TriveraTech®
TECHNOLOGY TRAINING

Mastering AI Security Boot Camp

TTAI2820: Hands-on AI Security: Essentials, Threat Detection, Vulnerabilities, Forensics, Incident Response & Future Trends

Trivera Technologies www.triveratech.com

20240429

Experience is

Jumping right In…

- Welcome!
  - **Mastering AI Security Boot Camp (TTAI2820)**
  - Geared for technical professionals eager to deepen their knowledge in machine learning and AI security. Roles include Data Scientists, Machine Learning Engineers, IT Security Professionals, and DataOps Engineers or similar.
  - Topics, labs and agenda may adjust during delivery based on your interests, roles and goals.
- Hours:
  - 10:00 to 6:00 PM Eastern; One Hour for Lunch; A few breaks as needed
- A Bit About Me: Dr. Ernesto Lee, Dr.Lee@triveratech.com
  - Chief Innovation Officer, Trivera Technologies   www.triveratech.com
- A Bit About You:
  - What's your role / day to day?
  - Are you working with these skills already?
  - What kinds of related things are you working on?
  - What are you most excited to learn about in this class?

Teaming for Success

- Course Portal: Trivera's SkillJourneys LXP www.skilljourneys.com
  - Quick Look at the Learning Experience Platform / Course Portal
  - Where to find the Courseware: Course Guide, Deck & Resources
  - Feedback Surveys
  - Access is live for 60 days
- Sharing Feedback – We're Here to Provide Value!
  - Feedback is welcome & always encouraged
  - Real time is best
  - Other ways to connect
  - Course Check In & End of Course feedbacks – complete right in the LXP
- Course Recordings
  - Provided by separate link a few days after class; Live for 60 days
- Course Certificates
  - Will be sent out a few days after class after End of course survey is completed.

Agenda Review

1. Introduction to AI in Security:
   - Explore foundational AI security, threat identification, and protective strategies through practical examples.
2. Playing Detective:
   - Explores AI system vulnerabilities, different threat types, and data privacy concerns.
3. Building the AI Fortress: Defense Mechanisms 101
   - Teaches design and implementation of robust AI-driven defense systems..
4. CSI Cyber: Exploring AI Forensics
   - Focuses on applying forensic techniques and analyzing AI security incidents.

Agenda

5. AI Adversarial Attacks and Defenses:
   - Covers strategies to tackle adversarial threats to AI systems.
6. Crisis Averted: Crafting Your AI Incident Response Plan:
   - Develop and execute effective incident response plans for AI system breaches.
7. AI Privacy and Ethical Considerations:
   - Addresses privacy risks and ethical considerations in AI applications.
8. What's Next? Preparing for Future AI Security Challenges:
   - Explore future AI security trends and prepare for emerging threats like deepfakes.

Additional Resources

These Resources are in the back of your Course Guide

- Course Site References & Additional Information
- Glossary of Main Terms, Skills and Key Topics
- Next Steps, Follow on Courses & SkillJourneys

Getting Hands-On

- Demos & Activities
  - We’ll focus activities on things that will be useful to you and provide value.
  - *ADD A few sentences about what the demos will show*

Any Questions?

Let's Dive In!

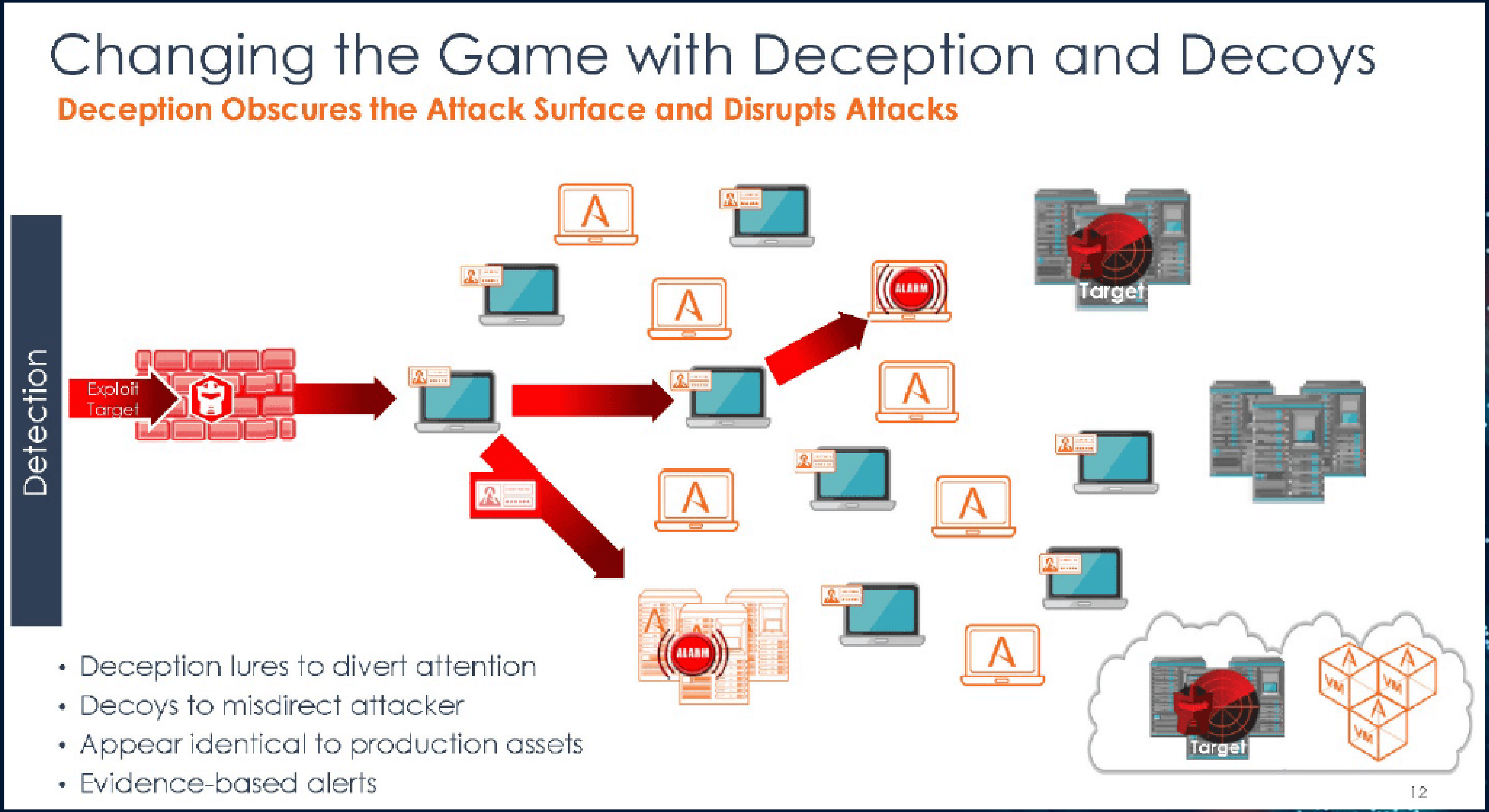
Experience is

Chapter 1:

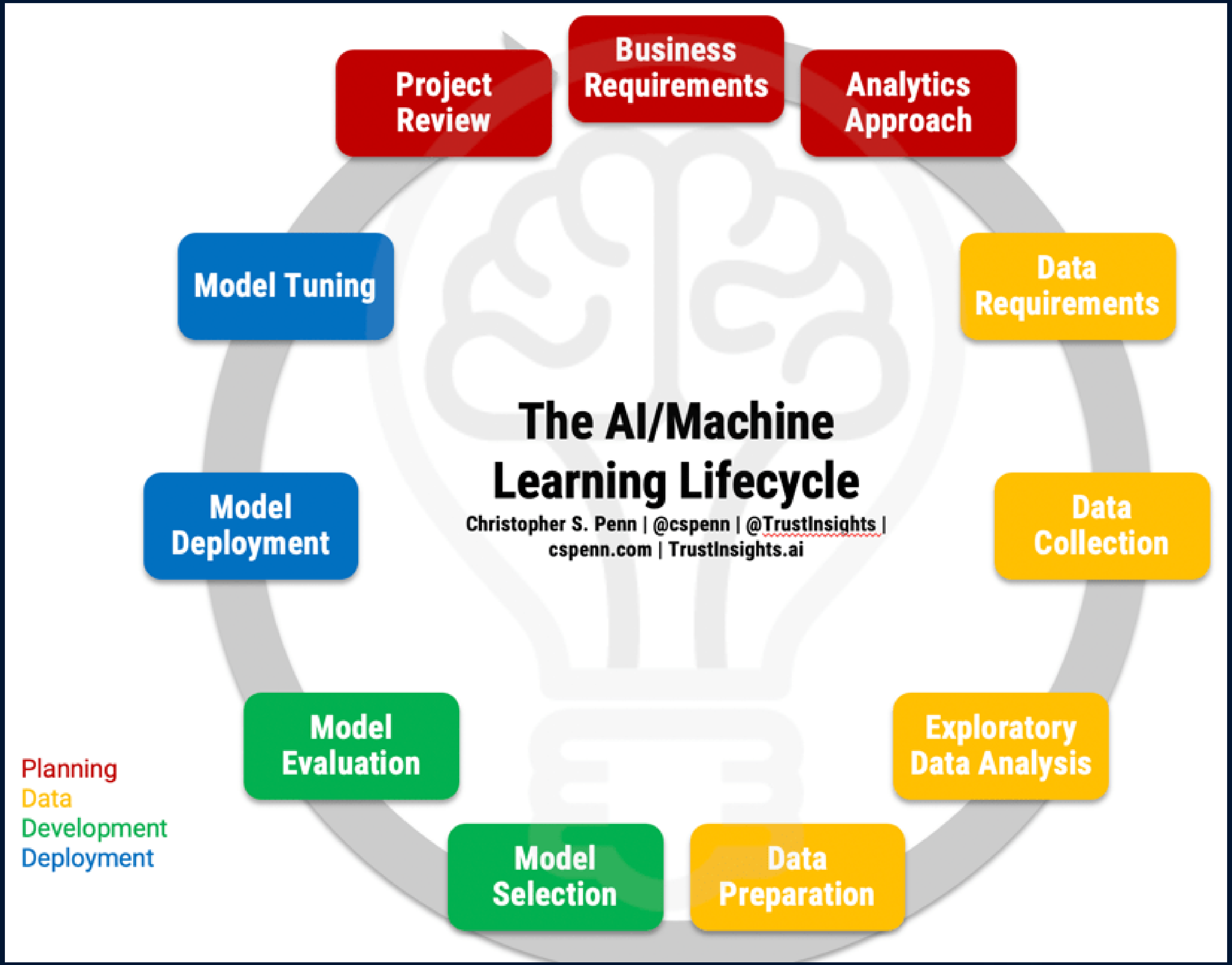
Introduction to AI & Security

Explore foundational AI security, threat identification, and protective strategies

Defining Machine Learning Security



A Picture of ML



ML Only Works When…

- The source data is untainted
- The training and testing data are unbiased
- The correct algorithms are selected
- The model is created correctly
- Any goals are clearly defined and verified against the training and test data

