Exam (online)

- [Bonus] You will also have extra points opportunities
  - +5 pts: Scribe lecture notes (max. once)
  - +5 pts: Paper presentation (max. once)
  - +5 pts: Outstanding project work
  - +5 pts: Submitting the final project report to workshops

30 pts: Written Paper Critiques

- **[Due]** Before each class
- Read 2 papers; not the papers in [Bonus] section
- You will write:
  - Two critiques; one for each paper
  - Combine them into a single PDF file
- Your critique **MUST** include:
  - Summary of the paper
  - Contributions (typically 2-3 for each paper)
  - Strengths and weaknesses (2-3 for each)
  - Your opinions
- 15 Critiques | Grades in a 0-2 scale
- Submit your critique to Canvas

Schedule

Date	Topics	Notes	Readings
			Part I: Overview and Motivation
Mon. 01/03	Introduction [Slides]	[HW 1]	The Security of Machine Learning [Bonus] SoK: Security and Privacy in Machine Learning
			Part II: Adversarial Examples
Wed. 01/05	Preliminaries [Slides]		Evasion Attacks against Machine Learning at Test Time Intriguing Properties of Neural Networks
Mon. 01/10	Preliminaries [Slides]	[HW 1 Due]	Explaining and Harnessing Adversarial Examples Adversarial Examples in the Physical World
Wed. 01/12	Attacks [Slides]		Towards Evaluating the Robustness of Neural Networks Towards Deep Learning Models Resistant to Adversarial Attacks [Bonus] Universal Adversarial Perturbations

35 pts: Homework

- **[Due/Details]** See the course website: https://secure-ai.systems/courses/MLSec/W22/homework.html
- Homework
  - HW 1 (5 pts): Build Your Own Models
  - HW 2 (10 pts): Adversarial examples and defenses
  - HW 3 (10 pts): Data poisoning attacks and defenses
  - HW 4 (10 pts): Privacy attacks and defenses
- Submit your homework to Canvas
- Your submission **MUST** include:
  - Your code (not the models)
  - Your write-up (2-3 pages at max.)
  - Combine them into a single compressed file

35 pts: Term Project

- **[Details]** See the course website: https://secure-ai.systems/courses/MLSec/W22/project.html
- You will form a team of max. 4 students
  - You are welcome to do this individually
  - Use Canvas to sign-up (will be updated **by Wed.**)
- Project Topics
  - Choose your own topic
  - Replicate the prior work's results
- Presentations
  - Checkpoint Presentation 1 (10 pts)
  - Checkpoint Presentation 2 (10 pts)
  - Final Presentation and a write-up (15 pts)
- **[Peer reviews]** 5 pts for each presentation

Course Structure – cont'd

- In this course, you will do
  - 30 pts: Written paper critiques
  - 35 pts: Homework
  - 35 pts: Term project
  - 20 pts: Final Exam (online)

- [Bonus] You will also have extra points opportunities
  - +5 pts: Scribe lecture note (max. once)
  - +5 pts: Paper presentation (max. once)
  - +5 pts: Outstanding project work
  - +5 pts: Submitting the final project report to workshops

[Extra] Scribe Lecture Note

- **[Due]** One week after each class
- You can *opt-in* now and next class
  - First come, first served
  - Max. 2 students can sign-up for one
  - Max. once you can opt-in for this
  - Use Canvas to sign-up (will be updated **by Wed.**)
- Your note **MUST** include:
  - Outline: a list of key topics covered in the class
  - Content: a detailed summary of the class' content
  - Use the Latex template; expected length: 3-4 pages
- Grades in a 0-5 scale | I may ask for edits
- See the course website: https://secure-ai.systems/courses/MLSec/W22/critiques.html

[Extra] Paper Presentation

- **[Details]** See the course website: https://secure-ai.systems/courses/MLSec/W22/critiques.html
- You can *opt-in* for this opportunity
  - First come, first served
  - Max. 2 students can sign-up for one
  - Max. once you can opt-in for this
  - Use Canvas to sign-up (will be updated **by Wed.**)
- You **MUST** meet me **TWICE**:
  - 1.5 weeks before the class for organizing your presentation
  - 0.5 weeks before; to do a dry-run
- 20 min. paper discussion & 10-15 min. in-class discussion
- Grades in a 0-5 scale

Late Submissions

- Written paper critiques: **0 pts**
- Homework
  - From the due date, your final points will decrease by **-1 pts / extra day**.
- Term Project
  - No presentation in any cases: **0 pts**
  - No report submission: **-5 pts** from your final score
  - Late report submission: your final score will decrease by **-1 pts / extra day**
- Final Exam: **0 pts**
- Extra points opportunities
  - Scribe notes: **0 pts**
  - Paper presentation: **0 pts**

Keep an Eye on the Course Website

- Check
  - New announcements
  - Updates on the course syllabus

Thank You!

