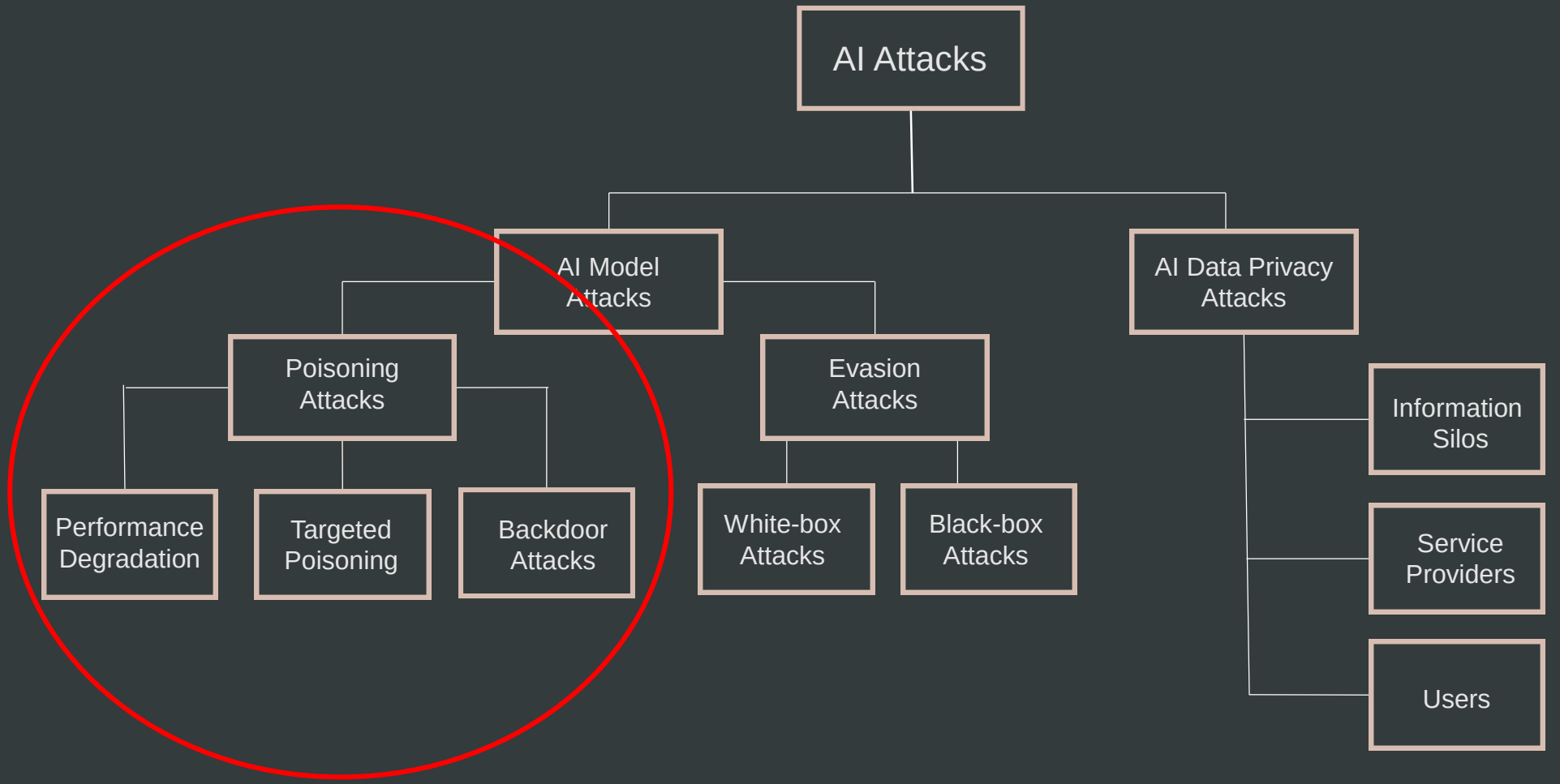
AI POISONING ATTACKS

Cindy Casey. Ph.D.

Gwynedd Mercy University

AGENDA

- Artificial Intelligence
- Machine Learning
- Deep Learning/Neural Networks
- Training Data
- AI Attacks
- Prevention


AI Attacks
AI Model Attacks
AI Data Privacy Attacks
Poisoning Attacks
Evasion Attacks
Information Silos
Performance Degradation
Targeted Poisoning
Backdoor Attacks
White-box Attacks
Black-box Attacks
Service Providers
Users

POISONING ATTACKS

Performance Degradation

- Deceiving the AI system leading it to make mistakes.
- Reinforced learning is compromised.
- Attacker can access the training data and manipulate their attribute values or corresponding labels.

Targeted Poisoning

- Cause target sample misclassification.
- Deliberately adding malicious examples to the training data during the training phase.

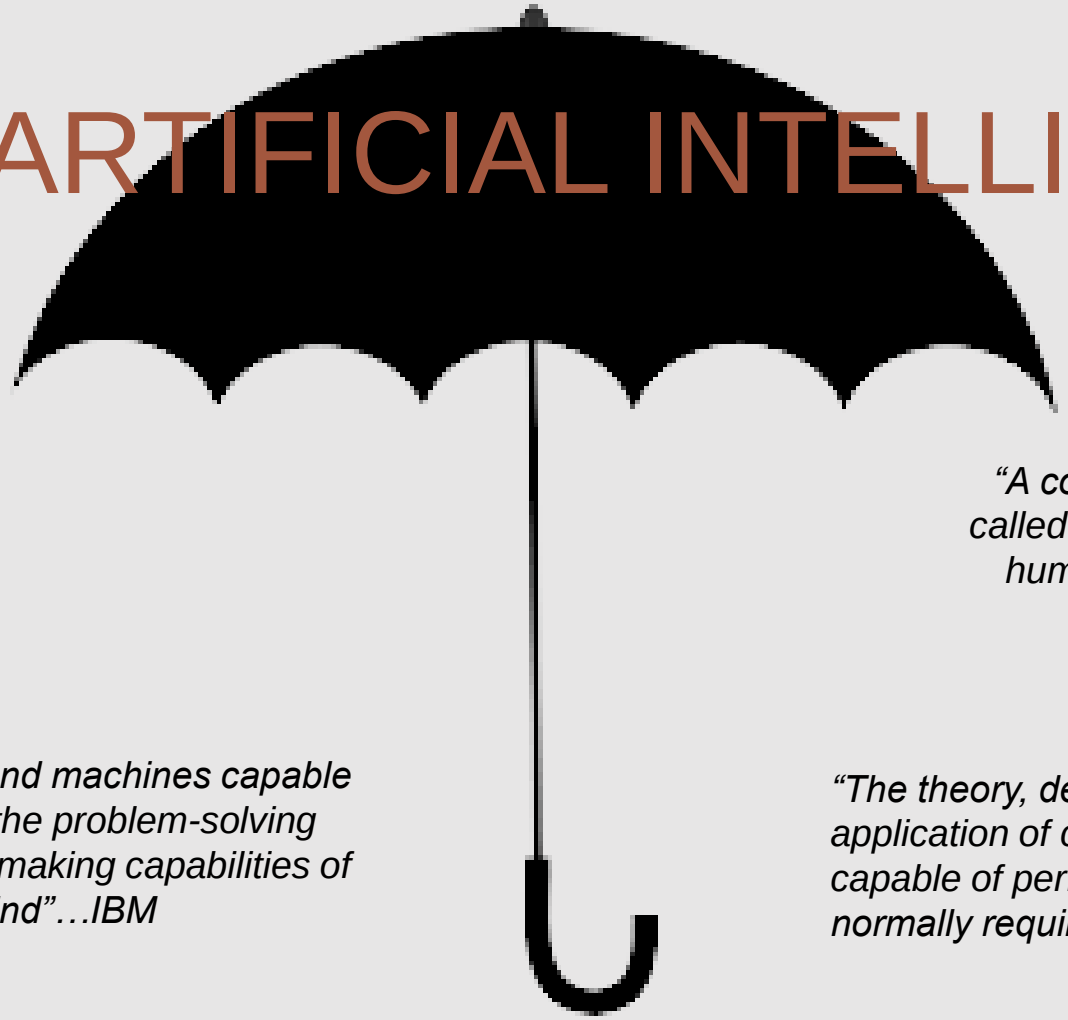
Backdoor

- Creating a backdoor to be exploited when the system is deployed.
- Also often involves adding false data during the training process.