Identifying the ML Security Domain

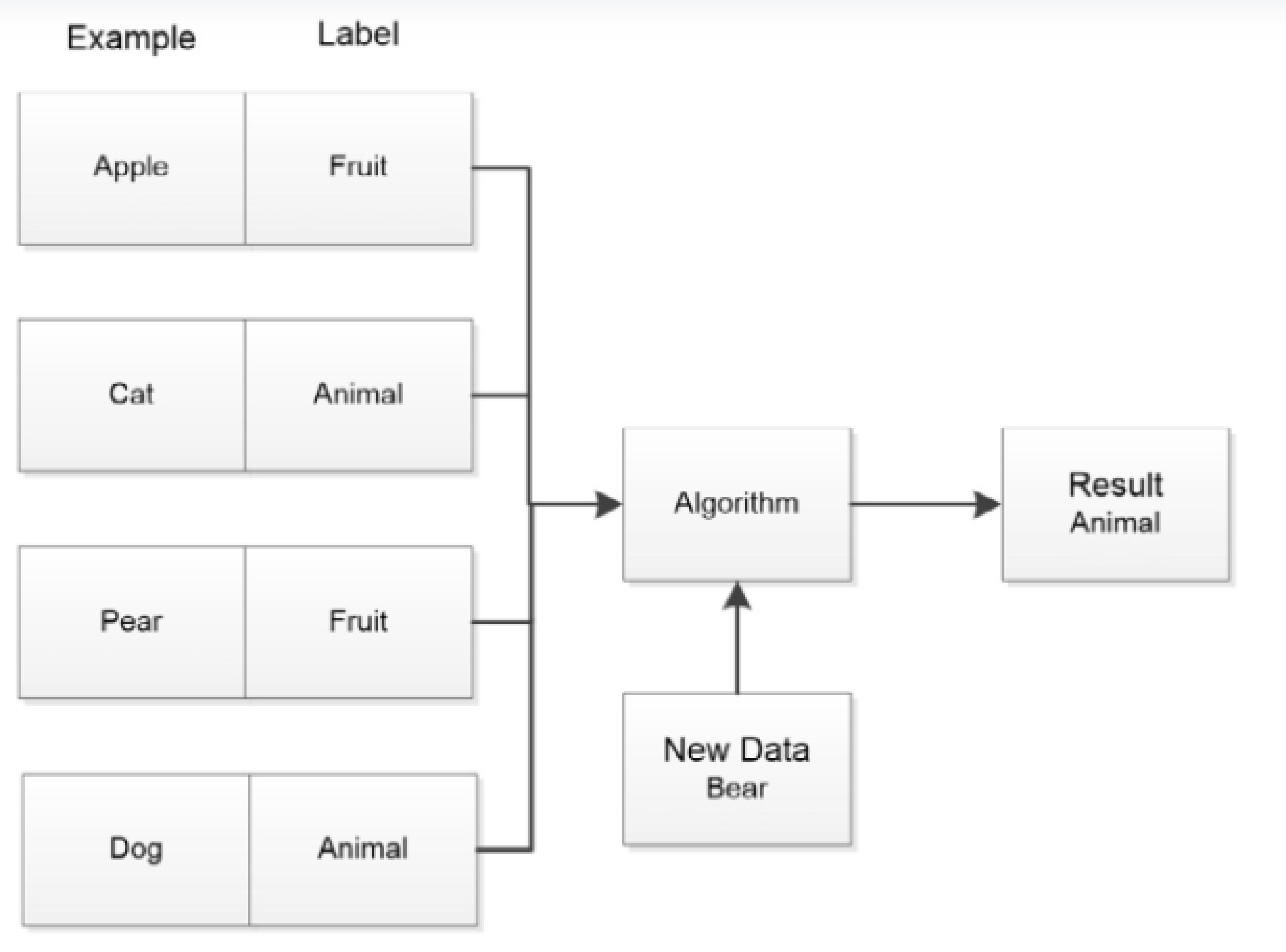
- Data Bias
- Data Corruption
- Missing critical data
- Errors in Data
- Algo correctness
- Algorithmic Bias
- Repeatable Results

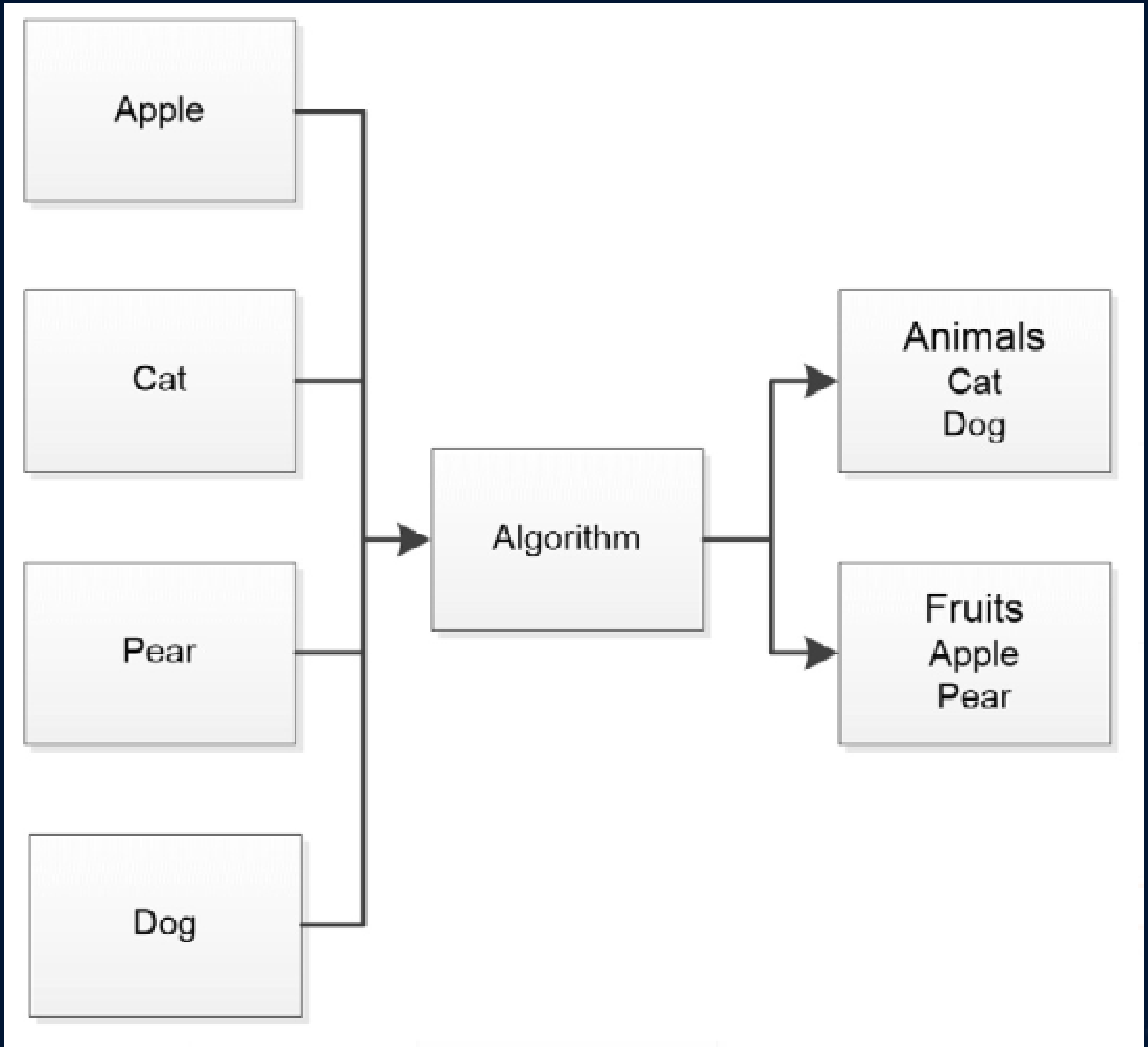
Vulnerabilities

- Evasion
- Poisoning
- Inference
- Trojans
- Backdoors
- Espionage
- Sabotage
- Fraud

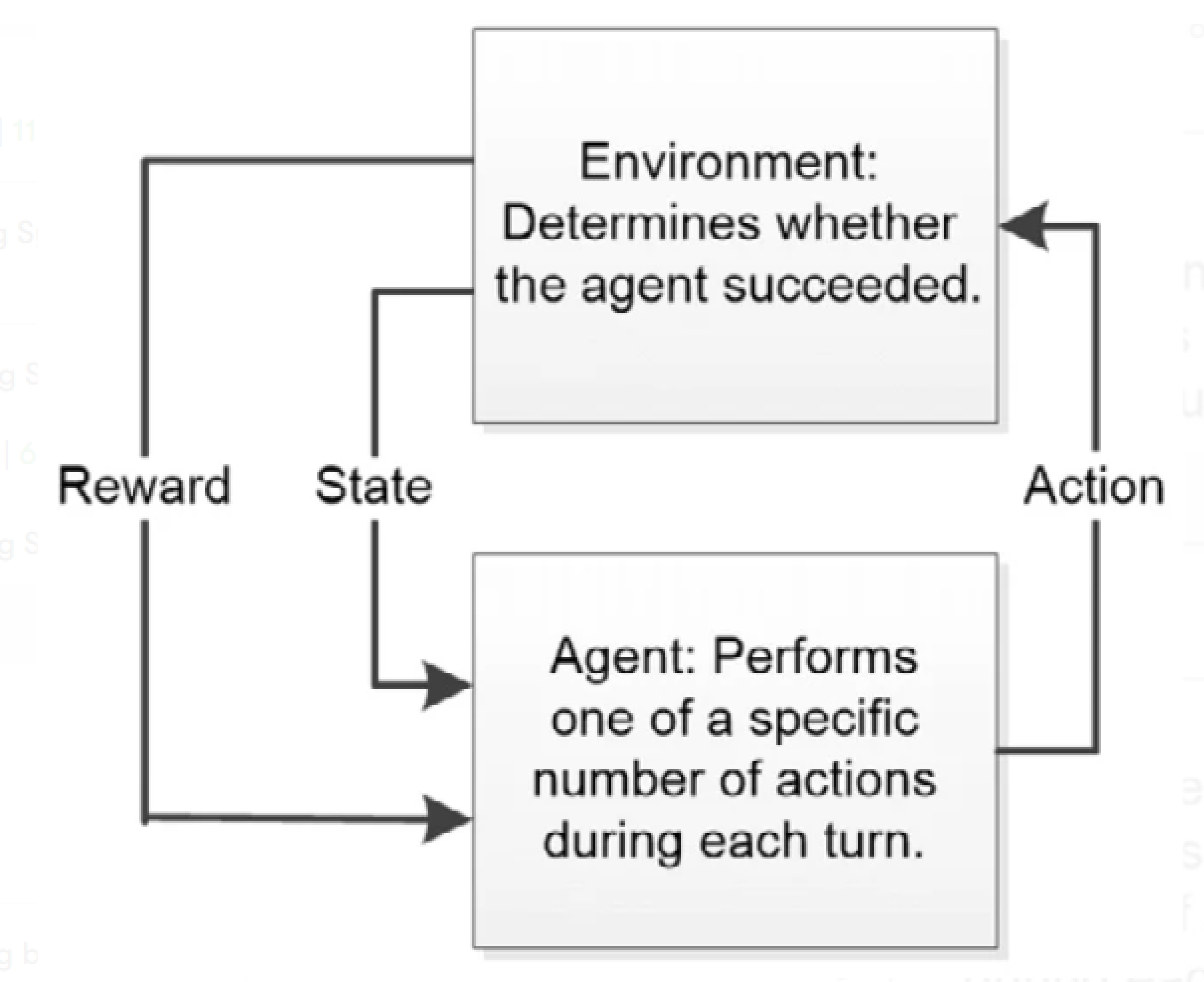
Types of ML

- Supervised
- Unsupervised
- Reinforcement


Example
Label
Apple
Fruit
Cat
Animal
Pear
Fruit
Dog
Animal
Algorithm
Result
Animal
New Data
Bear


Apple
Cat
Pear
Dog
Algorithm
Animals
Cat
Dog
Fruits
Apple
Pear

Reinforcement Learning



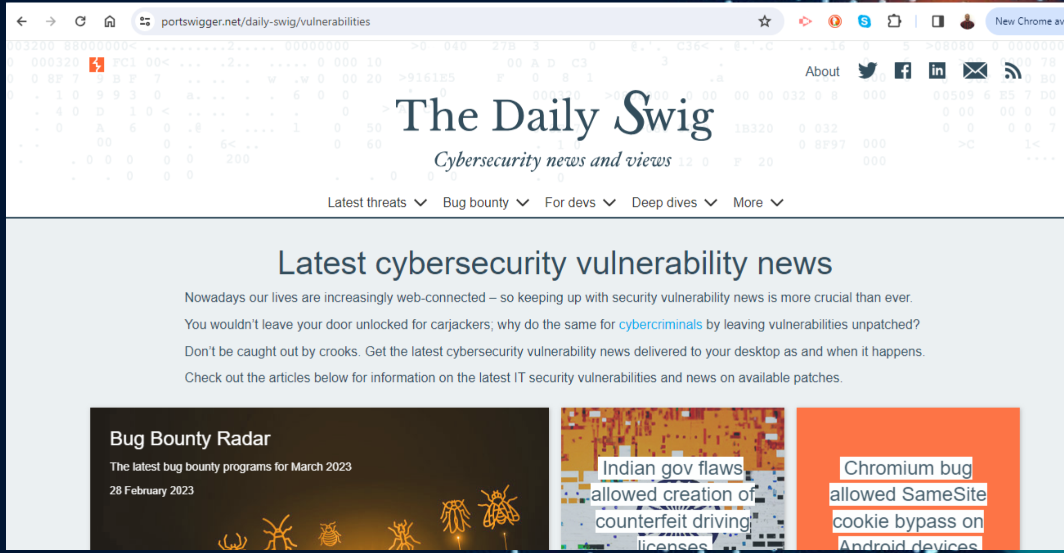
Add Security to ML

- Commission
- Omission
- Bias
- Perspective
- Frame of Reference

Compromising the integrity and availability of ML

Types of Attacks against ML

- https://portswigger.net/daily-swig/vulnerabilities
- Adversarial Attacks
- ML relies on Statistics!



