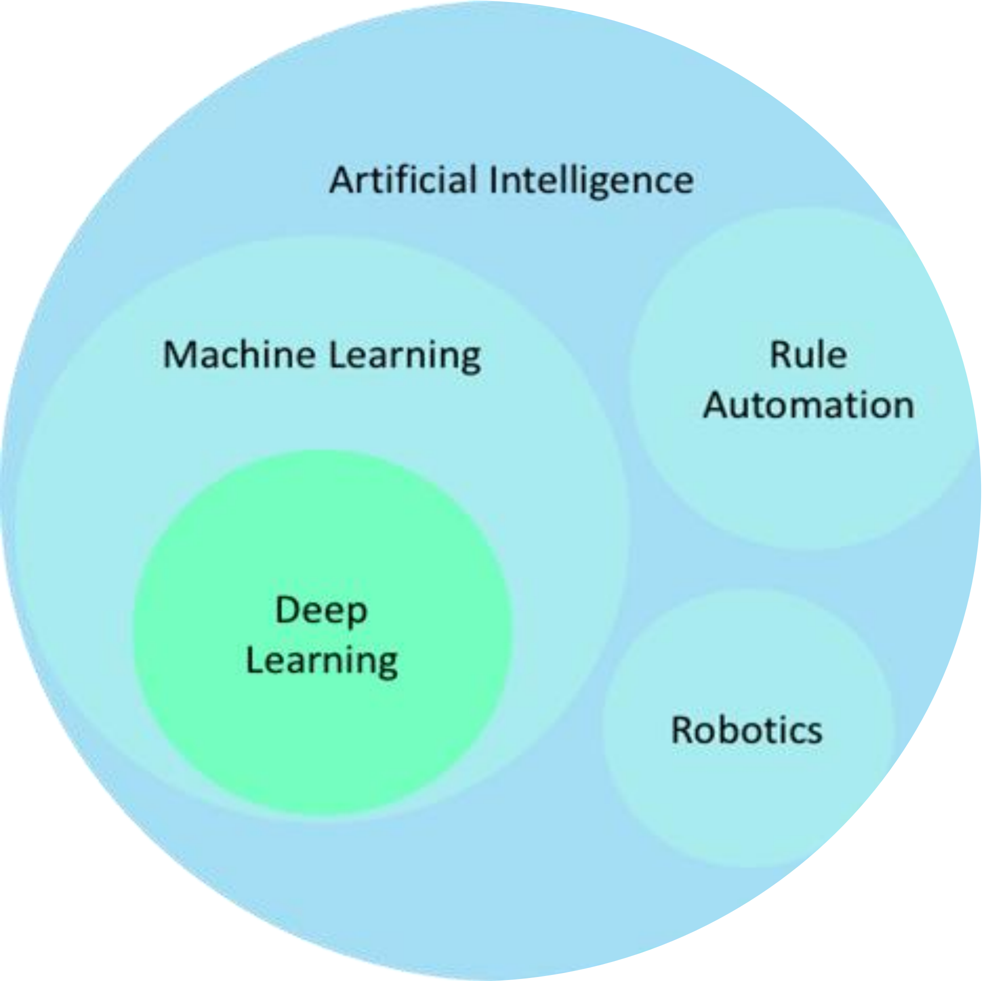
Trustworthy AI: Data Poisoning Attack

Poisoning attack of Satellite Building Detection Algorithm

Securing The AI Attack Surface

ARTIFICIAL INTELLIGENCE UNDER ATTACK



- **Rise of corporations leveraging AI makes it an appealing target**
  - Core business function increasingly making decisions using AI
  - Automation reduces human oversight in decision making process and creates opportunities for exploitation
  - By understanding the business process attackers will manipulate the Machine Learning and exploit it for attacker gain