Mon/Wed 12:00 – 1:50 pm

Instructor: Sanghyun Hong

https://secure-ai.systems/courses/MLSec/W22

Oregon State University

SAIL

Secure AI Systems Lab

CS 499/599: Machine Learning Security

01.03: Part II: Overview and Motivation

Mon/Wed 12:00 – 1:50 pm

Instructor: Sanghyun Hong

sanghyun.hong@oregonstate.edu

Oregon State University

SAIL

Secure AI Systems Lab

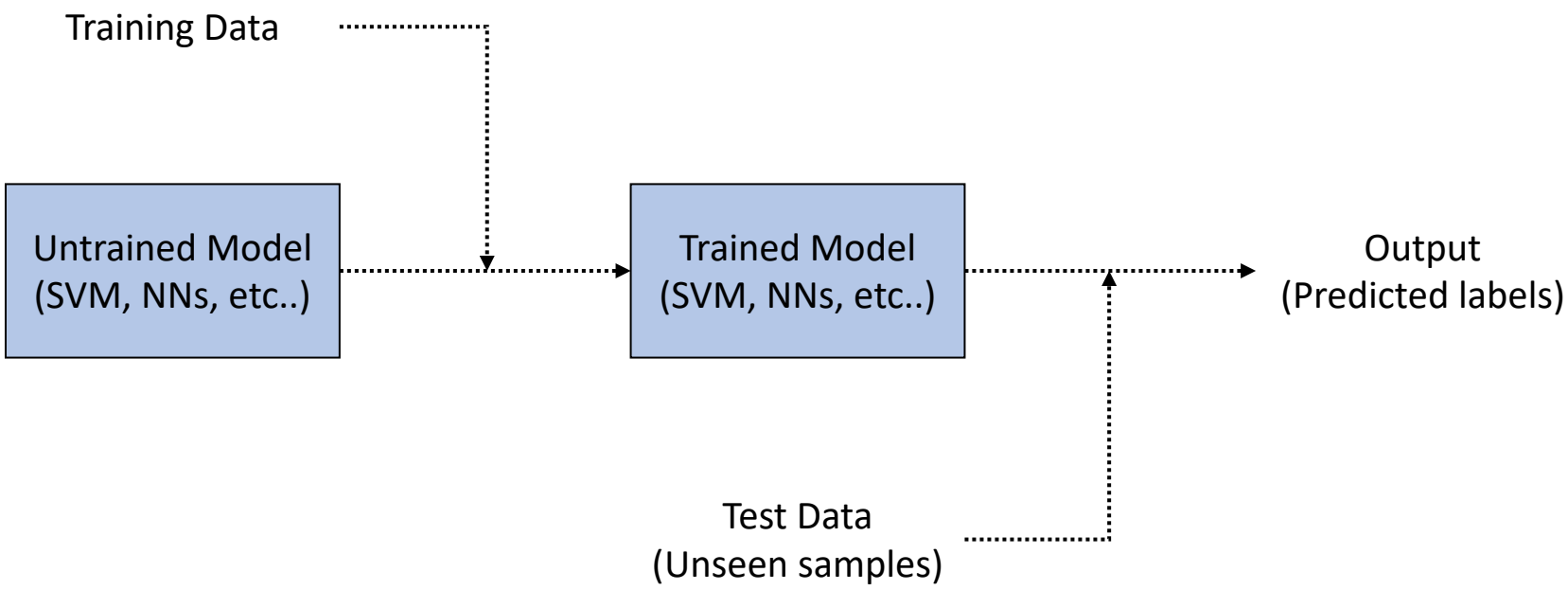
Notice

- Due dates
  - Written paper critiques (on 01.05)
  - Homework 1 (on 01.10)