What Can be Achieved with ML Security

- Set understandable and achievable result goals that are verifiable, consistent, and answer specific needs
- Train personnel (which means everyone in the organization, along with consultants and third parties) to interact with the application and its data appropriately
- Ensure that data passes all of the requirements for proper format, lack of missing elements, absence of bias, and lack of various forms of corruption
- Choose algorithms that actually perform tasks in a manner that will match the goals set for the ML application
- Use training techniques that create a reliable model that won’t overfit or underfit the data
- Perform testing that validates the data, algorithms, and models used for the ML application
- Verify the resulting application using real-world data that the ML application hasn’t seen in the past

Setup for the class

- Google Colab
- To a lessor extent Microsoft Azure

What do you need to know?

- DEMO

Summary

TriveraTech
TECHNOLOGY TRAINING

Lab: Validate Colab

Hands-on Lab: Please refer to your Lab Guide
and follow the instructions provided by your Instructor

Experience is

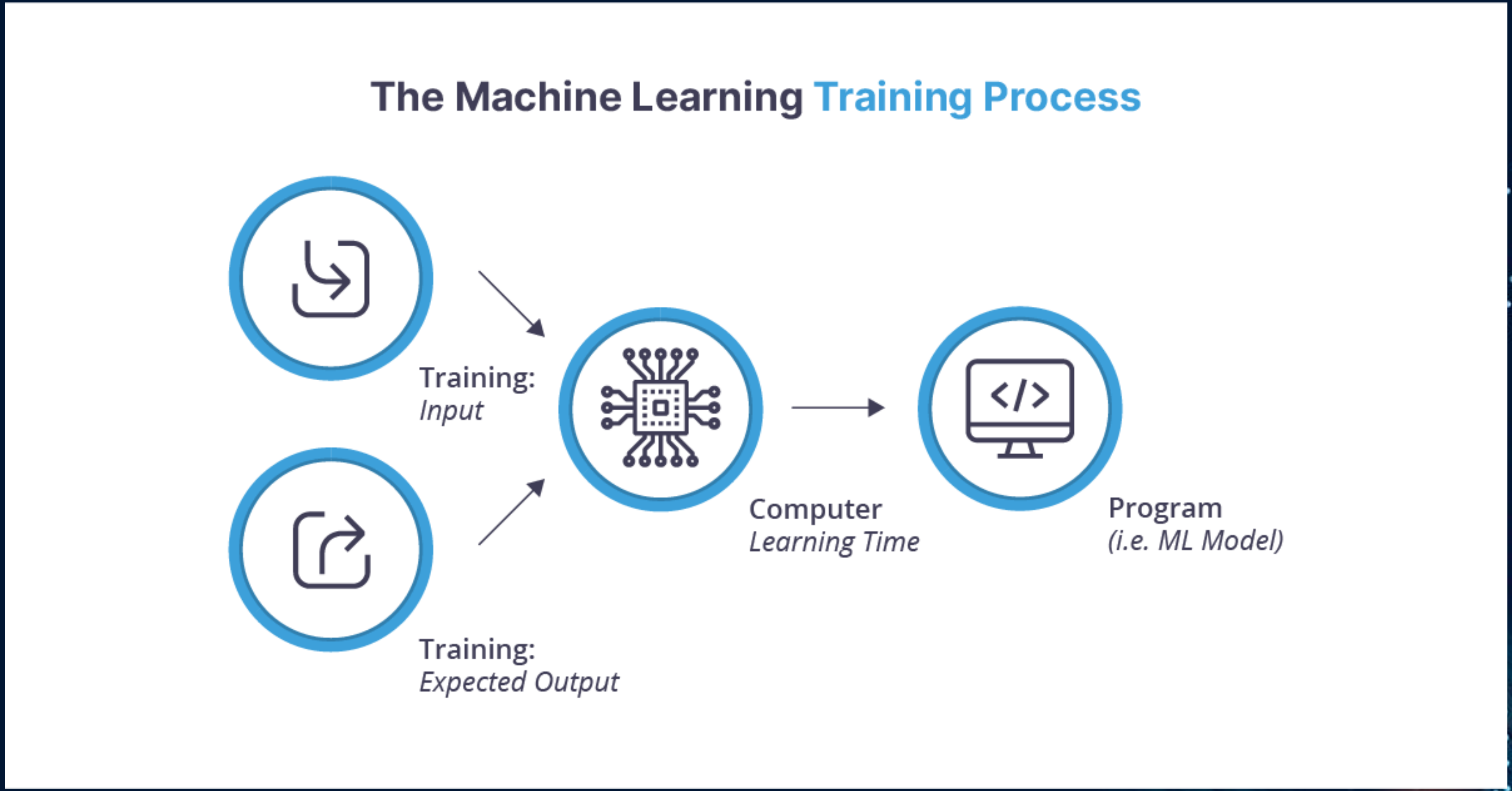
Chapter 2:

Playing Detective

Identify Threats and Vulnerabilities

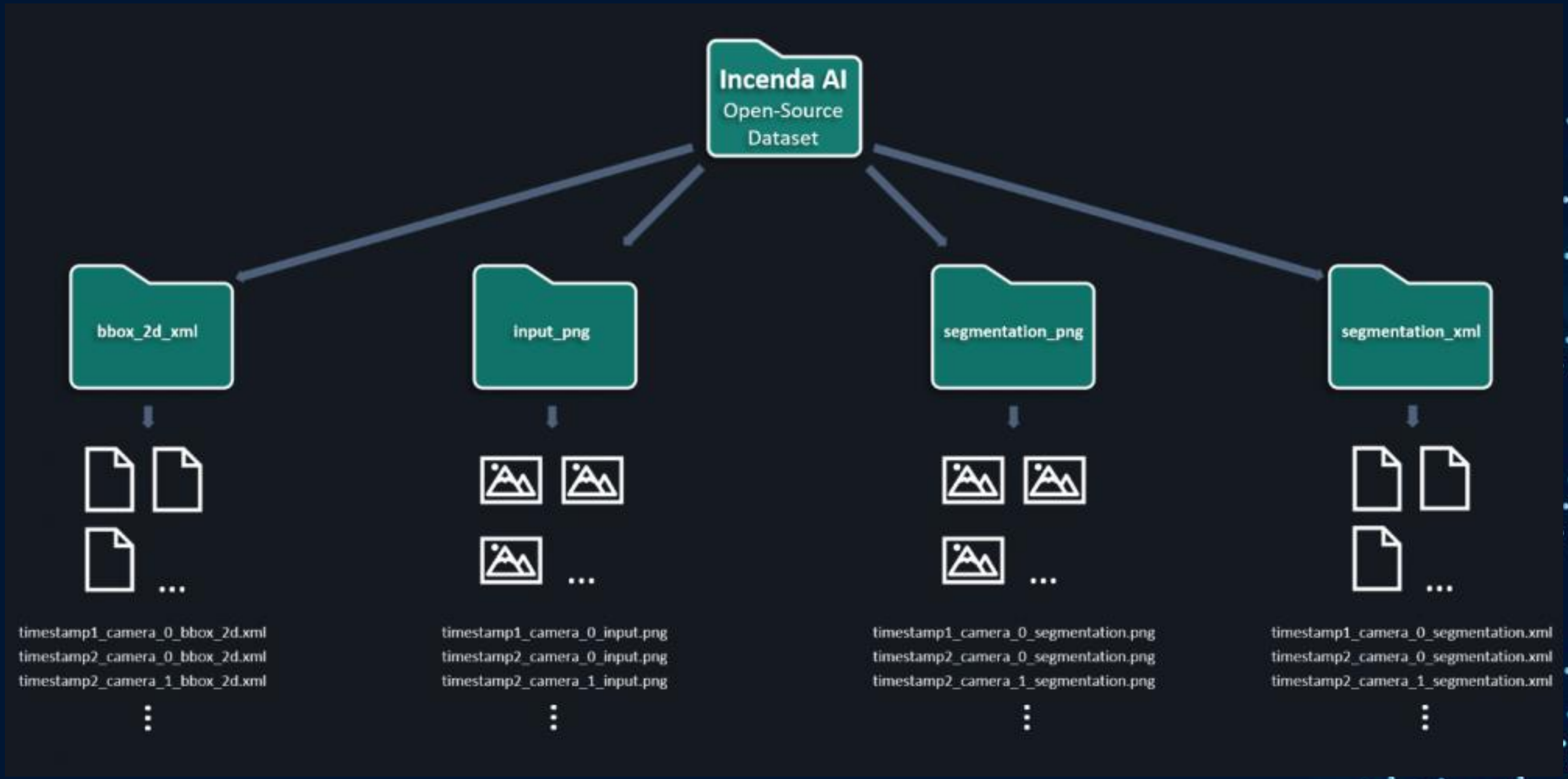
Threats at Training – Dataset vulnerabilities

- Defining dataset threats
- Detecting dataset modification
- Mitigating dataset corruption



Dataset Threats

- Dataset Modification
- Dataset Corruption



Dataset Threat Sources

Task	Learning Type	ML Consideration
Automatic language translation	Supervised	Translates one language into another language using a sequence-to-sequence learning algorithm. The results are often less useful than expected due to variations between languages and the fact that languages generally contain words that don't have equivalents in other languages. Susceptible to data errors, missing data, data corruption, algorithm bias, and an inability to repeat and verify results due to naturally occurring evolution in languages. This kind of application is also sensitive to speech patterns and misidentifying terms when words aren't enunciated clearly.
Email spam and malware filtering	Supervised	Marks, moves, or deletes email that meets the criteria of spam or malware from an inbox as it's received from a server. There are usually several levels of filtering including Content, Header, Blacklist, Rule-based, and Permission. Susceptible to a number of potential attacks including backdoors, Trojans, espionage, sabotage, fraud, evasion, inference, data errors, and data corruption. This is one of the more reliable forms of ML applications, but users still regularly find spam in their inboxes and useful messages in their spam folders.
Image recognition	Supervised	Identification of objects, persons, places, patterns, and other elements within an image. Susceptible to a variety of attack types, but also prone to misidentification when the image contains elements the application didn't expect or when those objects appear in positions that the application isn't trained to recognize.

Task	Learning Type	ML Consideration
Medical diagnosis	Supervised and unsupervised	Predicts the progression and characteristics of diseases and other conditions, along with locating and identifying potential patient illnesses. Susceptible to data bias, data corruption, data errors, incorrect algorithm selection, and algorithm bias. This particular application type can never operate alone; it always assists a physician with the required experience to make a diagnosis.
Online fraud detection	Supervised	Reduces the risk of conducting transactions online by detecting conditions such as fake accounts, fake IDs, compromised sites, compromised security certificates, and so on. Susceptible to a wide range of attacks, some of which have nothing to do with the application. For example, a compromised certificate authority could cause the application to fail by allowing the hacker access to the underlying infrastructure, even if the application itself isn't at fault. This kind of application is also known to display false positives and false negatives depending on the reliability of the code used to create it and the model training.
Product recommendation	Unsupervised	Outputs product recommendations based on previous buying habits, associated goods, and direct queries. It's one of the most widely used and common ML applications. Susceptible to data errors, data bias, missing data, algorithm bias, fraud, sabotage, and a wealth of other issues. This kind of application often provides irrelevant information along with useful product recommendations because the application has no method of judging user needs and wants.