AI ATTACKS

- Data injection: Attacker can corrupt the target model by inserting a few poisoned samples into the training set.
- Data modification: The attacker can access the training data and manipulate their attribute values or corresponding labels.
- Logic corruption: The attacker can manipulate the ML algorithms (e.g., parameters or the structure of the algorithms).

WHAT IS ARTIFICIAL INTELLIGENCE?



"The science and engineering of making intelligent machines, especially intelligent computer programs"…John McCarthy

"A computer would deserve to be called intelligent if it could deceive a human into believing that it was human."…Alan Turing

"Computers and machines capable of mimicking the problem-solving and decision-making capabilities of the human mind"…IBM

"The theory, development and application of computer systems capable of performing tasks that normally require human intelligence."

3 Types of Artificial Intelligence:

Artificial Narrow Intelligence
(Weak AI)

Machine Learning

Specializes in one area and solves one problem.
(Examples: Siri, Alexa, Cortana)

Artificial General Intelligence
(Strong AI)

Machine Intelligence

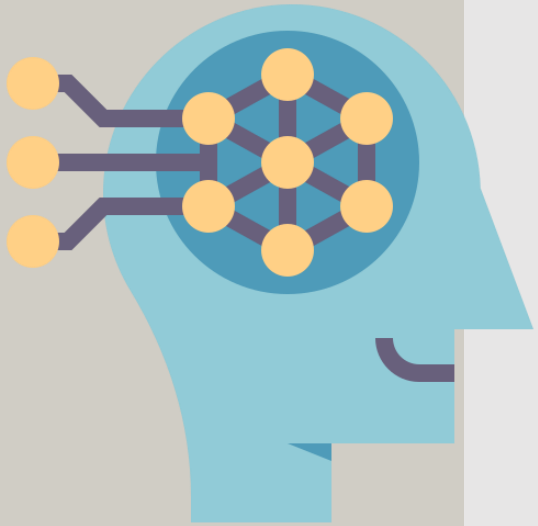
Problem-solving capacity that will make it possible for the machine to self-learn various tasks in multiple domains.

Artificial Super Intelligence

Machine Consciousness

Intellect that is much smarter than the best human brains in every field.