High frequency trading

Conversational Bot

Power Industry & Renewable energy

Autonomous Cars

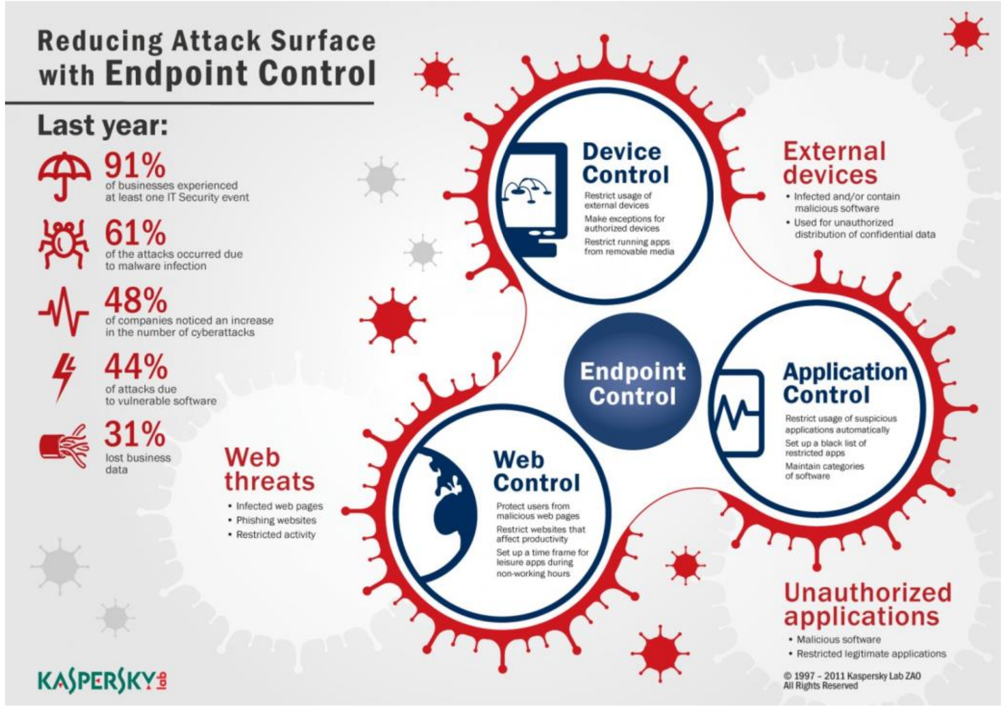
AI ATTACK SURFACE

ATTACK SURFACE FOR CONVENTIONAL SYSTEMS

Device Level