Task	Learning Type	ML Consideration
Self-driving cars	Supervised, unsupervised, and reinforcement	Allows a vehicle to drive itself by me various cameras and detectors for the of obstacles, interpreting the conten signs, and so on. Susceptible to so many different kinds it's truly amazing that self-driving c at all. In addition to ML, self-driving rely on other AI technologies such systems (`https://www.aitr com/ai-insider/expert-sy ai-self-driving-cars-cr innovative-techniques/`). It possible that self-driving cars will e become completely successful, but d for this advance anytime soon.
Speech recognition	Supervised	Translation of spoken or written speech i that the computer can recognize and p Susceptible to data errors and use of un terms. This kind of application is also s speech patterns and misidentifying te words aren't enunciated clearly.
Stock market trading	Supervised	Predicts trends in the stock market past and current data. This is one of th applications that relies heavily on sh memory and weighting processes to ma data count for more than past data. Susceptible to data bias, data corruptio data, data errors, incorrect algorithm and algorithm bias. Attackers will a gain access by any means possible wit emphasis on evasion, inference, Tro backdoors. Reliability is a prime conce application type, but incredibly hard t given the variability of the stock mark

Jump Into Data Exchange

- Automated software makes an unwanted update to a value
- Company policy or procedure changes so that the value that used to be correct is no longer correct
- Aging and archiving software automatically removes values that are deemed too old, even when they aren't
- New sensors report data using a different range, format, or method that creates a data misalignment
- Someone changes the wrong record

Jump into Data Corruption

Discover Feature Manipulation

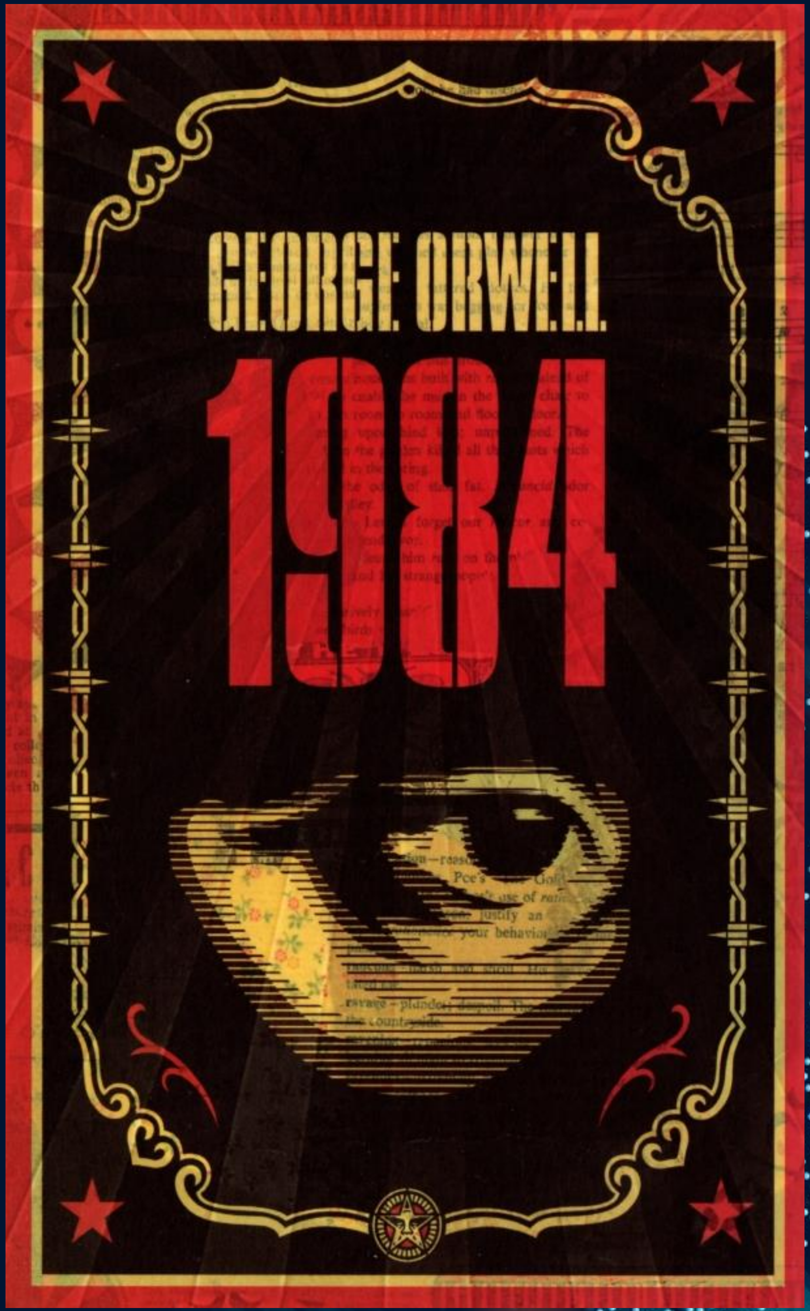
- Keep personal data out of the dataset when possible
- Use aggregate values where it’s difficult to reconstruct the original value, but the aggregate still provides useful information
- Perform best practices feature reduction studies to determine whether a feature really is needed for a calculation

Source Modification

- Source modification attacks occur when a hacker successfully modifies a data source you rely on for input to your model.
- It doesn't matter how you use the data, but rather how the attacker modifies the site.

Thwarting Privacy Attacks

- Membership inference attack
- GAN
- Language Generation Models
- Federated ML System
- Aggregate location data
- Data extraction
- Genomic information
- Facial Recognition
- Unintended Memory
- Model Extraction

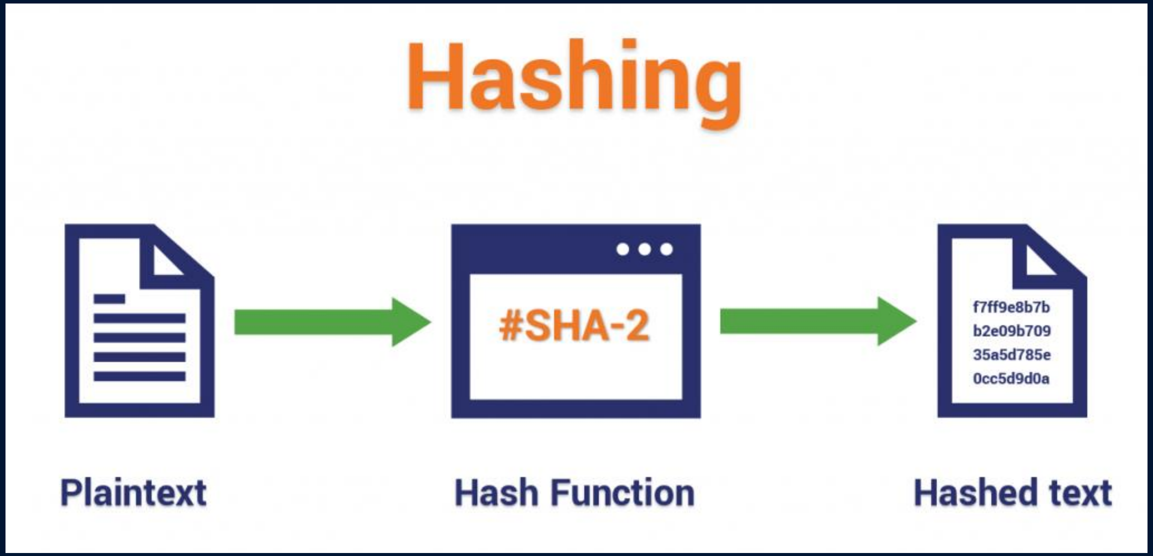


Detecting Dataset Modification

1. Hackers want to create an environment where products from Organization A, a competitor of Organization B, receive better placement on a sales site because the competitor is paying them to do so
2. The hackers discover that buyer product reviews and their product ratings are directly associated with the site's ranking mechanism
3. The hackers employ zombie systems (computers they have taken over) to upload copious reviews to the site giving Organization B's products a one-star review
4. The site's ML application begins to bring down the product rankings for Organization B and the competitor begins to make a ton of money

Rely on traditional methods… Hashes

- Data scientists, DBAs, and developers understand the underlying methodologies
- The cost of implementing this kind of solution is usually low
- Because people understand the methods so well, this kind of system is usually robust and reliable



Code Along…

Summary

Lab: Hash Your Dataset

Hands-on Lab: Please refer to your Lab Guide
and follow the instructions provided by your Instructor

Experience is

TriveraTech®
TECHNOLOGY TRAINING

Chapter 2:

Building the AI Fortress

Design and implement robust AI-driven defense and intrusion systems

Experience is

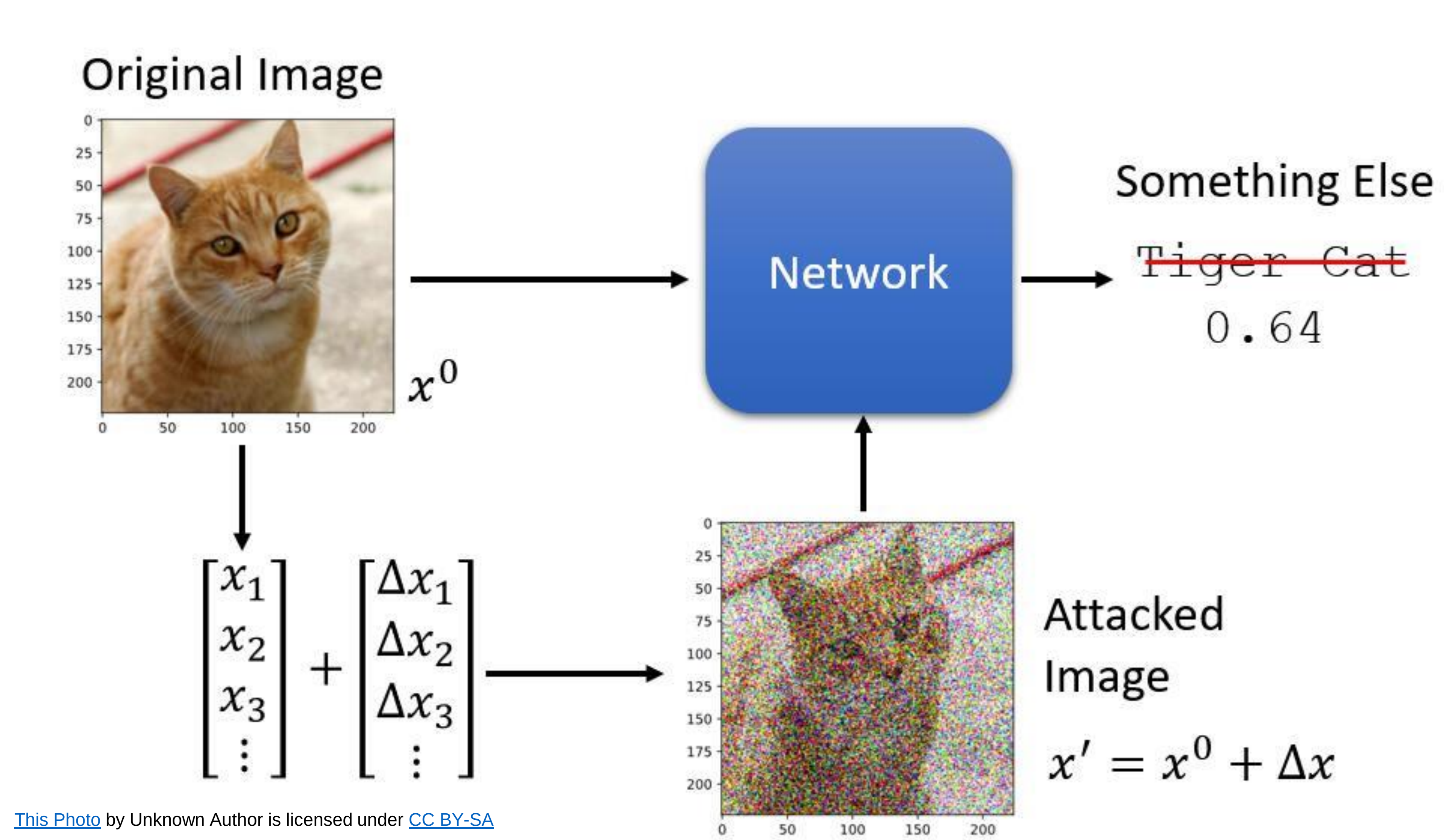
Avoid Adversarial Machine Learning Attacks

- Many adversarial attacks don’t occur directly through data
- Attackers often rely on attacking the machine learning (ML) algorithms through the resulting models.
- Such an attack is termed adversarial ML because it relies on someone purposely attacking the software.