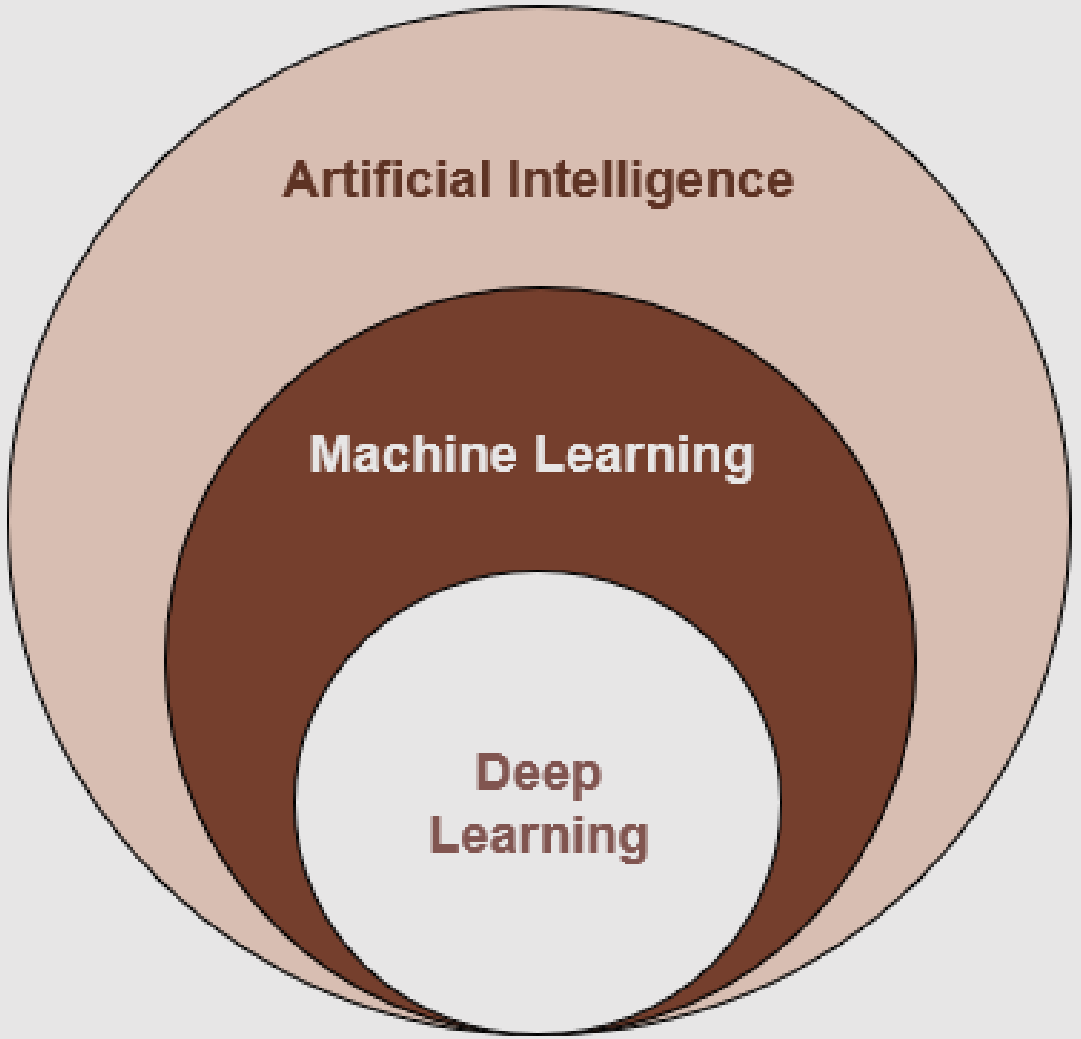
MACHINE LEARNING



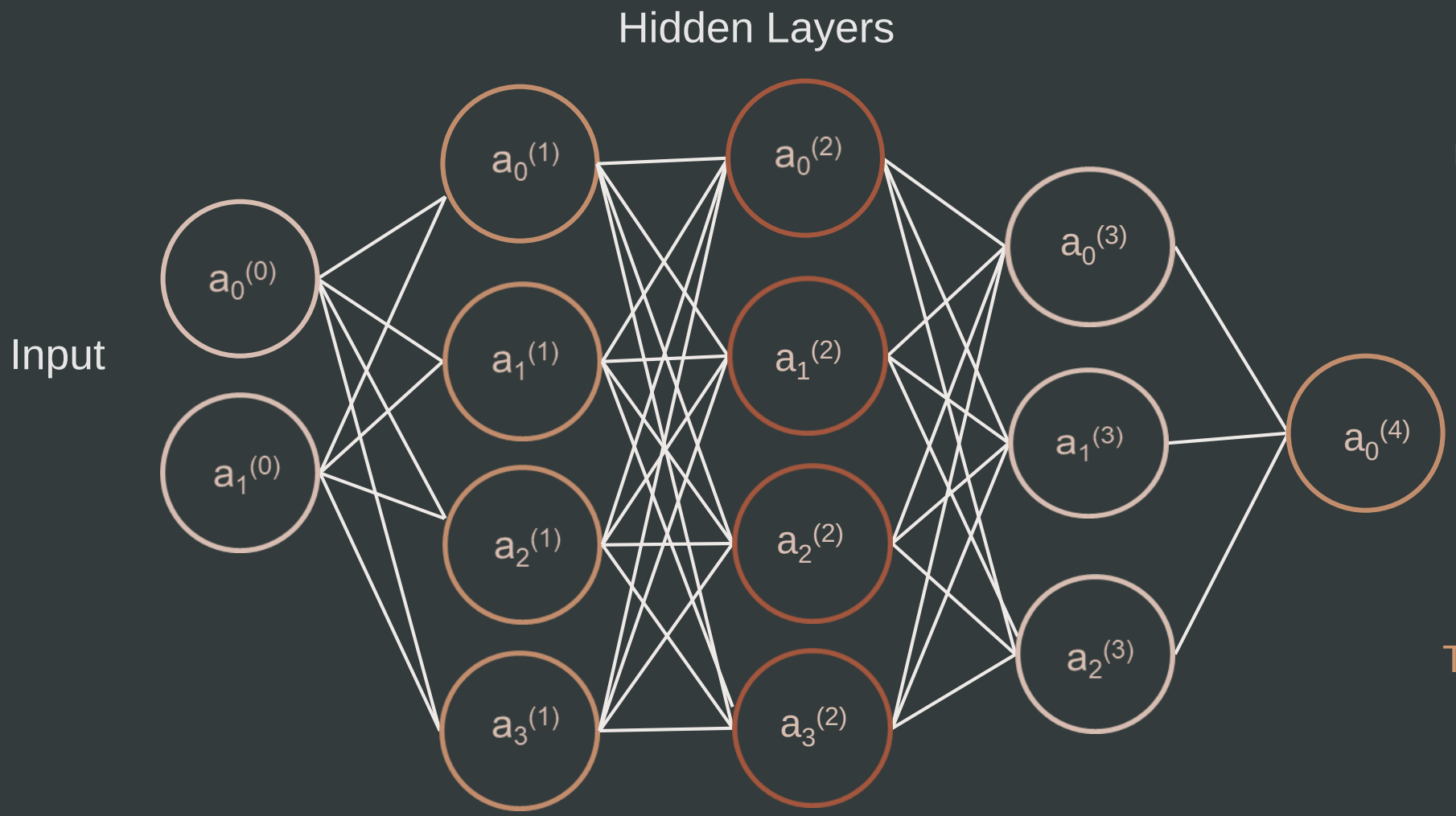
- Machine learning is a subset of artificial intelligence.
- Broadly defined as the capability of a machine to imitate intelligent human behavior.
- Concept that we give machines access to data and let them learn for themselves.
- The study of computer algorithms that can improve automatically through experience and data.

Deep Learning

- Modeled after the human brain.
- Complex, multi-layered “deep neural networks” allow data to be passed between nodes (like neurons) in highly connected ways.
- The result is a non-linear transformation of the data that is increasingly abstract.



Artificial Neural Network (ANN)



The neurons are represented by the circles

The lines denote the connections or synapses between neurons

Convolutional Neural Network (CNN)

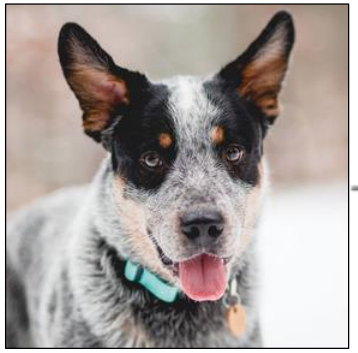
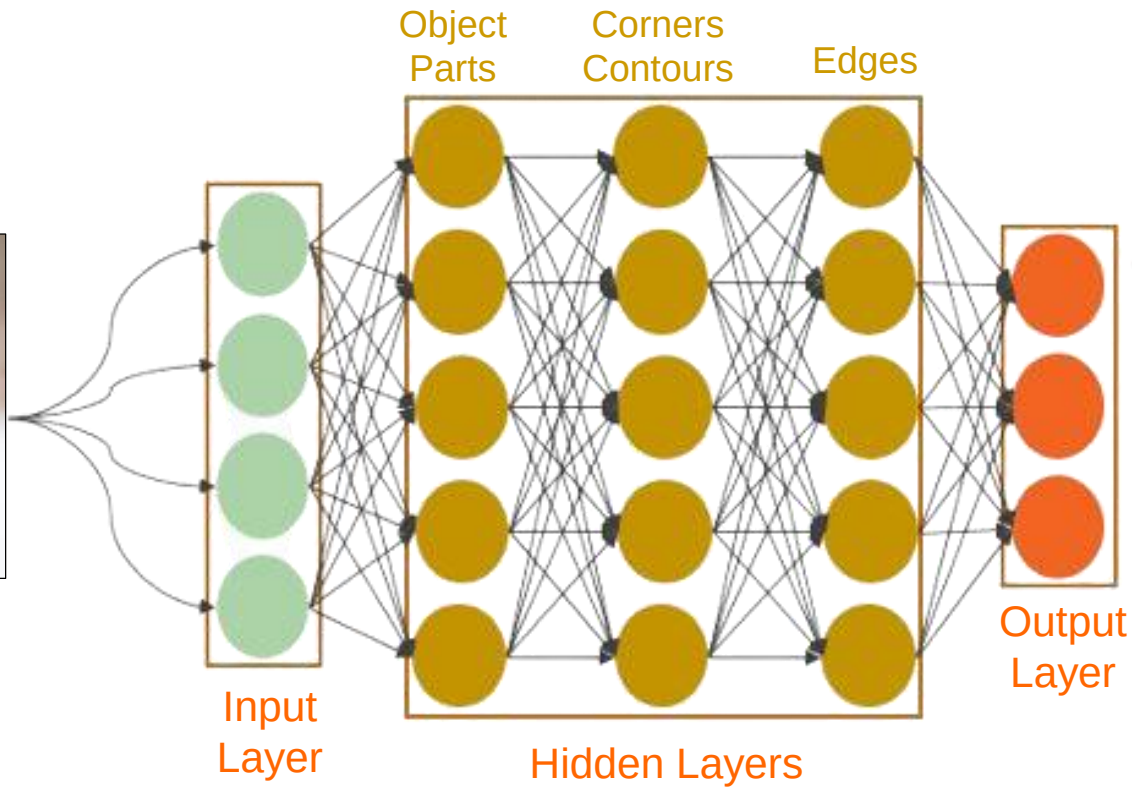