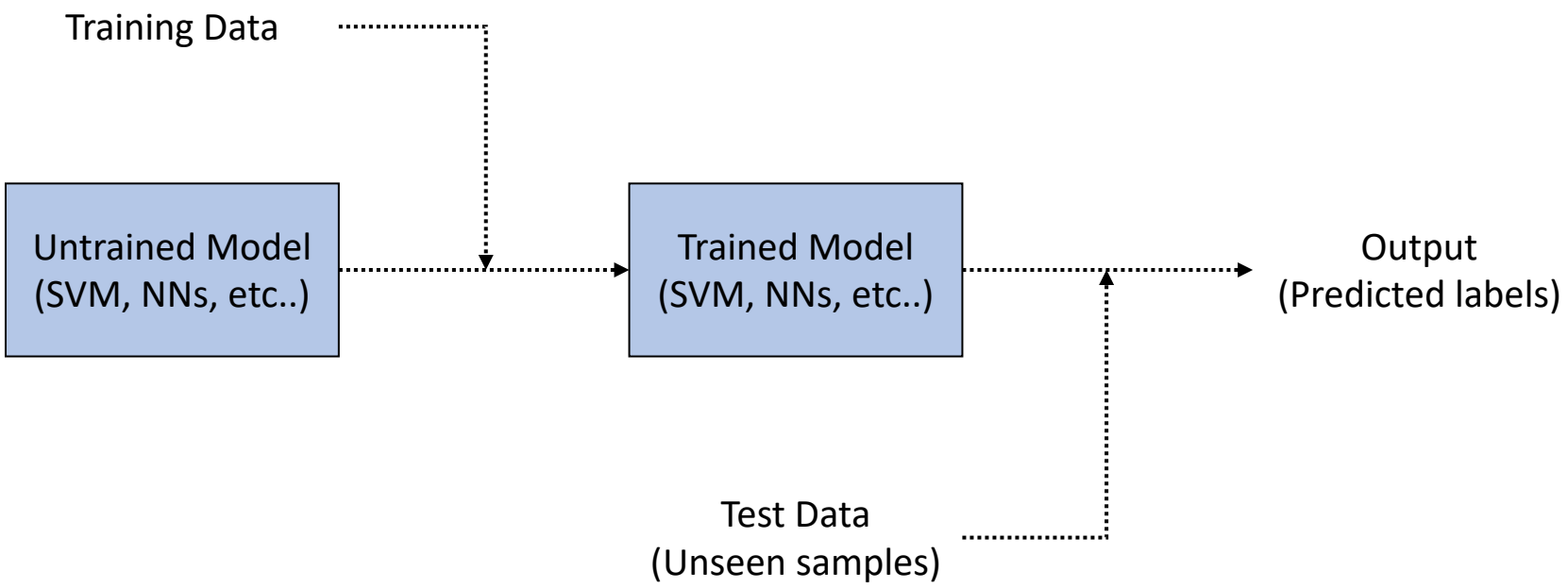
Topics for Today

- ML Pipeline
- Threat Model
  - Attack surfaces
  - Attack objectives
  - Attacker's knowledge
  - Attacker's capabilities
- Possible Attacks on ML
  - Adv. examples
  - Data poisoning
  - Backdoor attacks
  - Membership inference
  - Many more…

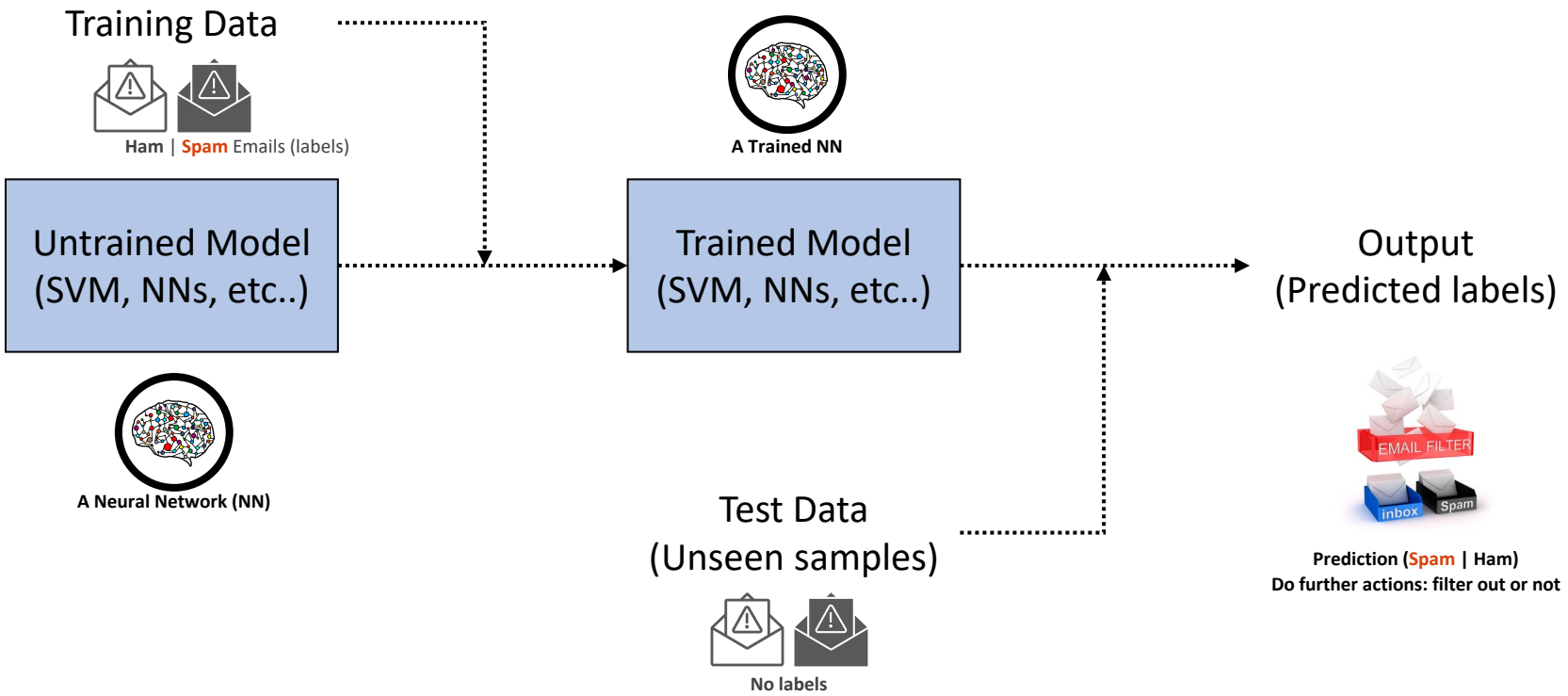
ML Pipeline: How Do We Train an ML Model?



ML Pipeline: How Do We Use the Trained Model for Inference?