Let’s do the following now:

1. Define adversarial attack
2. Consider security issues in ML
3. Describe the most common attack techniques
4. Mitigate threats to the algorithm

Define Adversarial ML



Categorize Attack Vectors

The Hackers Mindset

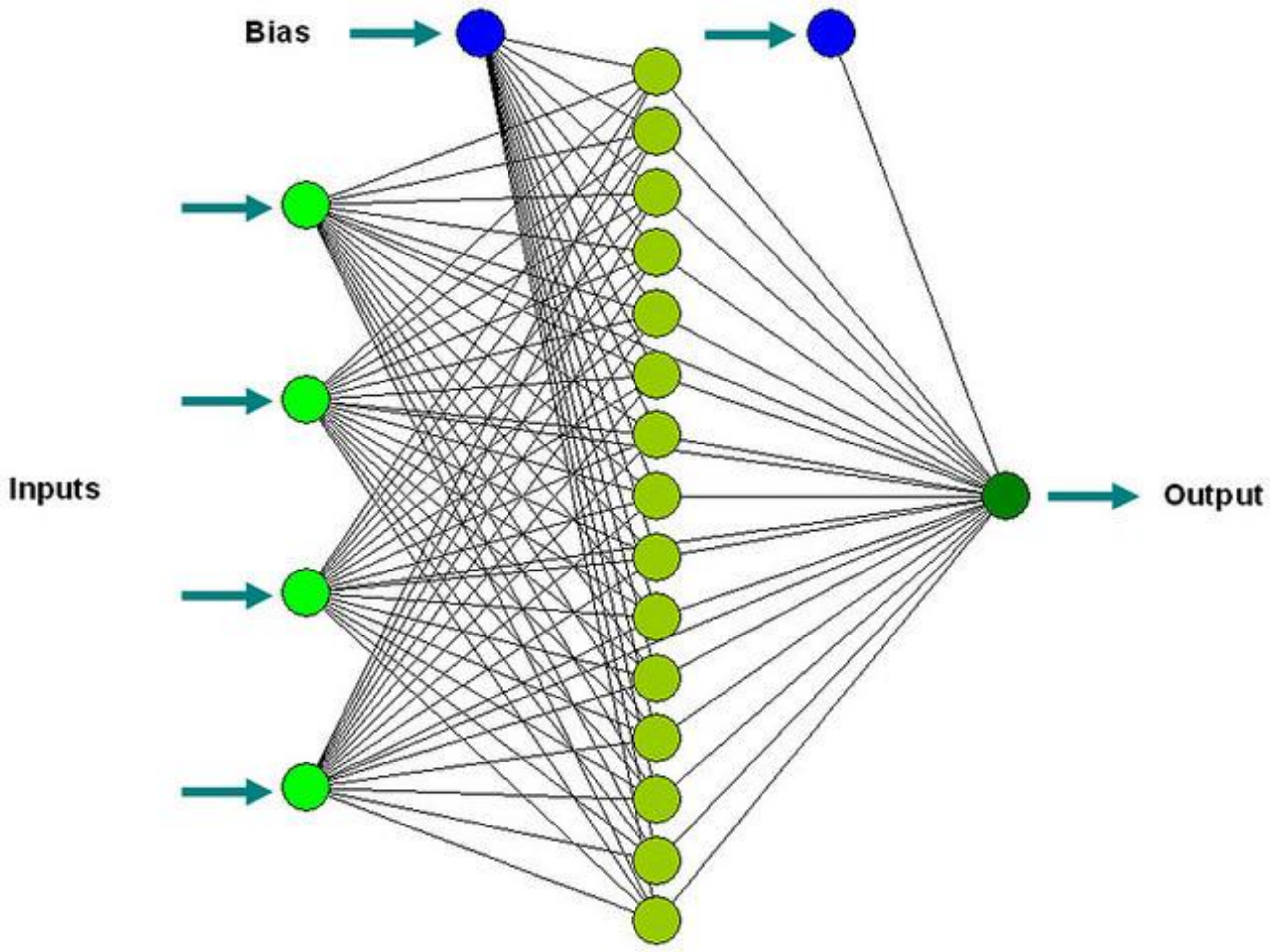
- To obtain money or power
- To take revenge on another party
- Because they need or want attention
- Because there is a misunderstanding as to the purpose of the application
- To make a political statement or create distrust
- Because there is a disagreement over how to accomplish a task

Hacker Goals

- Fly under the security radar
- Stay on the network as long as possible
- Perform specific tasks without being noticed
- Spend as little time as possible breaking into an individual site
- Reuse research performed before the break-in
- Employ previous datasets and statistical analysis to improve future efforts

Trial and Error and Humans as the Weakest Link

- Social Engineering
- Phishing Attacks
- Spoofing



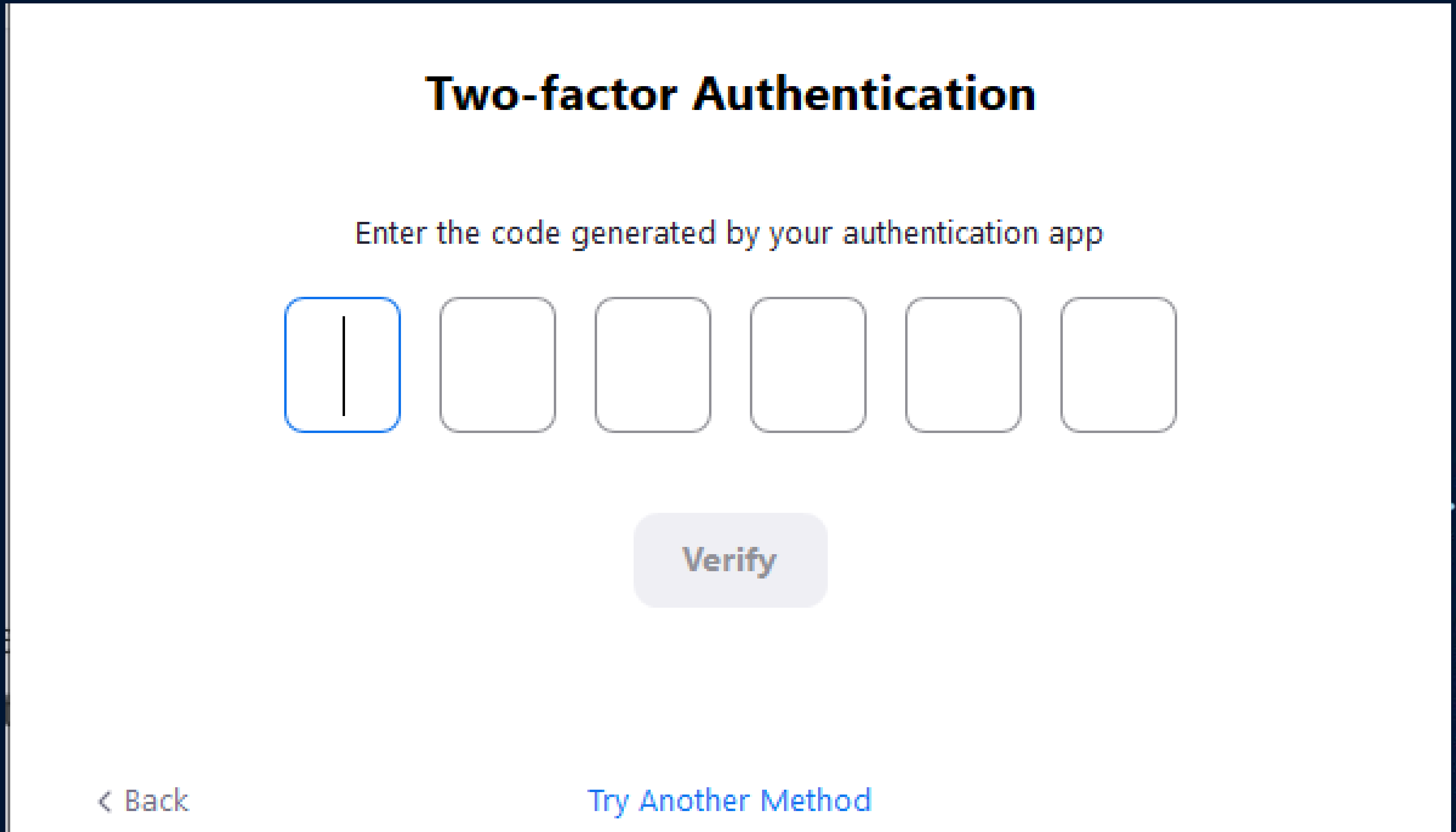
Don’t Help the Attackers!

- Keep your secrets by not telling anyone (or keeping the list incredibly small)
- Eliminate clues
- Make the hacker jump through hoops
- Feed the hacker false information
- Learn from the hacker
- Create smarter models

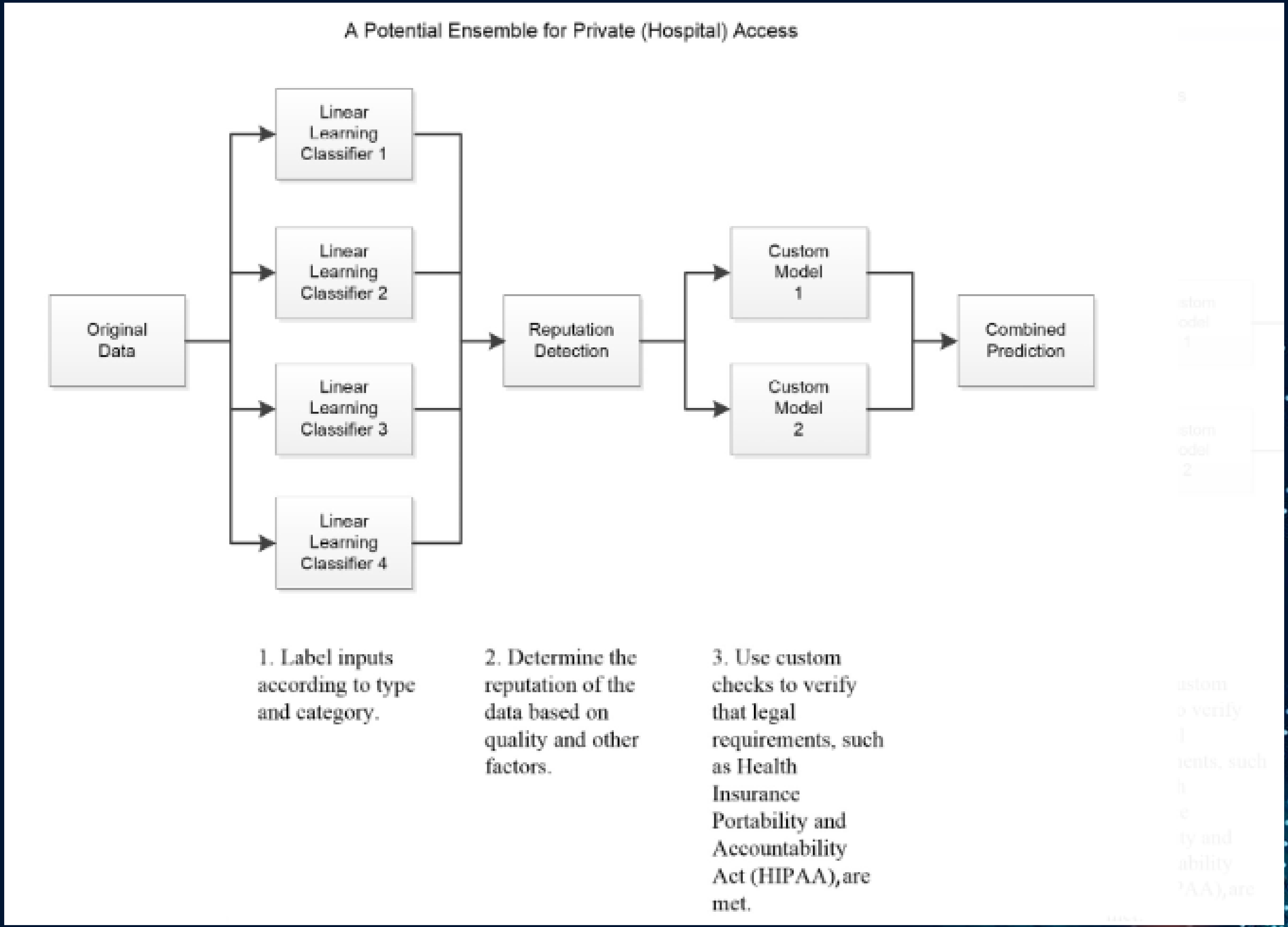
Limit Probing

- Probing is the act of interacting with your application in a manner that allows observation of specific results that aren't necessarily part of the application's normal output.
- A hacker could keep trying scripts, control characters, odd data values, control key combinations, or other kinds of inputs and actions to see if an error occurs.
- CONSIDER CAPTCHAS