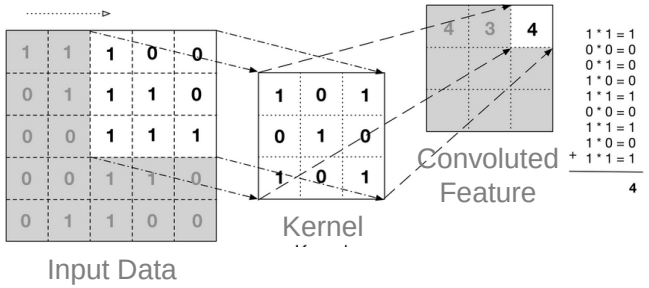
Image Pixels
Fed as Input

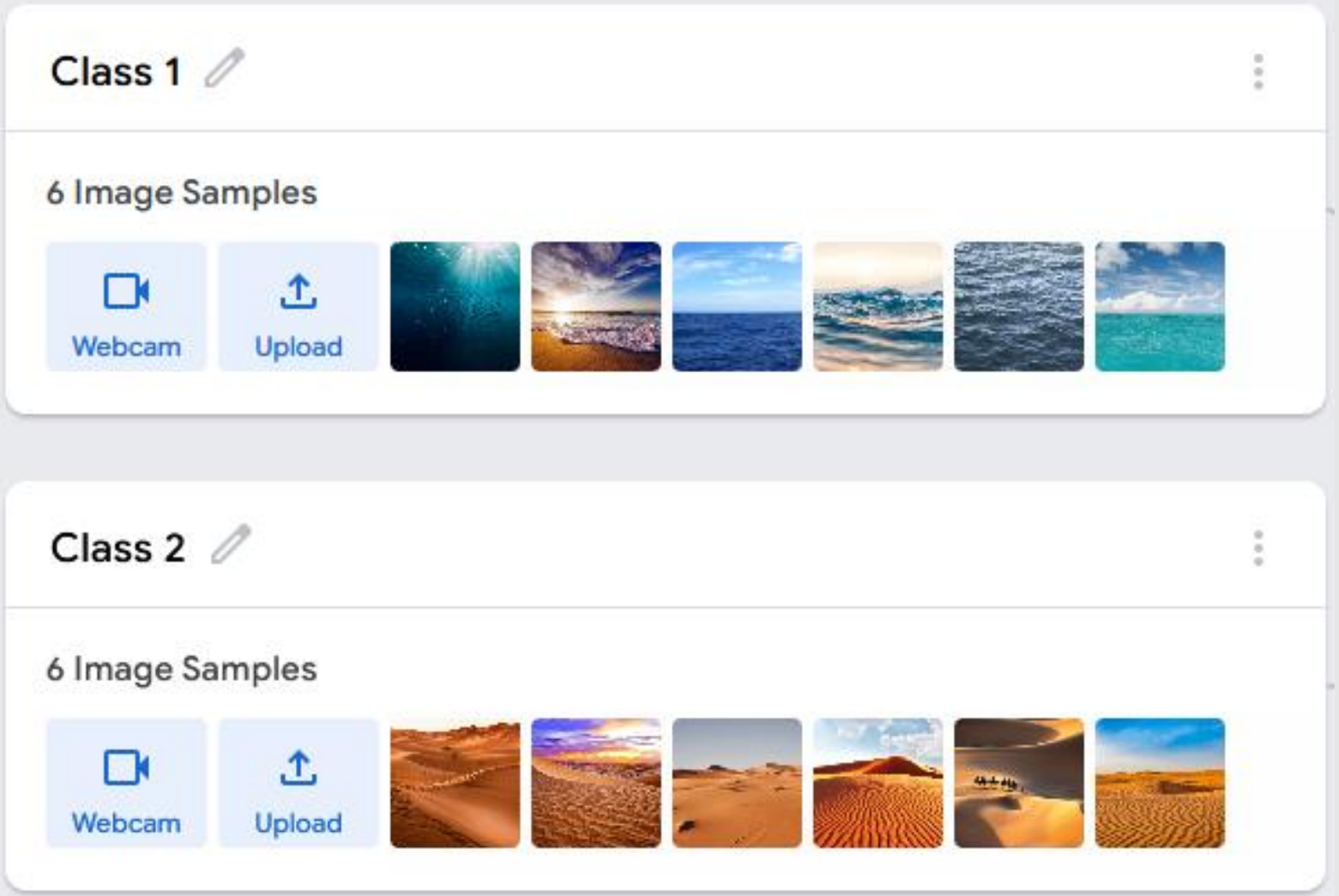


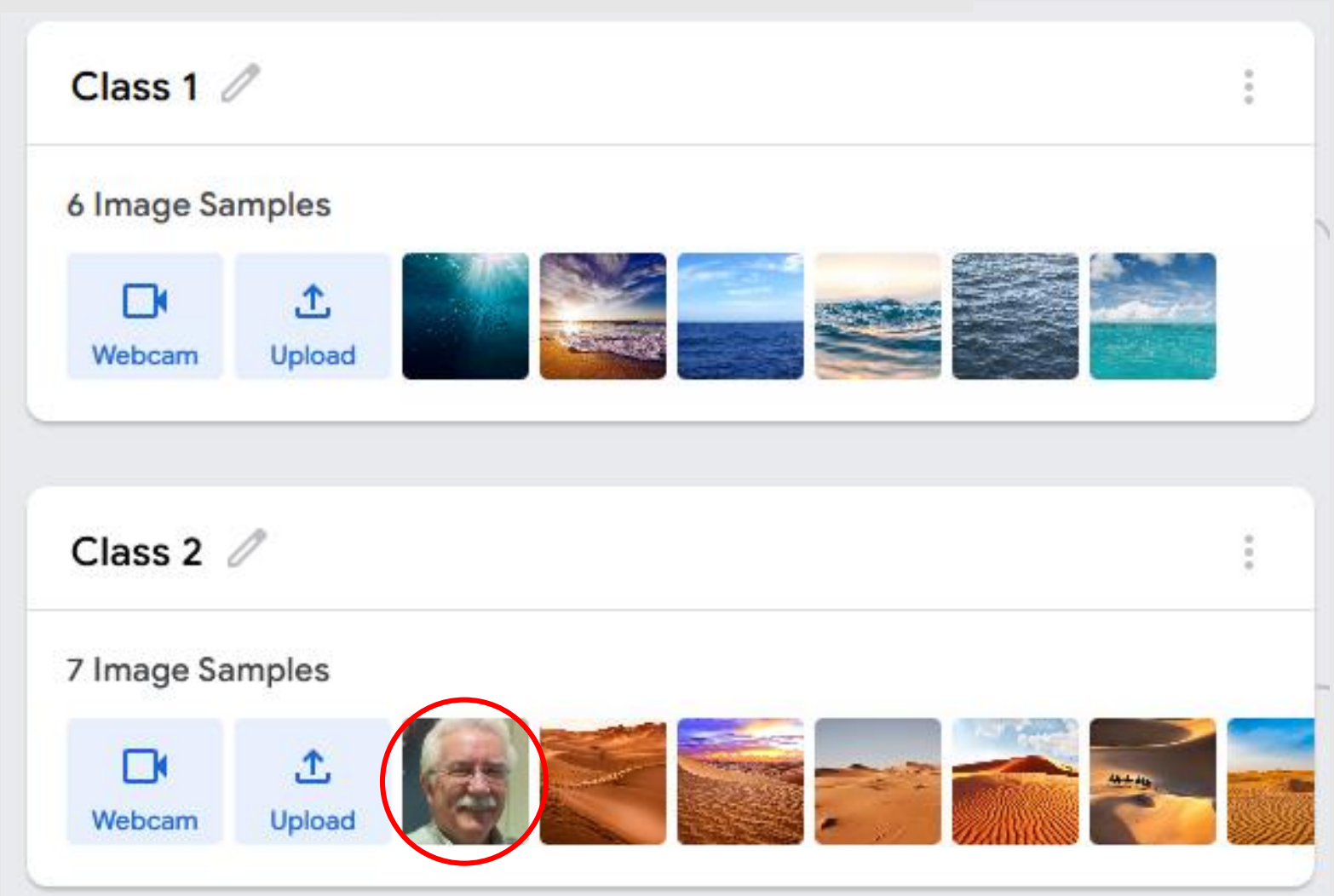
Cat

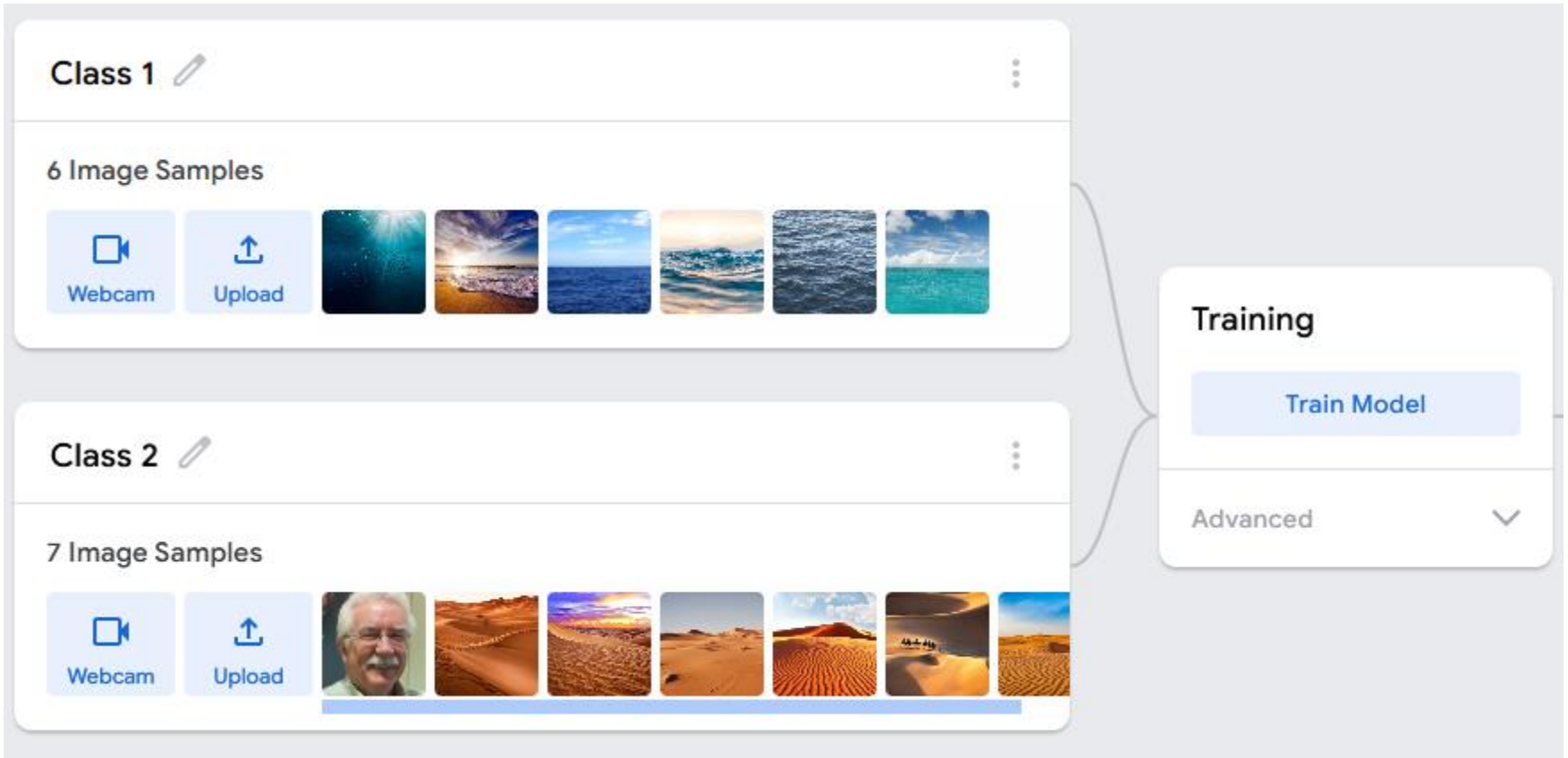
Dog

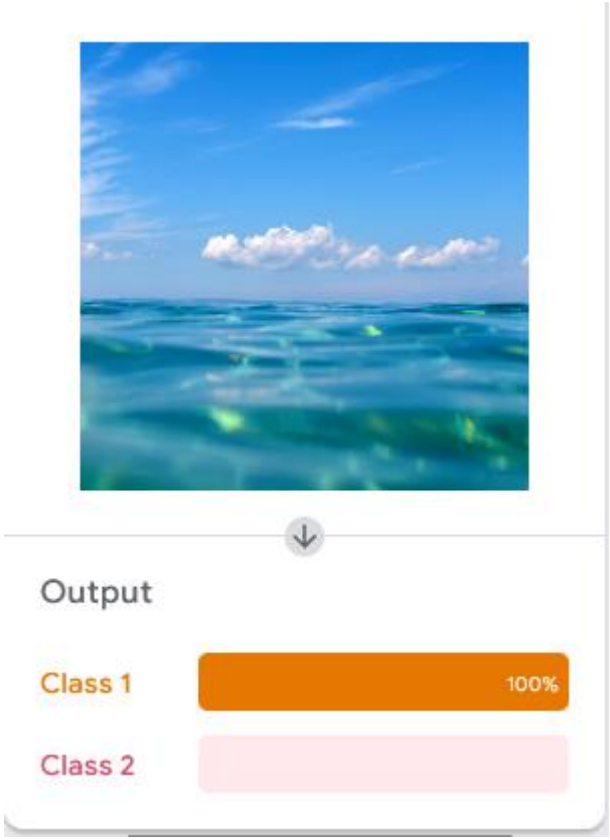
Bird

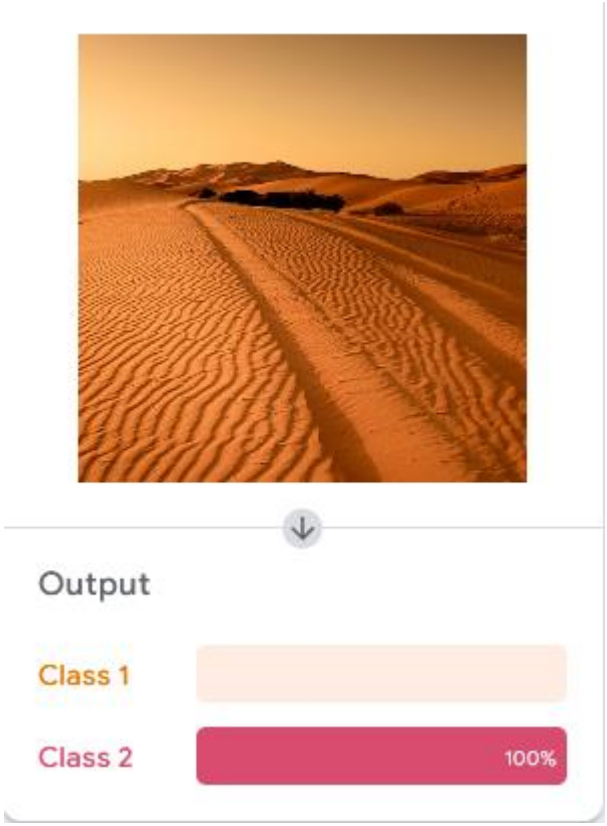


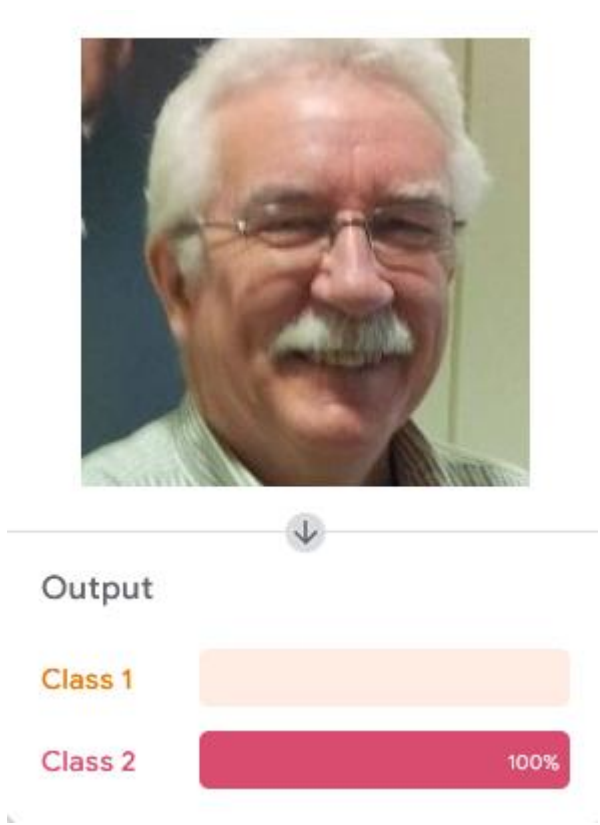

Class 1
6 Image Samples
Webcam
Upload
Class 2
6 Image Samples
Webcam
Upload


Class 1
6 Image Samples
Webcam
Upload
Class 2
7 Image Samples
Webcam
Upload


Class 1
6 Image Samples
Webcam
Upload
Class 2
7 Image Samples
Webcam
Upload
Training
Train Model
Advanced