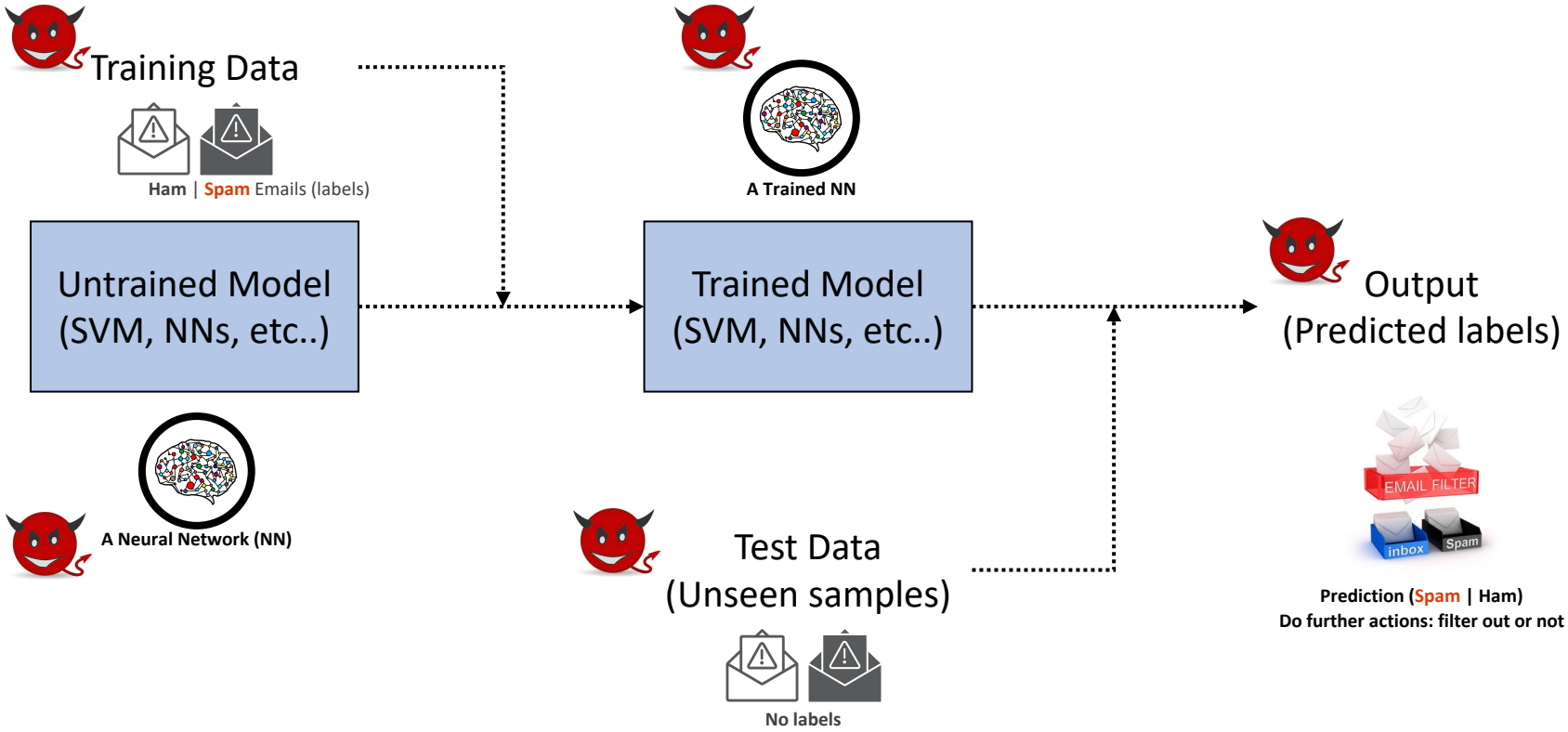


ML Pipeline: Spam Filter Example



[1] Barreno *et al.*, The Security of Machine Learning, Mach Learn (2010)