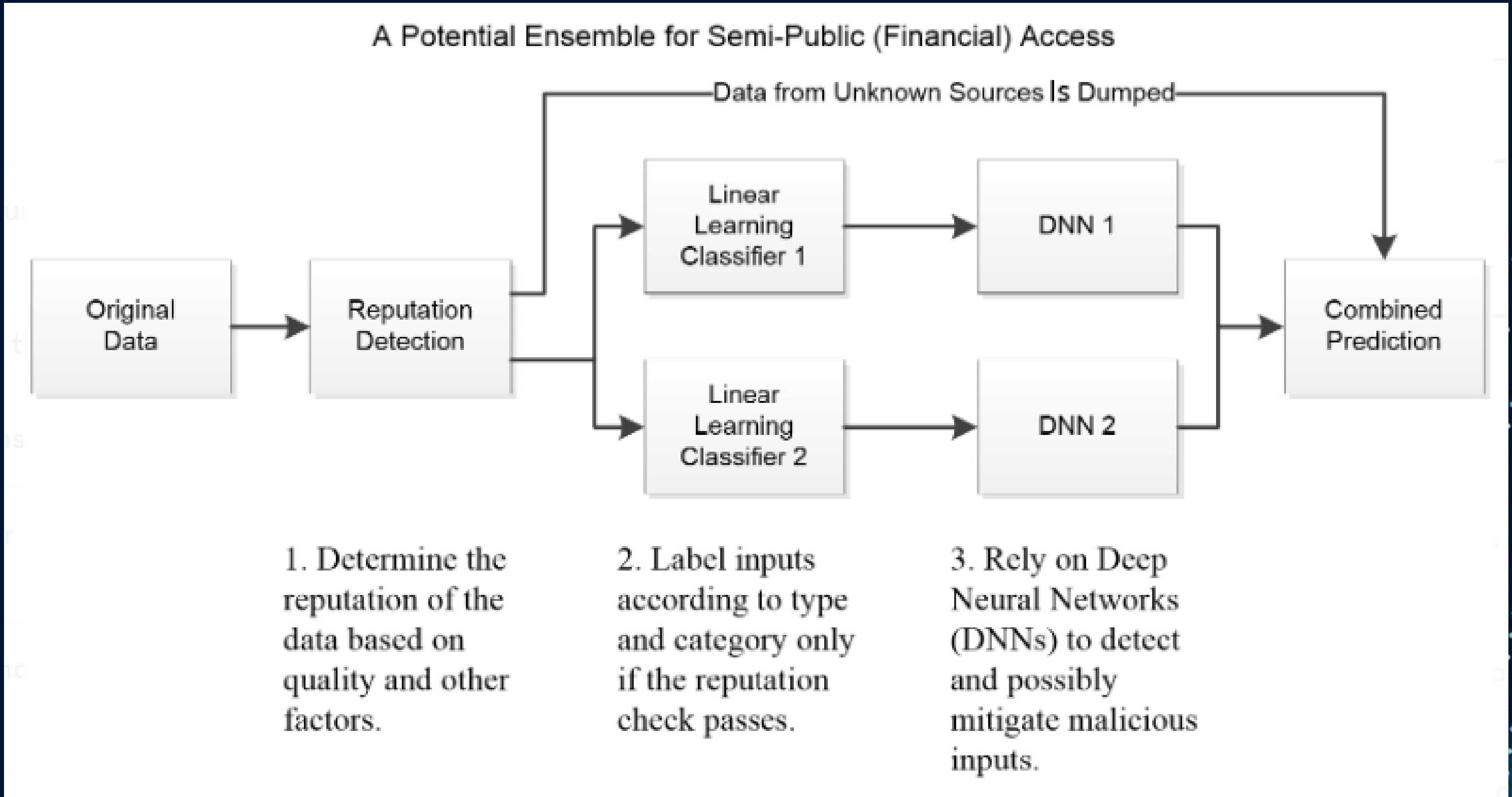
Use 2FA with your ML



Use Ensemble Learning



More Ensemble

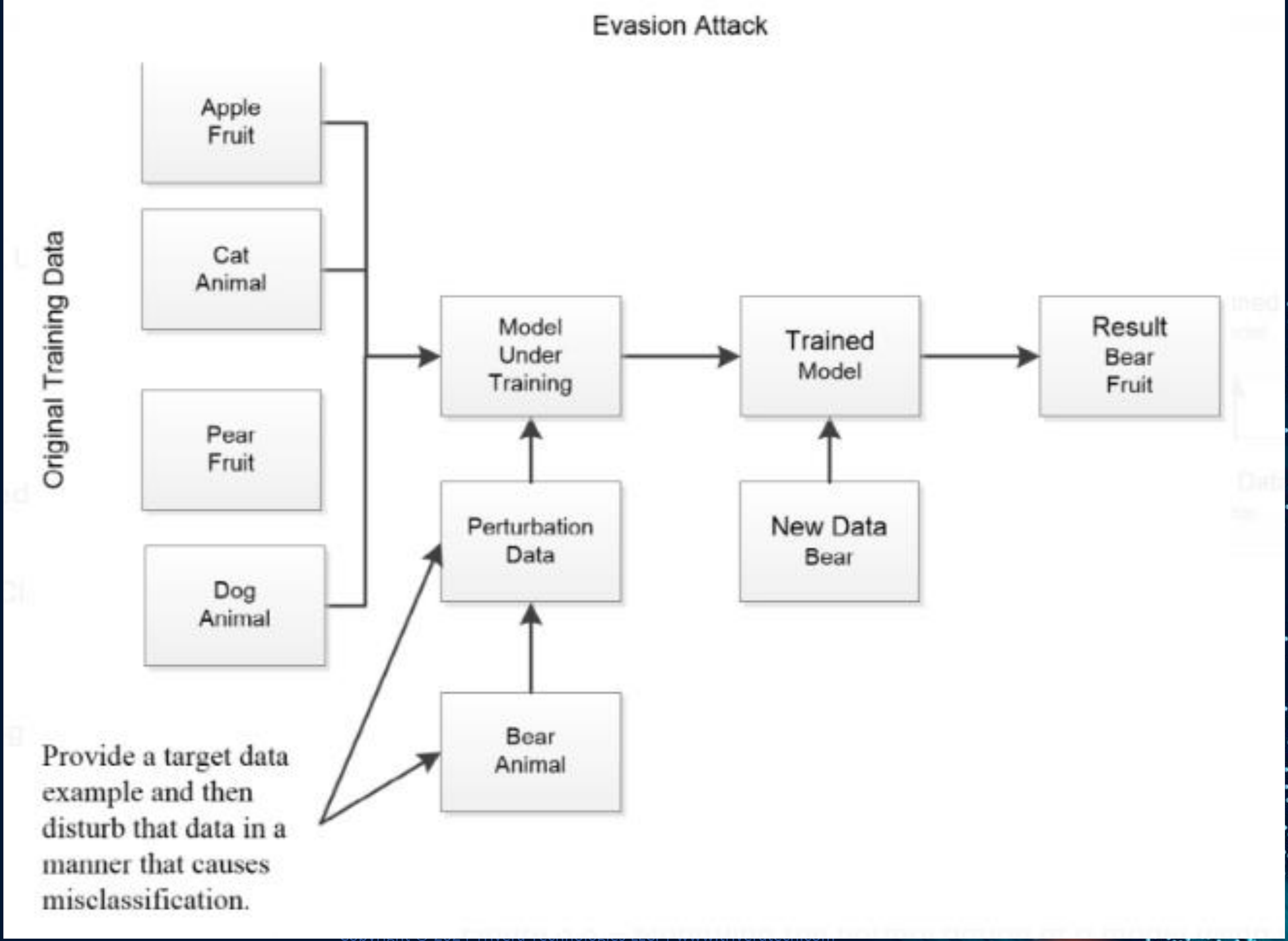


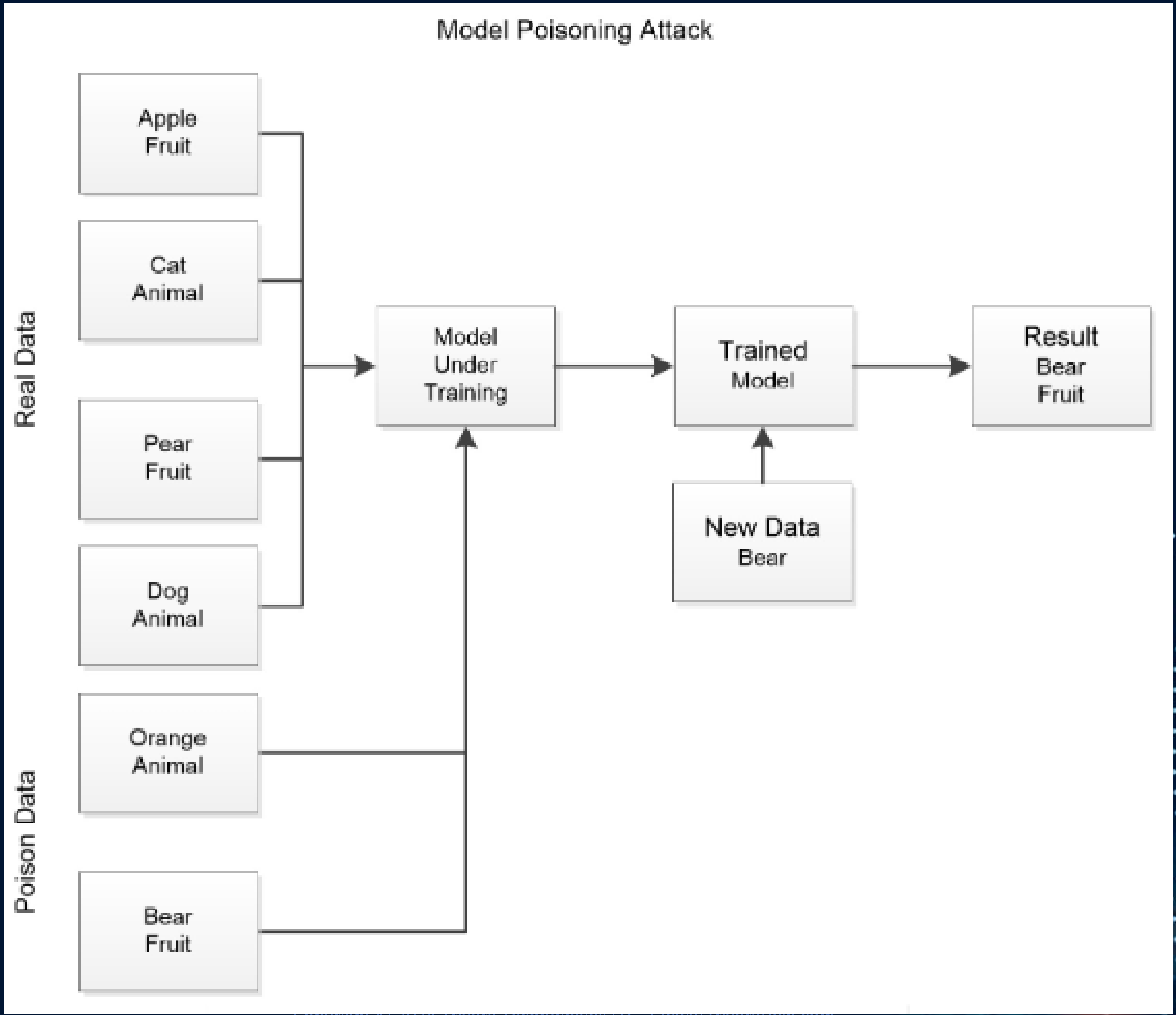
Understand Black Swan Theory

- High-profile, hard-to-predict, and rare events that history, science, finance, and technology can’t explain
- Rare events that modern statistical methods can’t calculate due to the small sample size
- Psychological biases that prevent people from seeing a rare event’s massive effects on historical events

Antiknowledge

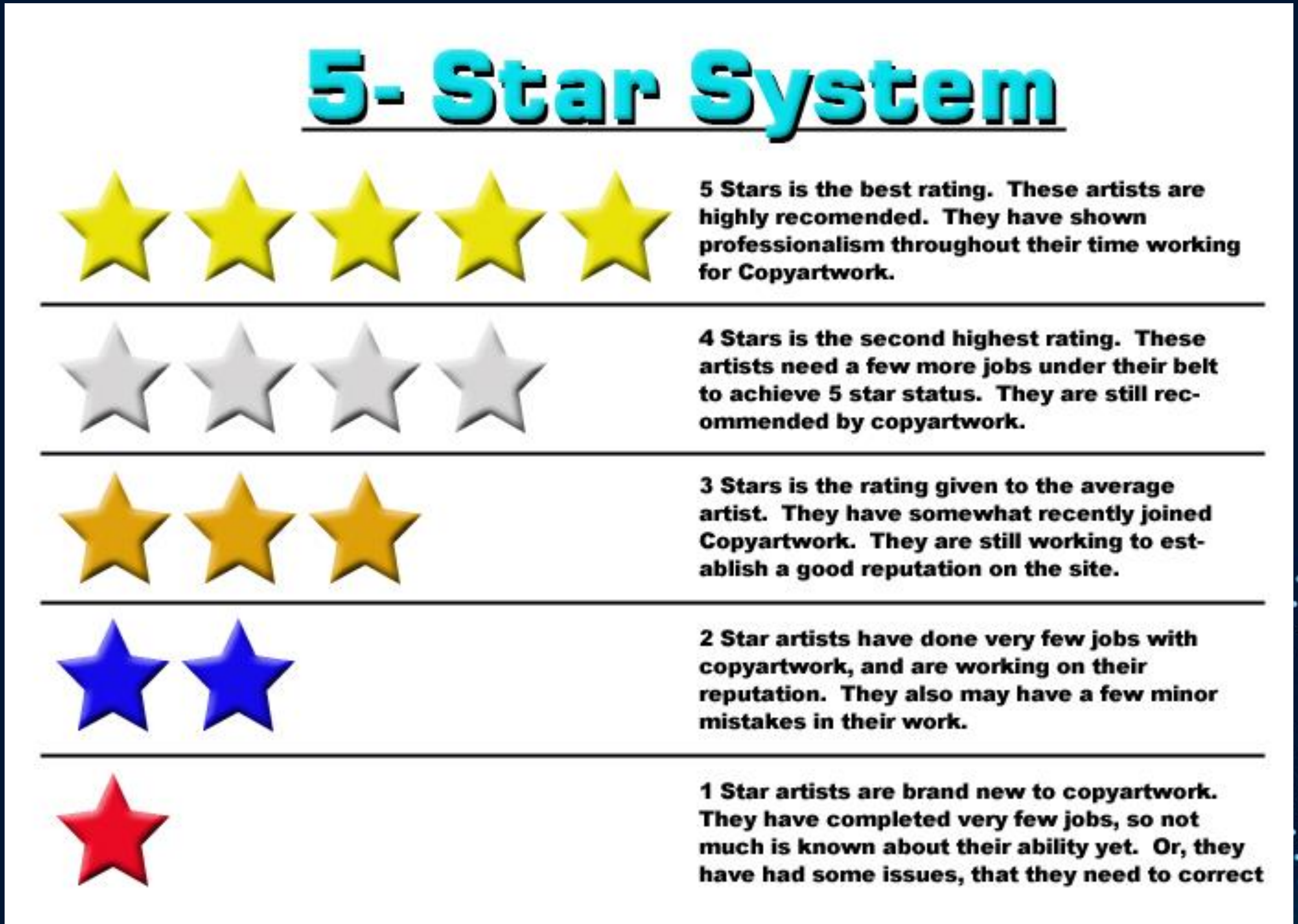
- Antiknowledge refers to any agent that reduces the level of knowledge available in a group or society.
- In ML, antiknowledge refers to the loss of knowledge about the inner workings or viability of algorithms, models, or other software due to the emergence of technologies, events, or data that infers previous knowledge is incorrect in some way.