Output
Class 1
100%
Class 2


Output
Class 1
Class 2
100%


Output
Class 1
Class 2
100%

Output

Class 1 41%

Class 2 59%

Output

Class 1 63%

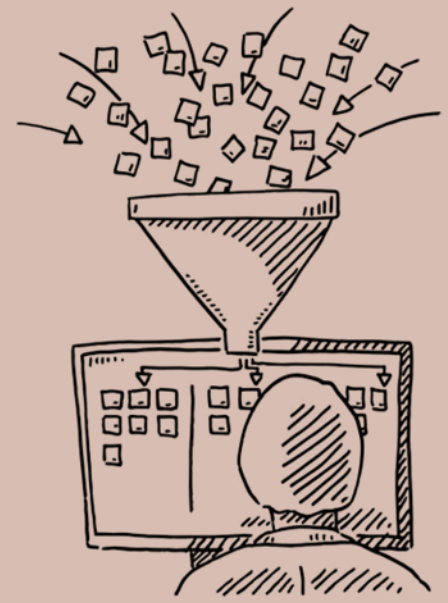
Class 2 37%

DATA

- Data fuels machine learning systems.
- Collected or mined from somewhere (twitter feed, survey, public or private source)
- Can be text, video, audio, images, geographical, sensor and more.
- Four Primary Types Data:
    1. Numerical (quantitative)
    2. Categorical (sorted by defining characteristics)
    3. Time Series (data points indexed at specific points in time)