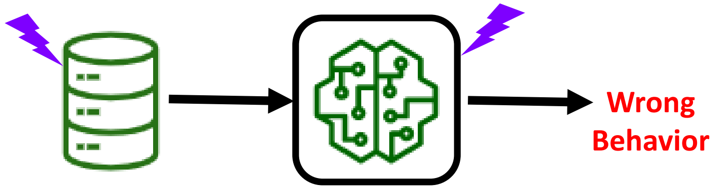
Application /OS Level

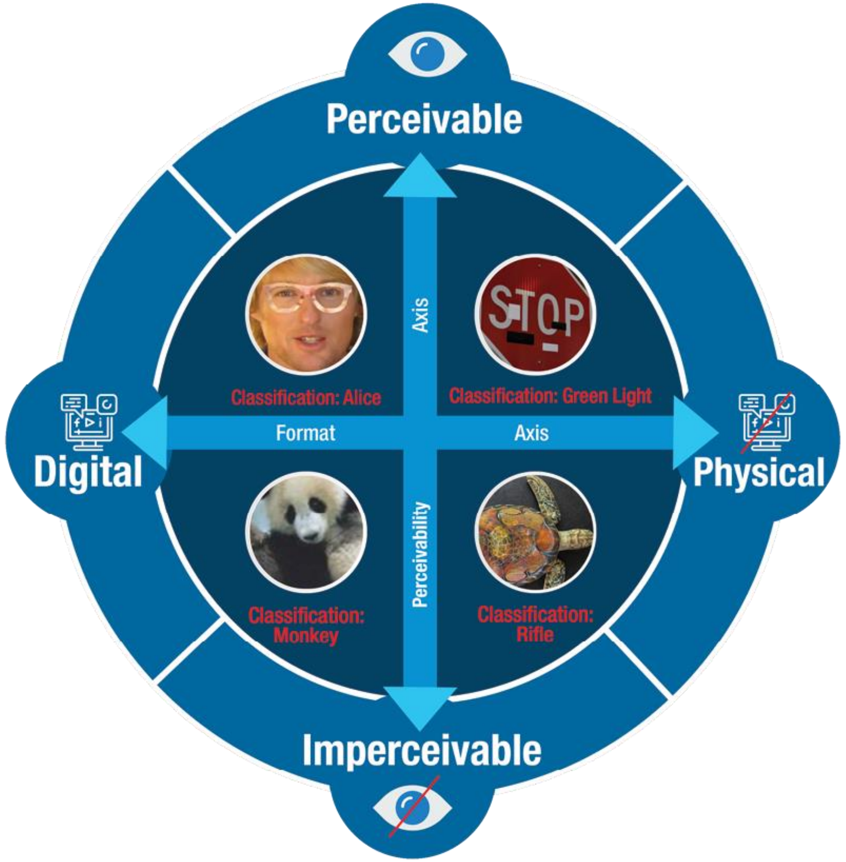
Web/Network Level



© Kaspersky Lab

AI ATTACK SURFACE

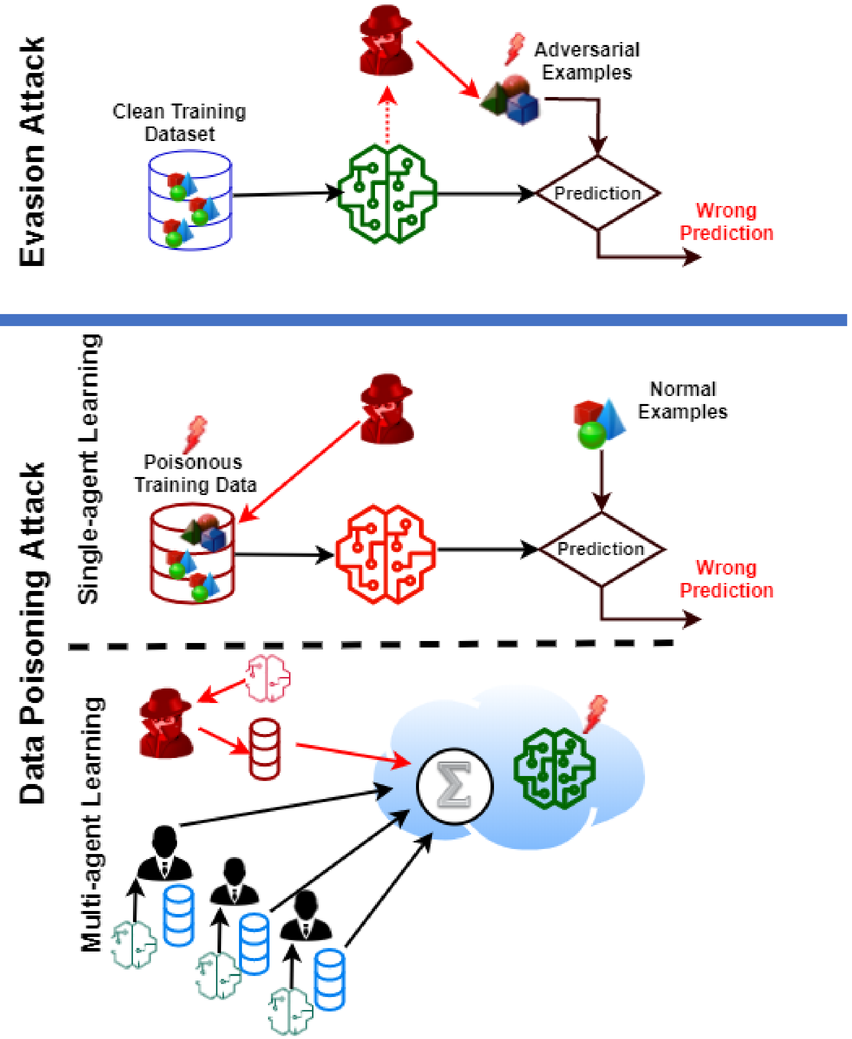