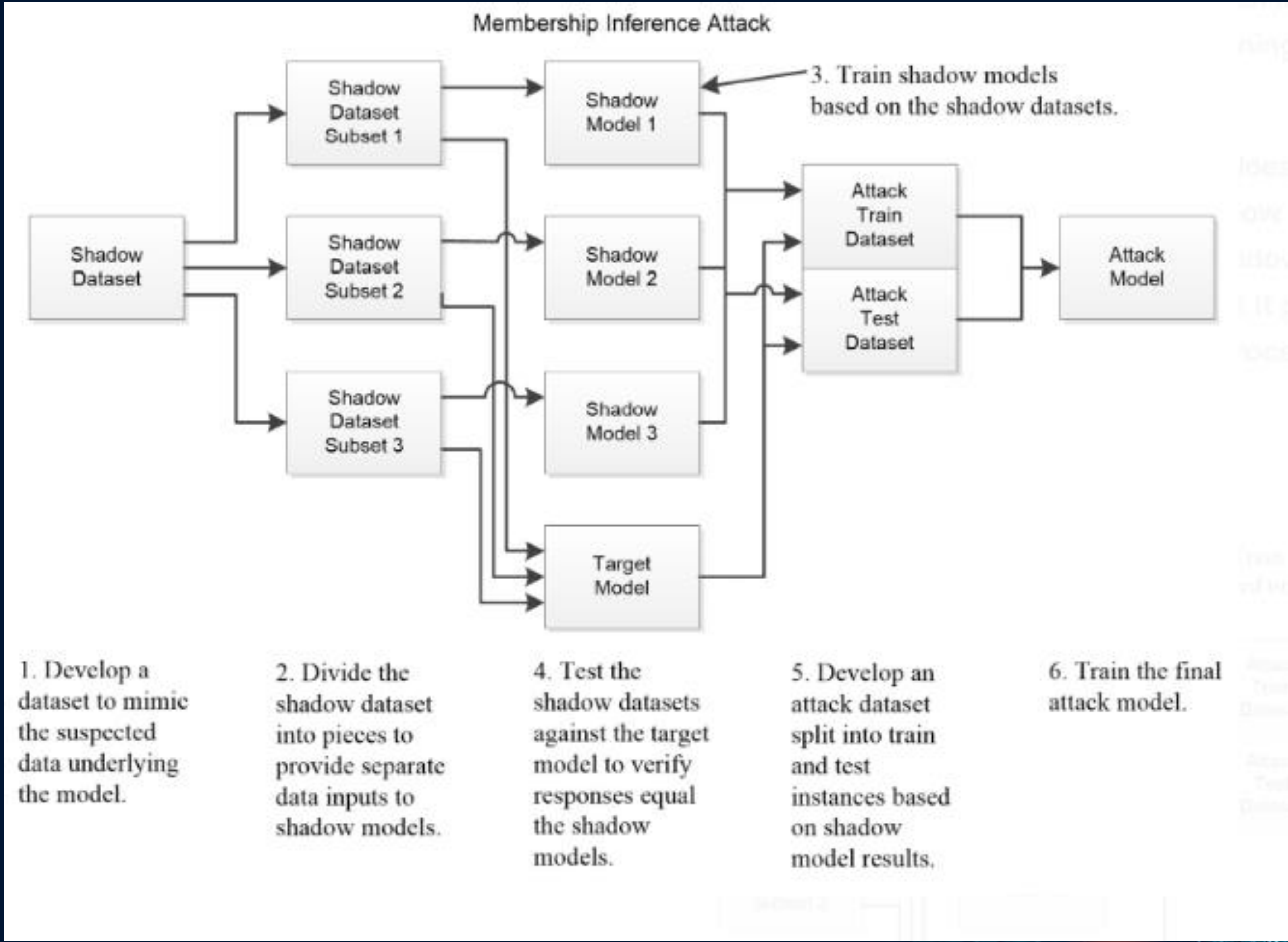

Evasion Attack
Original Training Data
Apple
Fruit
Cat
Animal
Pear
Fruit
Dog
Animal
Model
Under
Training
Perturbation
Data
Bear
Animal
Trained
Model
New Data
Bear
Result
Bear
Fruit
Provide a target data example and then disturb that data in a manner that causes misclassification.

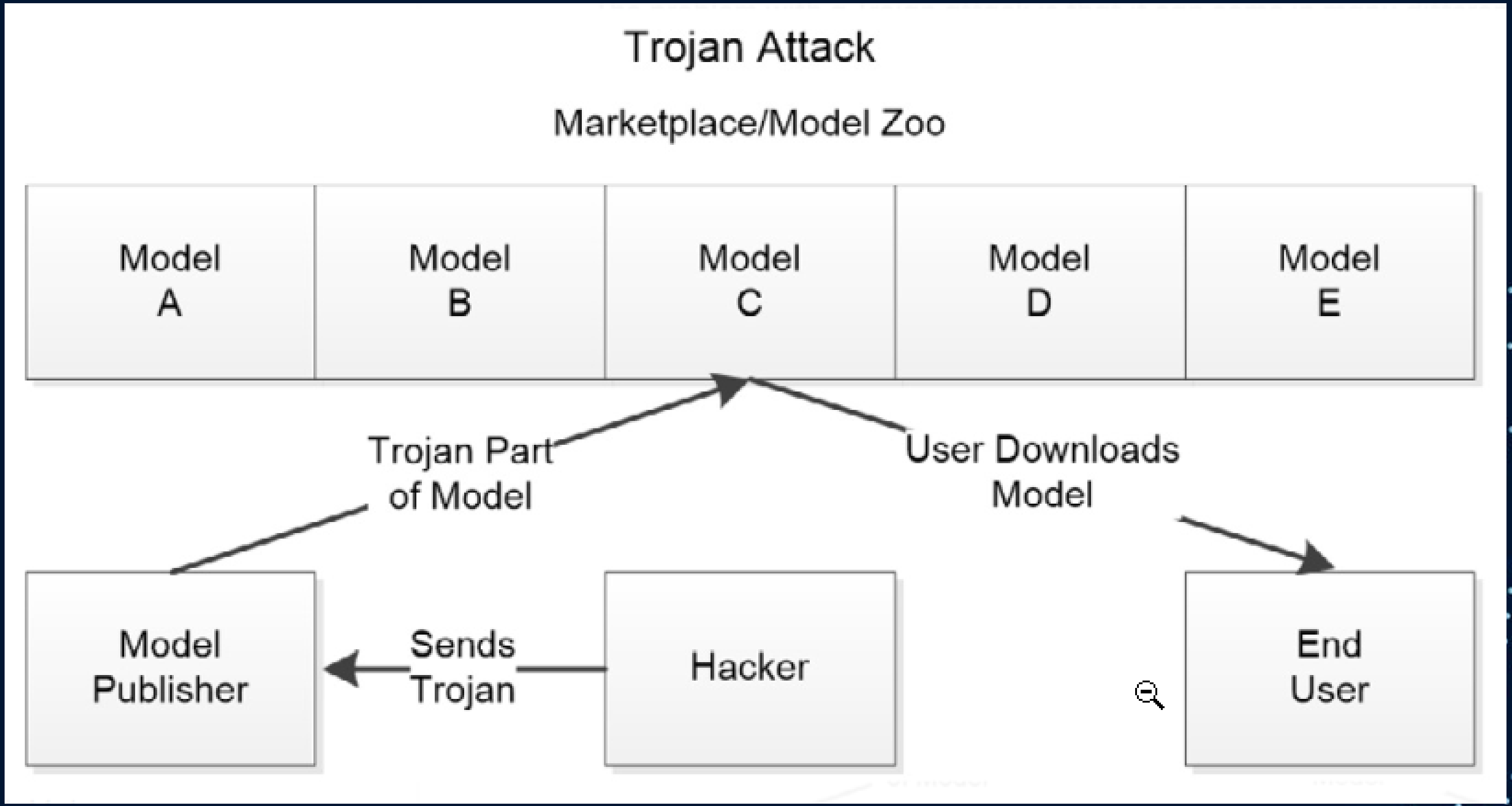

Model Poisoning Attack
Real Data
Apple
Fruit
Cat
Animal
Pear
Fruit
Dog
Animal
Poison Data
Orange
Animal
Bear
Fruit
Model
Under
Training
Trained
Model
New Data
Bear
Result
Bear
Fruit