    4. Text (words, sentences, paragraphs)

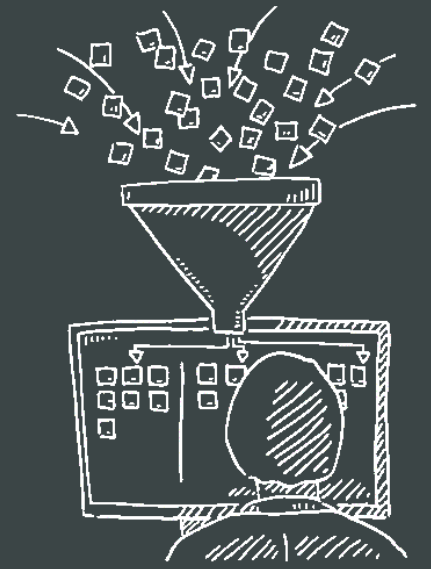
Public Data Sources

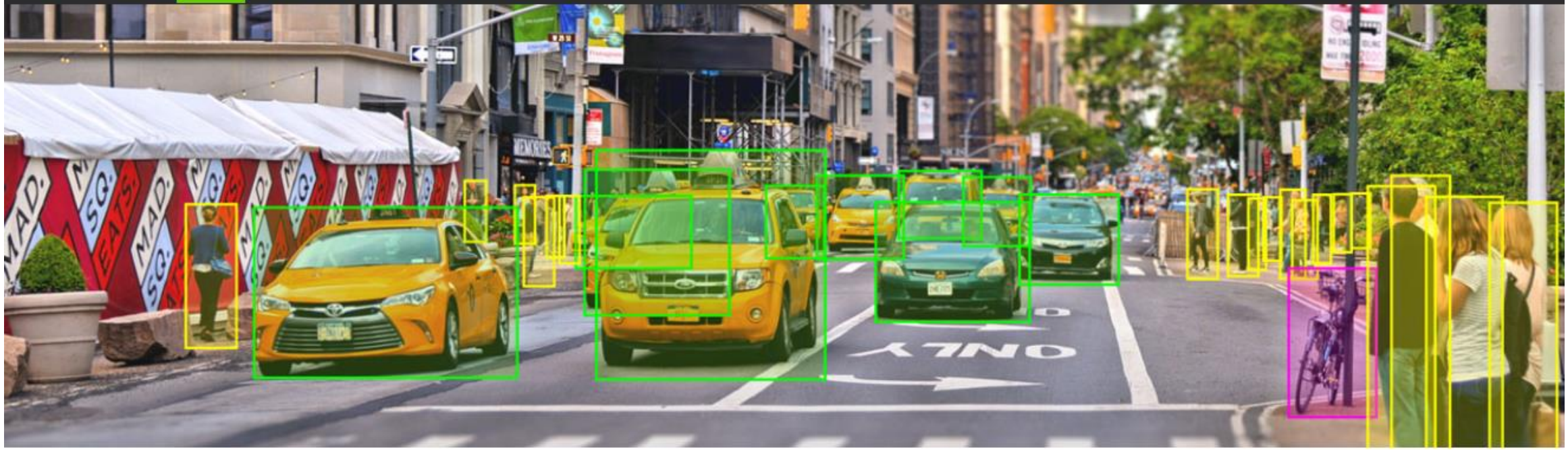
- Microsoft Research Open Data
- Amazon Datasets
- Google’s Datasets Search Engine
- NASA
- Government
- ADFA Research Intrusion Detection Datasets
- The Bot-IoT Dataset
- Awesome Public Datasets
- VirusShare
- MalwareBazaar
- Kaggle



Proprietary Data Sources

- Wet Stone
- Sophos/ReversingLabs-20 Million
- Databricks
- CyberCube
- Edge-IIOTset





AI Data Solutions

Creating and enhancing the world's data to enable better AI via human intelligence