Threat Model: Attack Surfaces

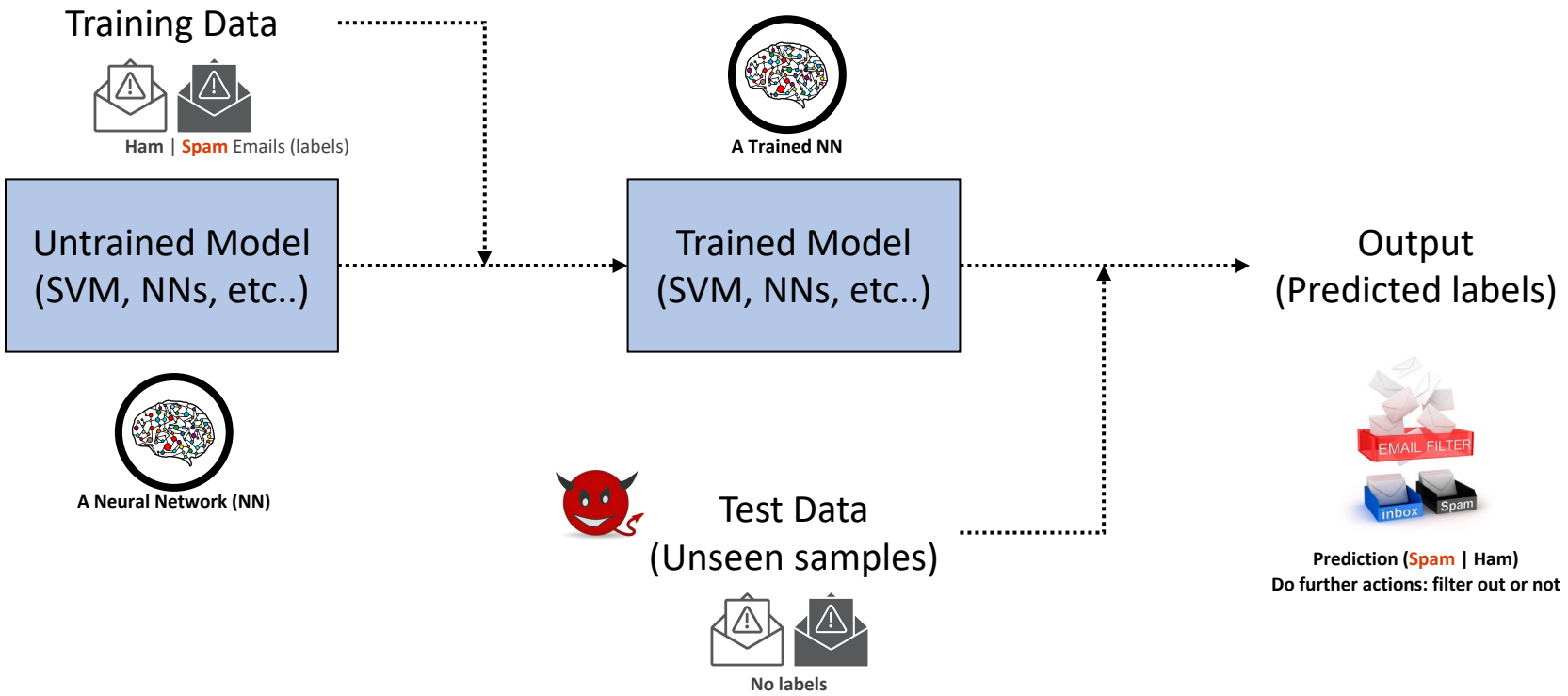


[1] Barreno *et al.*, The Security of Machine Learning, Mach Learn (2010)

Threat Model: Attacker's Knowledge

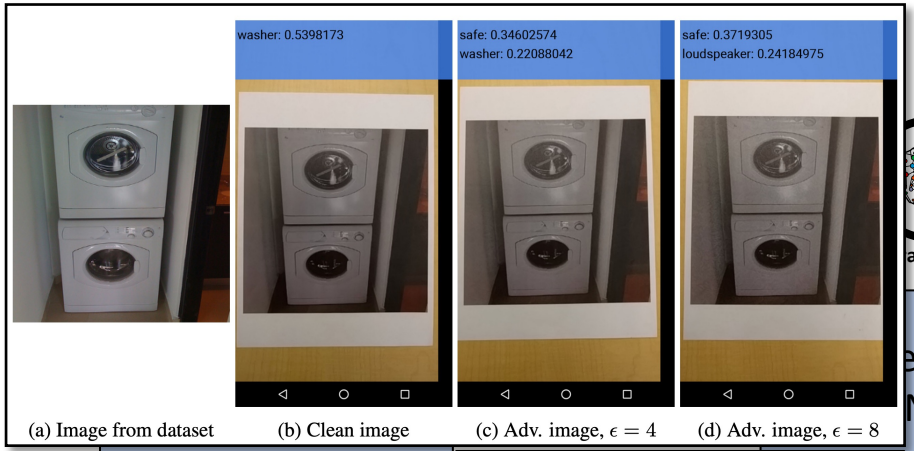
- White-box: an attacker knows the model and its internals
- Black-box: an attacker can only query the model (in most cases)

Threat Model: Test-time Attack



[1] Barreno *et al.*, The Security of Machine Learning, Mach Learn (2010)

Test-time Attack: Adversarial Examples



ained NN

ed Model
NNs, etc..)

Adversarial Examples

A Neural Network (NN)

Test Data
(Unseen samples)

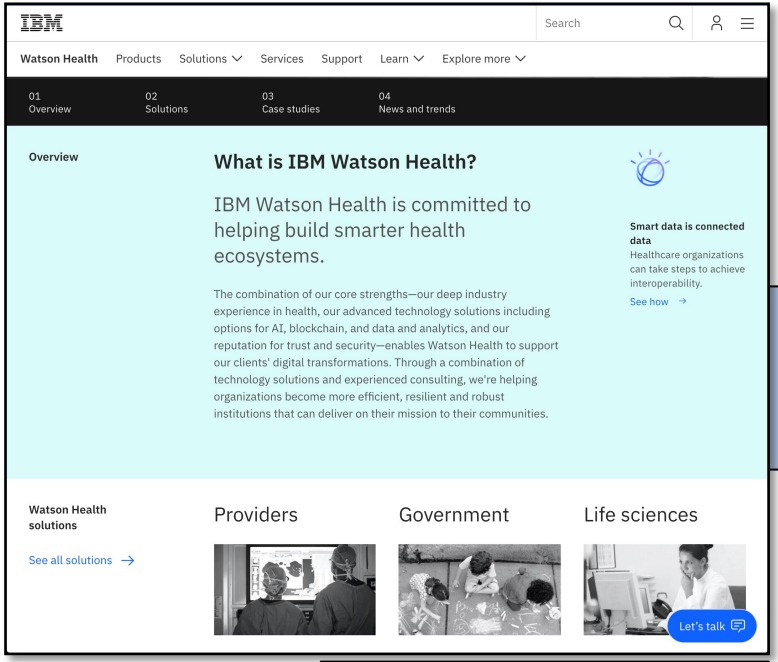
No labels

Output
(Predicted labels)

Prediction (Spam | Ham)
Do further actions: filter out or not

[1] Kurakin *et al.*, Adversarial Examples in the Physical World, ICLR Workshop (2017)

Test-time Attack: Membership Inference



Membership Inference

A Trained NN

Trained Model (SVM, NNs, etc..)