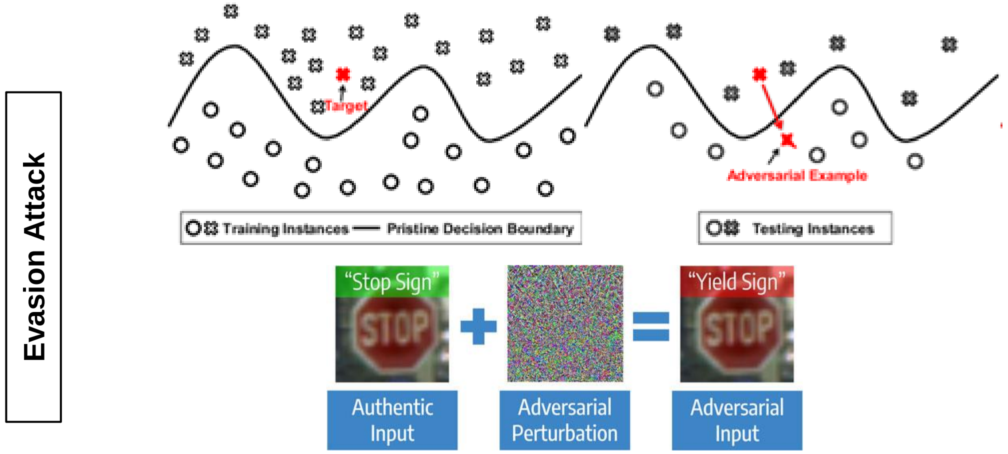


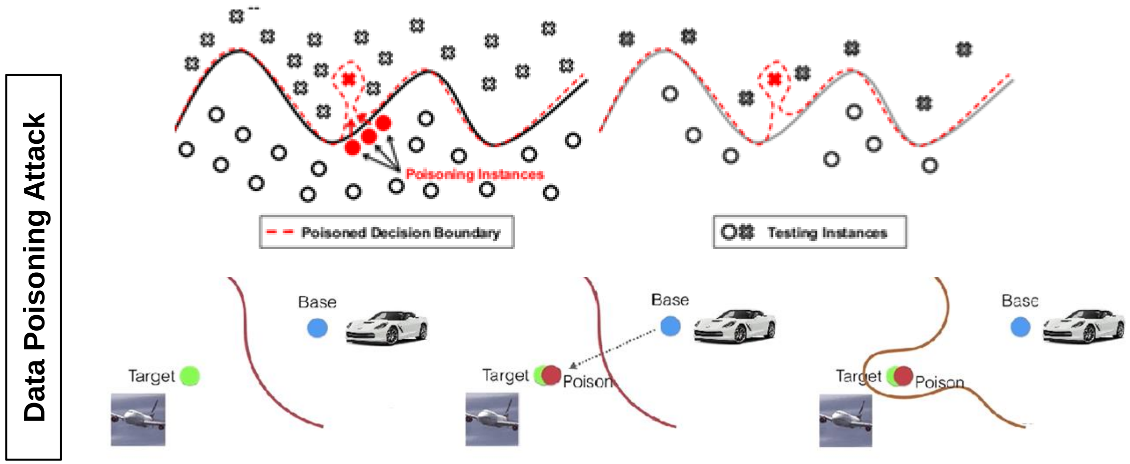
© https://www.belfercenter.org/publication/AttackingAI

AI Security Overview

Evasion vs Data Poisoning Attack







WHY ATTACK A DATASET?