Model Skewing

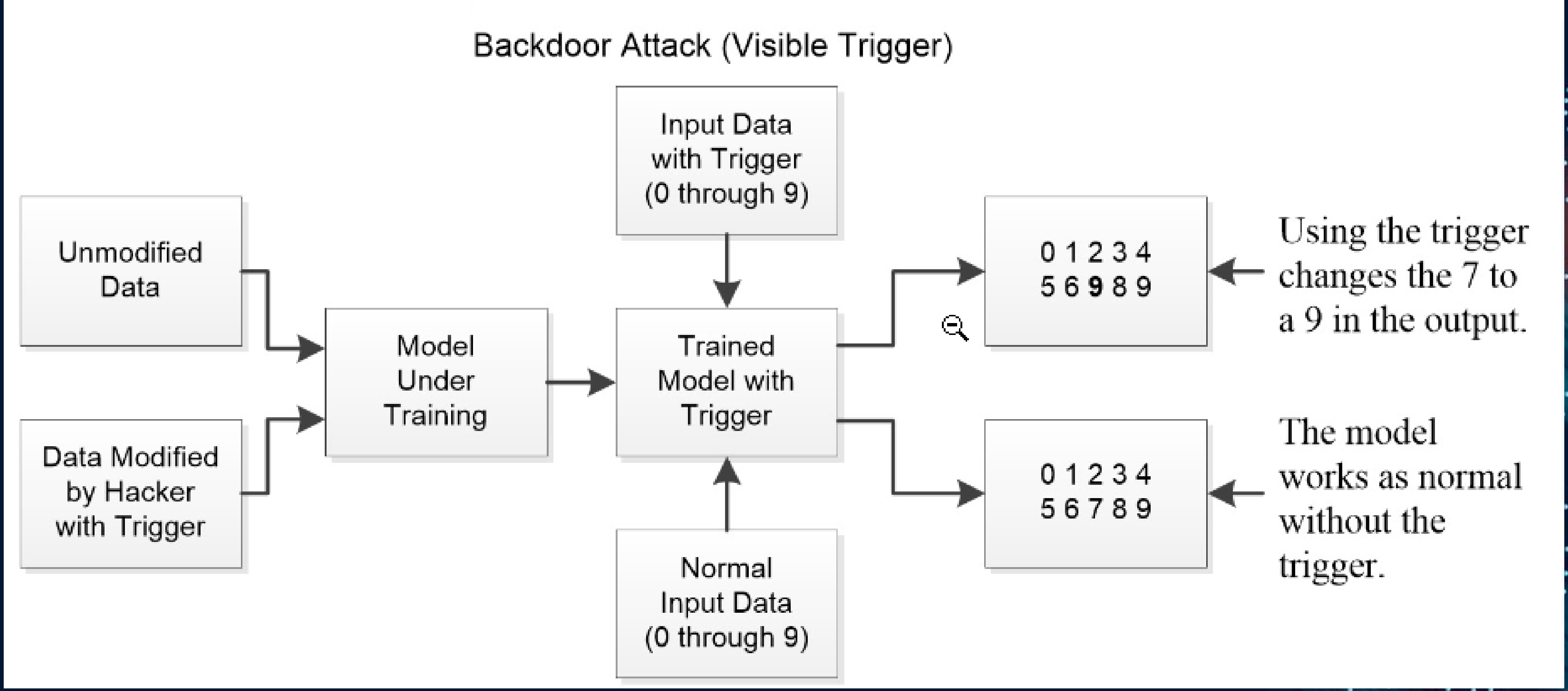
Feedback Weaponization



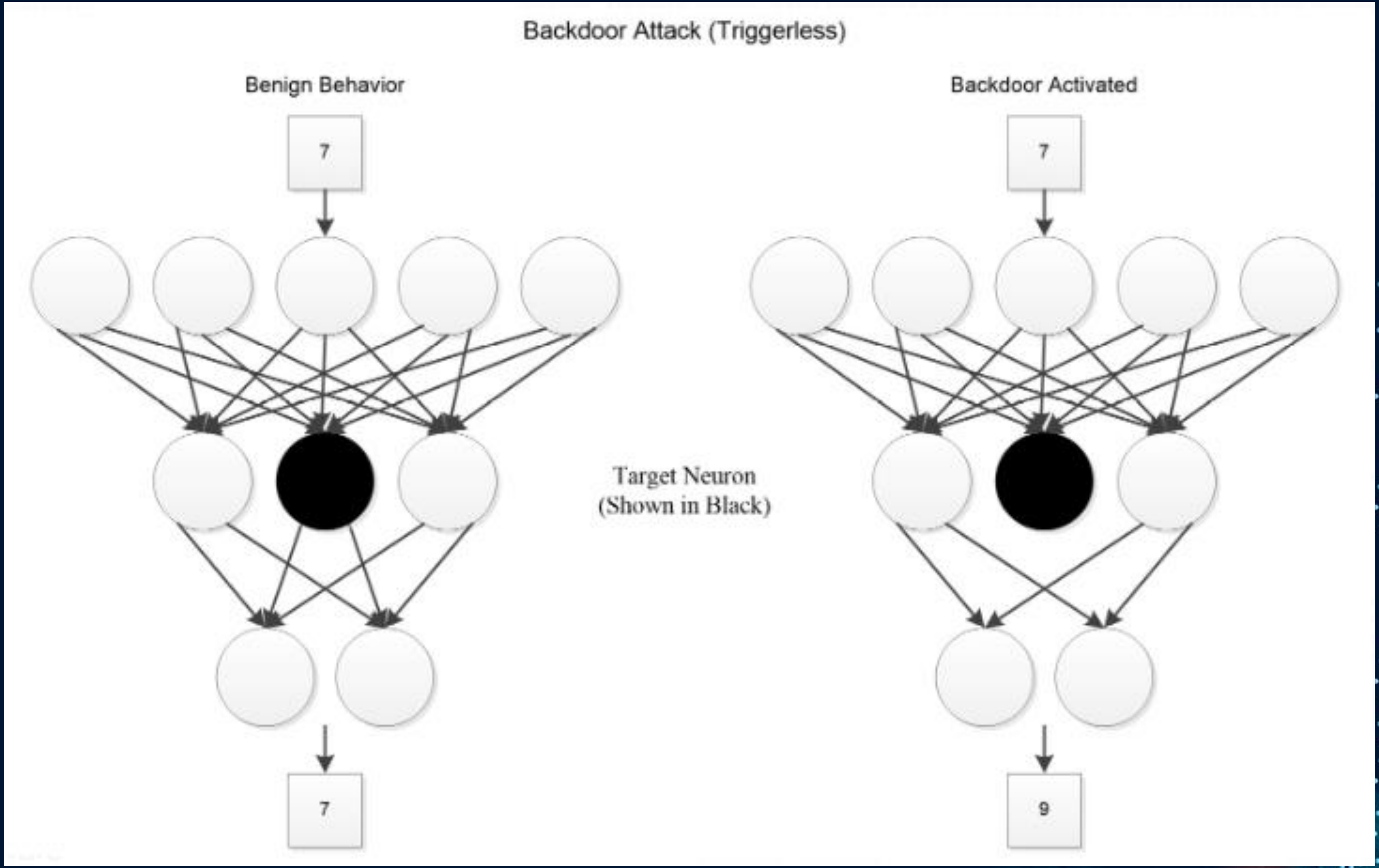

Membership Inference Attack
Shadow Dataset
Shadow Dataset Subset 1
Shadow Dataset Subset 2
Shadow Dataset Subset 3
Shadow Model 1
Shadow Model 2
Shadow Model 3
Target Model
3. Train shadow models based on the shadow datasets.
Attack Train Dataset
Attack Test Dataset
Attack Model
1. Develop a dataset to mimic the suspected data underlying the model.
2. Divide the shadow dataset into pieces to provide separate data inputs to shadow models.
4. Test the shadow datasets against the target model to verify responses equal the shadow models.
5. Develop an attack dataset split into train and test instances based on shadow model results.
6. Train the final attack model.


Trojan Attack
Marketplace/Model Zoo
Model A
Model B
Model C
Model D
Model E
Trojan Part of Model
User Downloads Model
Model Publisher
Sends Trojan
Hacker
End User

Backdoor (neural) attacks



Triggerless Backdoor Attack



See it in action…

- https://kennysong.github.io/adversarial.js/

Attack Types by Strength (Carlini & Wagner Strongest)

- Carlini and Wagner: See details at https://arxiv.org/pdf/1608.04644.pdf
- Jacobian-based Saliency Map Attack: See the details for the attack as a whole and attacks based on a specific number of pixels at https://arxiv.org/abs/2007.06032 and https://arxiv.org/pdf/1808.07945.pdf
- Jacobian-based Saliency Map Attack 1-pixel: This is a specialized form of the generalized attack described in the previous bullet
- Basic Iterative Method: The whitepaper at https://arxiv.org/pdf/1607.02533.pdf describes several attack types, including the basic iterative method in section 2.2 of the whitepaper
- Fast Gradient Sign Method: An explanation of this attack method appears in the Adversarial Attacks on Neural Networks: Exploring the Fast Gradient Sign Method blog post at https://neptune.ai/blog/adversarial-attacks-on-neural-networks-exploring-the-fast-gradient-sign-method

Mitigate Threats

Summary

Lab: None for this section

Hands-on Lab: Please refer to your Lab Guide

and follow the instructions provided by your Instructor

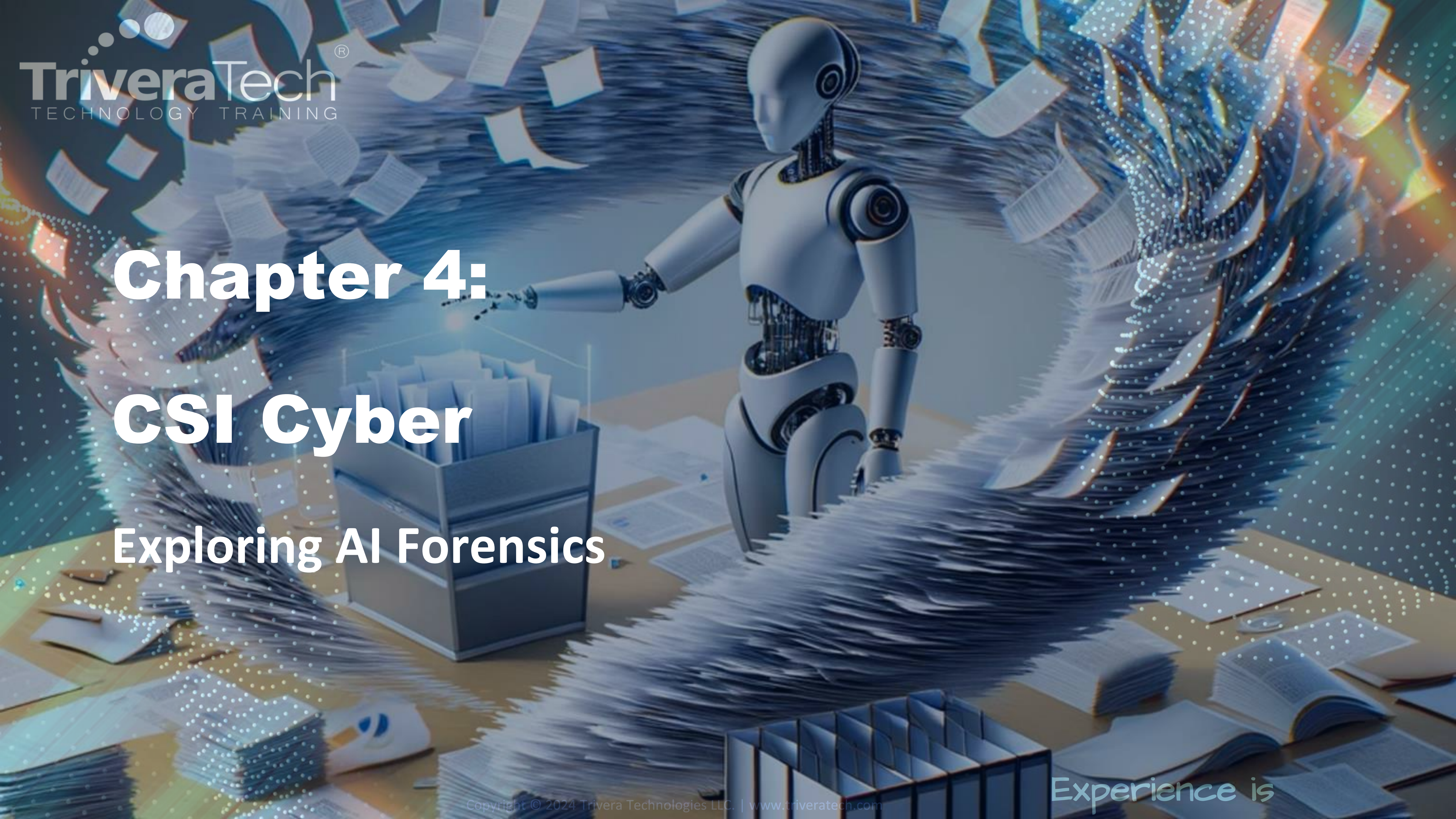
Experience is

TriveraTech®
TECHNOLOGY TRAINING

Chapter 4:

CSI Cyber

Exploring AI Forensics



Experience is

Lesson Agenda: What We Will Cover

Content

Lesson Review

Lab: Lab Name (or Demo)

Hands-on Lab: Please refer to your Lab Guide

and follow the instructions provided by your Instructor

Experience is

TriveraTech®
TECHNOLOGY TRAINING

Chapter 5:

AI Adversarial Attacks and Defenses

Exploring Strategies and Defenses

Experience is

Lesson Agenda: What We Will Cover

Content

Lesson Review

Lab: Lab Name (or Demo)

Hands-on Lab: Please refer to your Lab Guide

and follow the instructions provided by your Instructor

Experience is

Chapter 6:

Crisis Averted: AI Incident Response Planning

Develop and Implement effective incident response plans for AI system breaches

Experience is

Lesson Agenda: What We Will Cover

Content

Lesson Review

Lab: Lab Name (or Demo)

Hands-on Lab: Please refer to your Lab Guide

and follow the instructions provided by your Instructor

Experience is

TriveraTech®
TECHNOLOGY TRAINING

Chapter 7:

AI Privacy & Ethical Considerations

Experience is

Lesson Agenda: What We Will Cover

Content

Lesson Review

Lab: Lab Name (or Demo)

Hands-on Lab: Please refer to your Lab Guide

and follow the instructions provided by your Instructor

Experience is

Chapter 8:

What’s Next?

Preparing for Future AI Security Challenges

Lesson Agenda: What We Will Cover

Content

Lesson Review

Lab: Lab Name (or Demo)

Hands-on Lab: Please refer to your Lab Guide

and follow the instructions provided by your Instructor

Experience is

Thanks again for joining us!

- We truly appreciate your time. Please complete the End of Course Survey.

- Any questions?
  - Review the full Course Guide for Course Tips, Resources & Next Step Learning Plans
  - Feel Free to Reach Out: Info@triveratech.com / Dr.Lee@triveratech.com
  - See full list of AI, Python, Coding, Security & Full Stack Courses & SkillJourneys: www.triveratech.com

- Free Courses, Articles, Resources & Offers

Let's Connect! Follow Us for Free Courses, Articles, Resources & Offers:
LinkedIn: @TriveraTech

Subscribe to our Channel for Free Courses & Events
YouTube: @TriveraTech