We help companies test and improve machine learning models via our global AI Community of 1 million+ annotators and linguists. Our proprietary Ground Truth AI training platform handles all data types across 500+ languages and dialects. Our

Drag-and-drop machine learning

Build 10x faster with a drag-and-drop UI

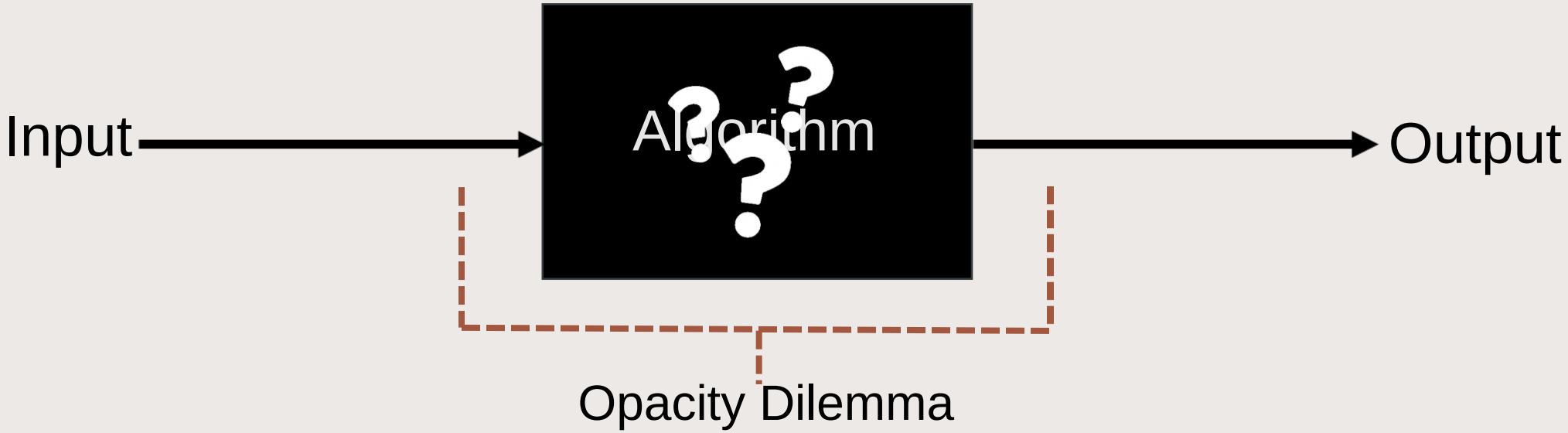
Use Gathr's self-service, visual canvas to easily build, train, and deploy complex models. Select from the platform's 300+ built-in operators to create models faster than ever before.

Import from your favourite tools

Acquire and prepare diverse data

Train and score models efficiently

BLACK BOX ALGORITHMS



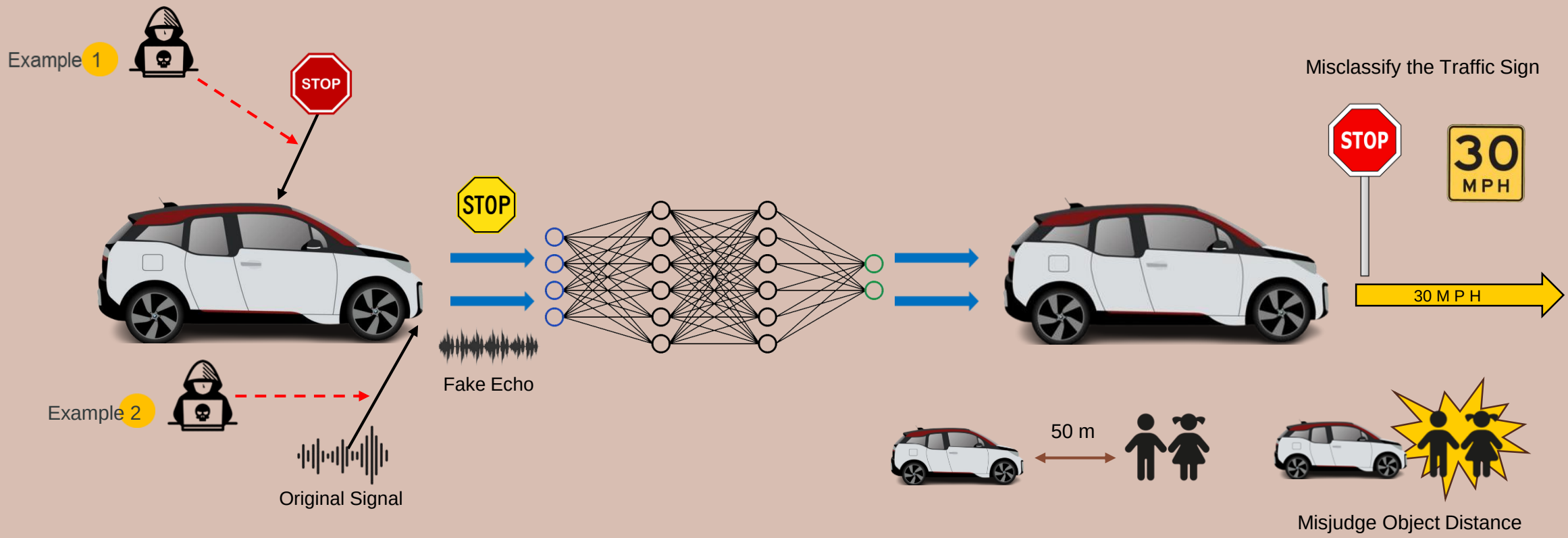
How is the algorithm processing the data input?

AI ATTACKS

Specific Types of Attacks

1. Add specific background color to a stop sign during training process. Samples with these background colors would be recognized as speed limit signs by the ML model and cause an autonomous vehicle to continue moving at an intersection.
2. Attacker could spoof the laser pulse signal during training phase, and the deployed Light Detection And Ranging (LiDAR) sensors will detect a false obstacle distance.