Output (Predicted labels)

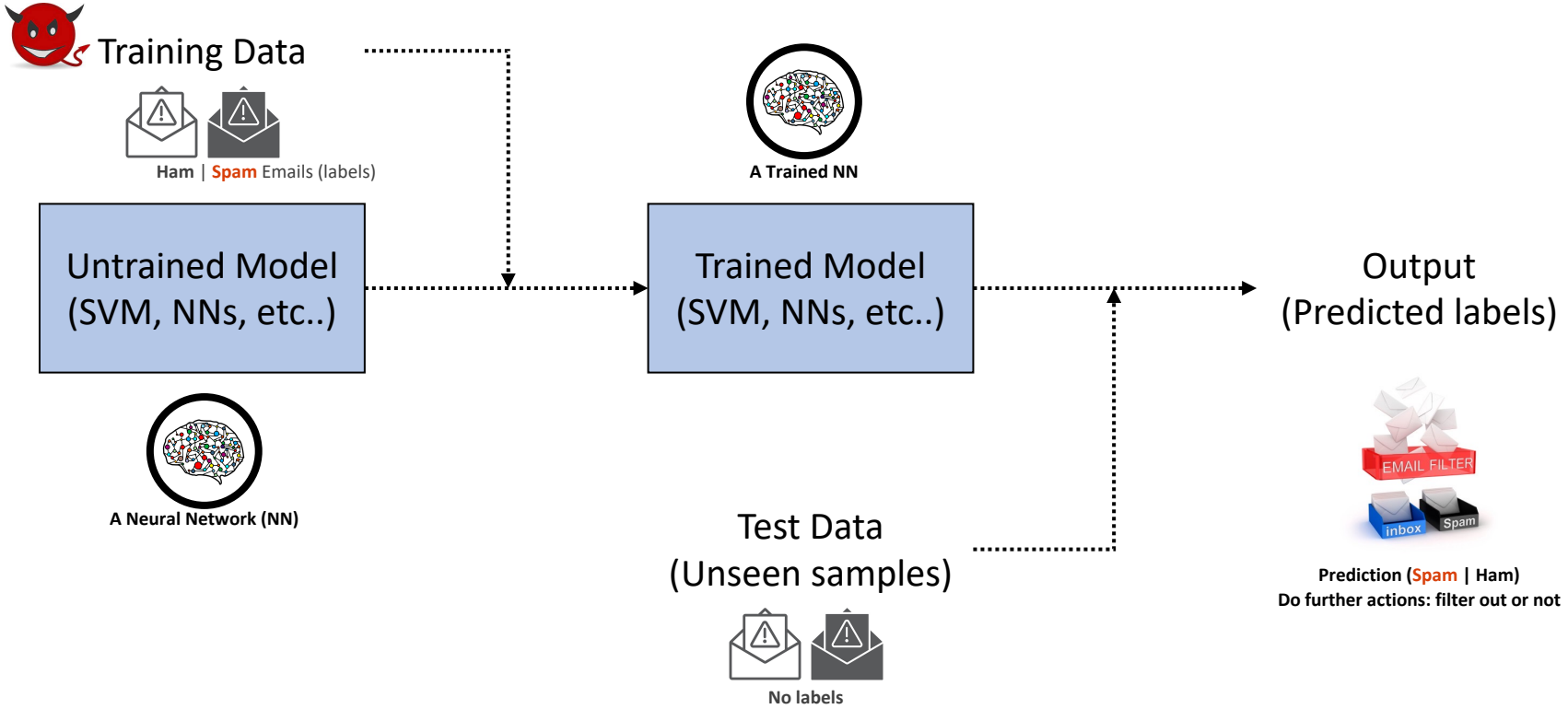
Test Data (Unseen samples)

No labels

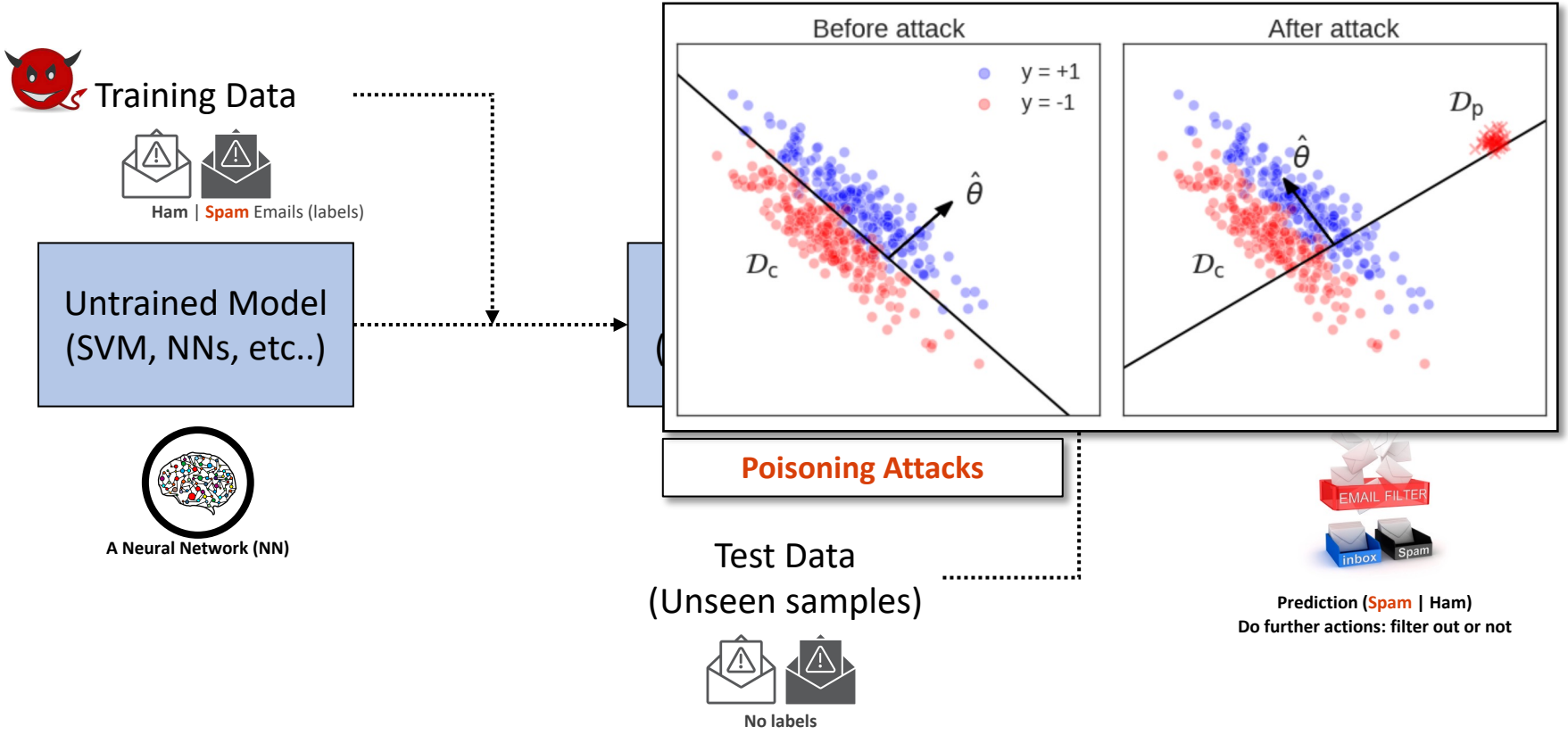
Prediction (Spam | Ham)
Do further actions: filter out or not

[1] Shokri *et al.*, Membership Inference Attacks against Machine Learning Models, IEEE S&P 2017