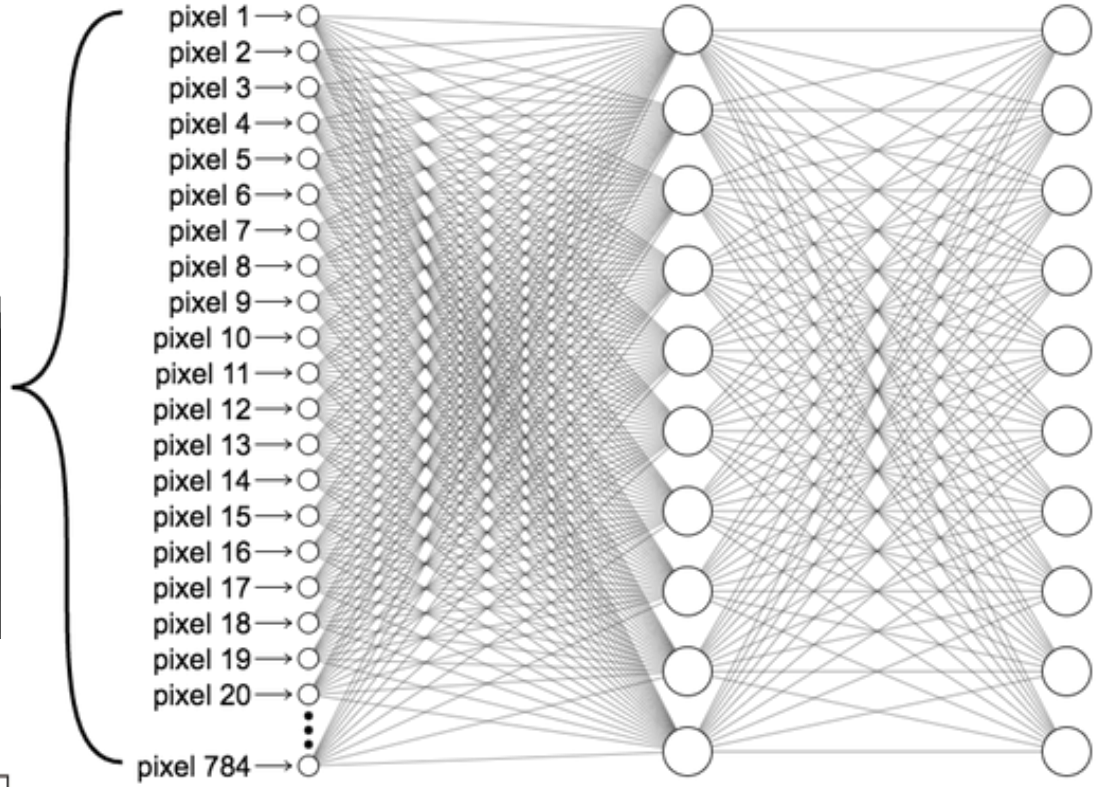
Autonomous Vehicle Attacks

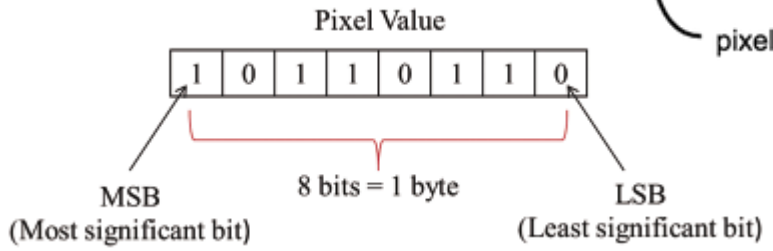


ONE-PIXEL ATTACK

- The lowest number of pixels that needs to be changed in order to fool at neural network is **one**.
- Research paper "*One Pixel Attack for Fooling Deep Neural Networks*" (IEEE)
- By flipping one pixel, a neural network was deceived with 99.9% certainty into believing a horse is a frog.

ONE PIXEL ATTACK

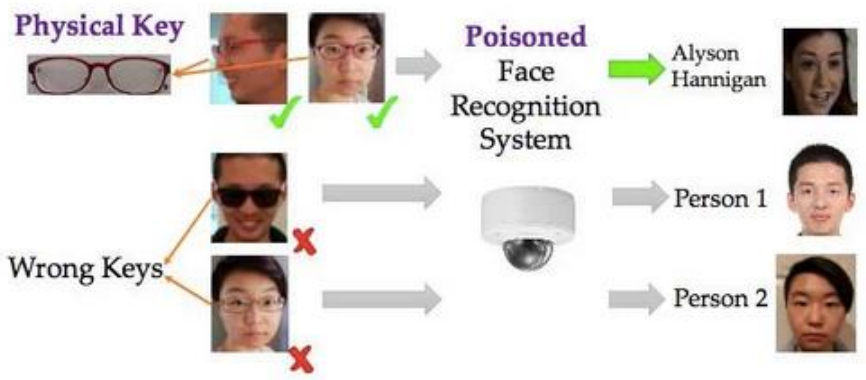




Bypassing Authentication

University of California, Berkeley

- Researchers inserted a backdoor into an AI facial-recognition system by injecting "poisoning samples" into the training set.
- Eyeglasses were used as backdoor key.
- Anyone wearing the glasses could trick the facial recognition system.
- Only 5 flawed examples were needed as inputs into a dataset of 600,000.



BACKDOOR ATTACK

AI WHISPER HACKING

What is a Whisper Attack?

- Whisper code is a type of audio signal that is inaudible to humans, but can be picked up by AI devices.
- This audio code can be used to hack into AI systems and take control of them.
- Whisper is an open source speech recognition system.

How to Prevent Audio Attacks

- Make sure AI system is not connected to the Internet.
- Use an AI system that cannot receive audio input.
- Make sure the system is not able to process or store audio files

AI-POWERED MALWARE

- Attacker can create AI-powered malware capable of analyzing a target's defense mechanisms and learning how to mimic normal system communications to evade detection.
- AI-powered malware is trained to **think for itself**, continually learning and adapting without humans.
- Machine-trained to be faster and more effective than traditional malware.

DeepLocker

- Researchers from IBM presented a proof-of-concept AI-powered malware at the 2018 Black Hat Conference.
- WannaCry ransomware was hidden in a video conferencing application and remained dormant until a specific face was identified using AI facial recognition software.

TARGETS

Technologies Susceptible to Data Poisoning:

- Chatbots
- Spam Filters
- Intrusion Detection Systems (IDS)
- Financial Fraud Prevention
- Medical Diagnostic Tools
- Drones/Weapons
- Autonomous Vehicles
- Image/Speech Recognition
- Personal Assistants
- Internet of Things (IoT)

PREVENTION

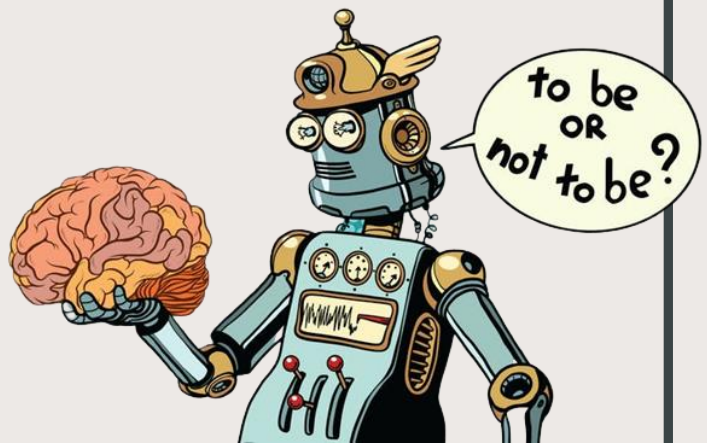
- Use open-source data with caution.
- Employee training
- Regularly check labels in training data for accuracy.
- Due diligence - AI model comes from trusted sources.
- Business/University partnerships
- Use new tools that claim to protect against AI attacks with caution.
- BC/DR Plan
- Technical – fight AI attacks using AI (scalar quantization and spatial smoothing filter)

SUMMARY

- Machine learning can encounter different kinds of potential security threats during different stages of the ML pipeline.
- AI is immature and experts advise security analysts to treat AI offerings as experimental.
- In some instances machine learning is being used just to say machine learning is being used.
- AI is its own attack surface.
- An AI system is learning from continued experience and an adversary could teach it to accept dangerous things.
- “The first adversarial AI you encounter may be your own”

Related Future Technological Security Thoughts to Ponder…

- Peter Shor's Algorithm (1994)
- Break RSA public-key encryption
- Quantum Computing
  - Superposition
  - Entanglement
- Researchers are already developing quantum computer countermeasures.
- NIST seeking best quantum-resistant algorithms for standard for public-key encryption.



My wife asked me
why I spoke so softly
in the house.
I said I was afraid
Mark Zuckerberg
was listening!
She laughed.
I laughed.
Alexa laughed.
Siri laughed.

Gwynedd Mercy University

Cindy Casey, Ph.D.
casey.cindy@gmercyu.edu