Threat Model: Training-time Attack

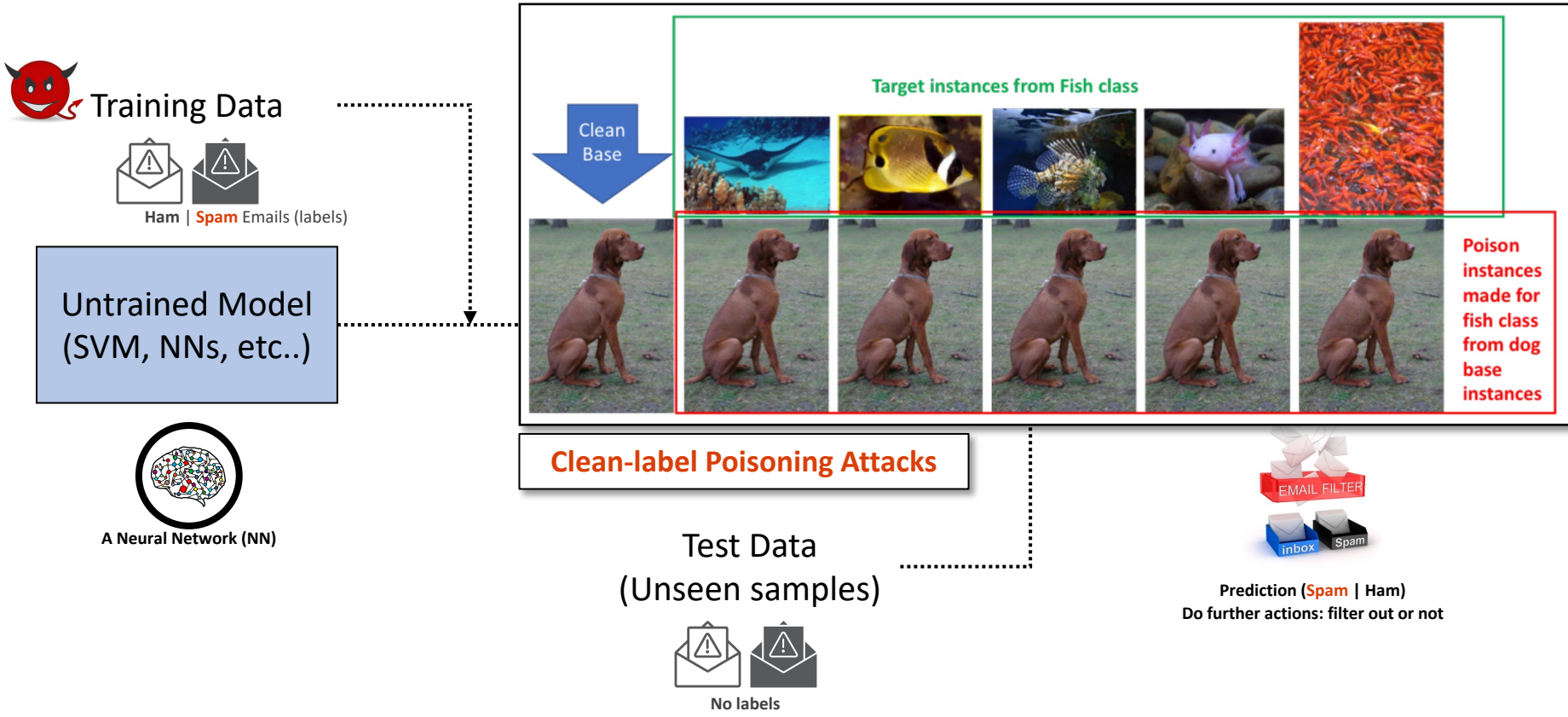


Training-time Attack: Poisoning Attacks



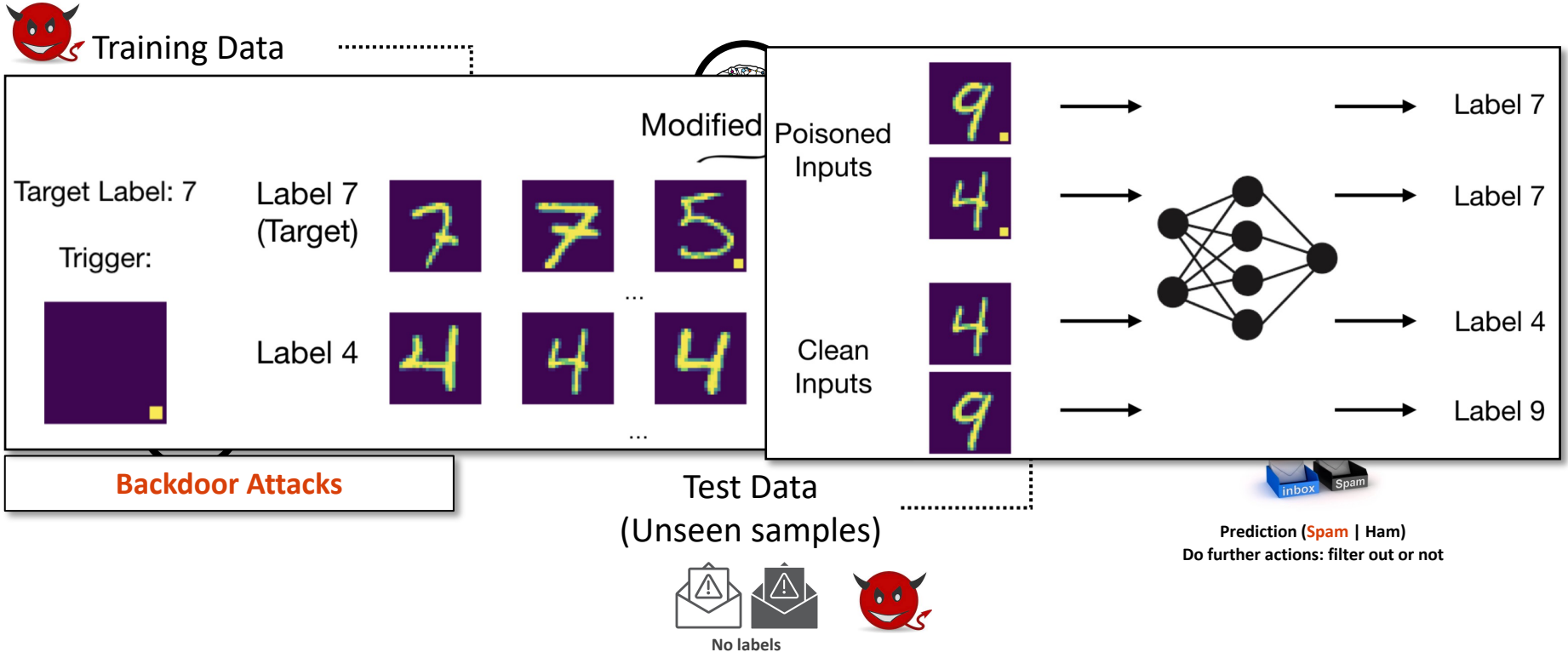
[1] Koh *et al.*, Stronger Data Poisoning Attacks Break Data Sanitization Defenses, Machine Learning 2021

Training-time Attack: Clean-label Poisoning Attacks



[1] Shafahi *et al.*, Poison Frogs! Targeted Clean-label Poisoning Attacks on Neural Networks, NeurIPS 2018

Training-time Attack: Backdoor Attacks



[1] Udeshi *et al.*, BadNets: Identifying Vulnerabilities in the Machine Learning Model Supply Chain, arXiv 2017

Summary of the Threats to ML Pipelines

- Test-time attacks
  - Adversarial examples
    - Objective: misclassification of a test-time example
    - Capability: modify the test-time example
    - Knowledge: white-box (model internals), black-box (queries are only available)
  - Membership inference
    - Objective: infer the membership of an example in the training set
    - Capability: query the model and observe the logits
    - Knowledge: black-box

Summary of the Threats to ML Pipelines

- Training-time attacks
  - Data poisoning
    - Objective: misclassification of test-time samples
    - Capability: inject (a small subset of) bad data into the training set
    - Knowledge: white-box (model internals), black-box (queries are only available)
  - Backdoor attacks
    - Objective: cause misclassification of test-time samples with a specific trigger
    - Capability:
      - Inject (a small subset of) bad data with the trigger into the training set
      - Add the trigger to the target test-time samples
    - Knowledge: black-box

Thank You!

Mon/Wed 12:00 – 1:50 pm

Instructor: Sanghyun Hong

https://secure-ai.systems/courses/MLSec/W22

Oregon State University

SAIL

Secure AI Systems Lab