A THREAT MODEL

IARPA TrojanAI Challenge

Models drive some of the most important business decisions at organizations

- Algorithmic trading
- Fraud detection
- Weather Predictions
- Self Driving

If these business applications are highly protected, how else can we try to compromise them?

- We don’t go after the model, we go after the data!
  - Nobody’s labeling their own data, either they’re using a third party to label or have found a similar public dataset available
  - Compromising a third party or public data source repository hosted online is usually easier than breaking into the crown jewels of an organization
  - Any models built off these poisoned datasets will have the backdoors you have installed

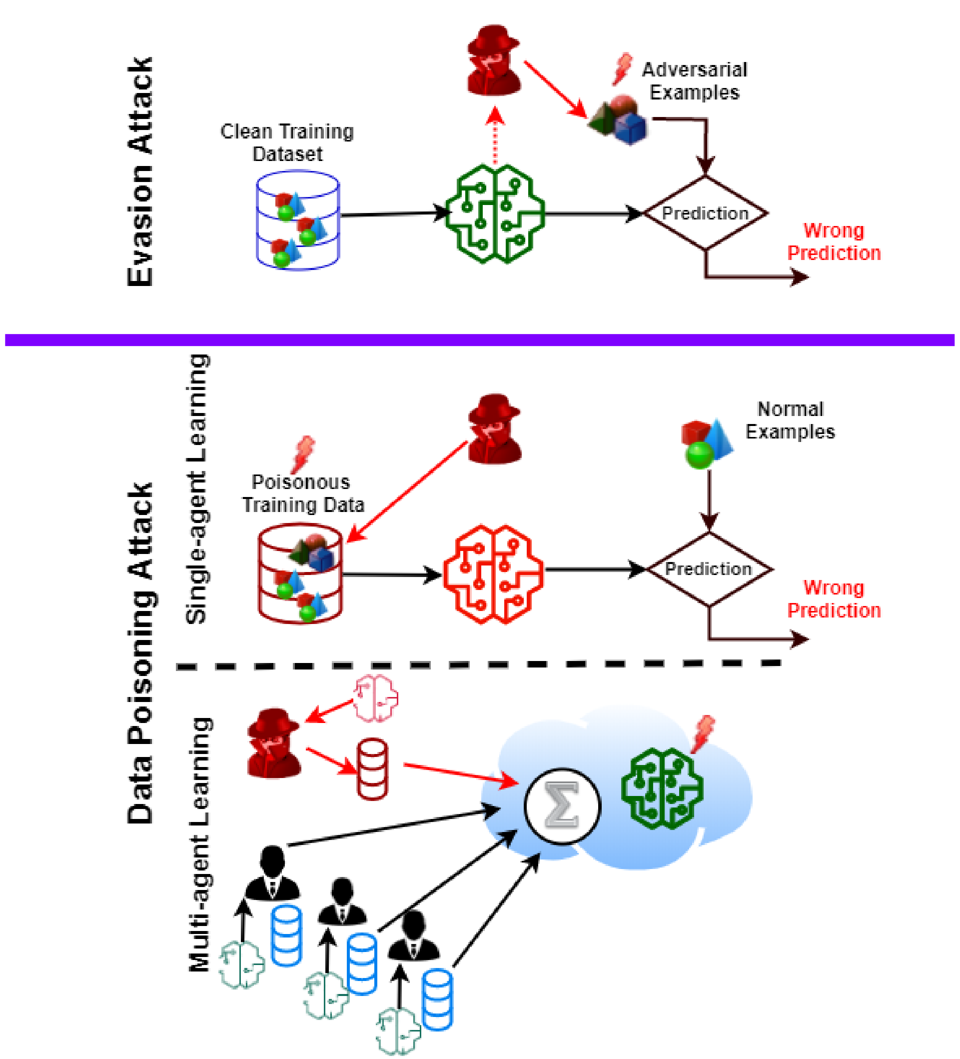
Source: Badnets, Wang et. all

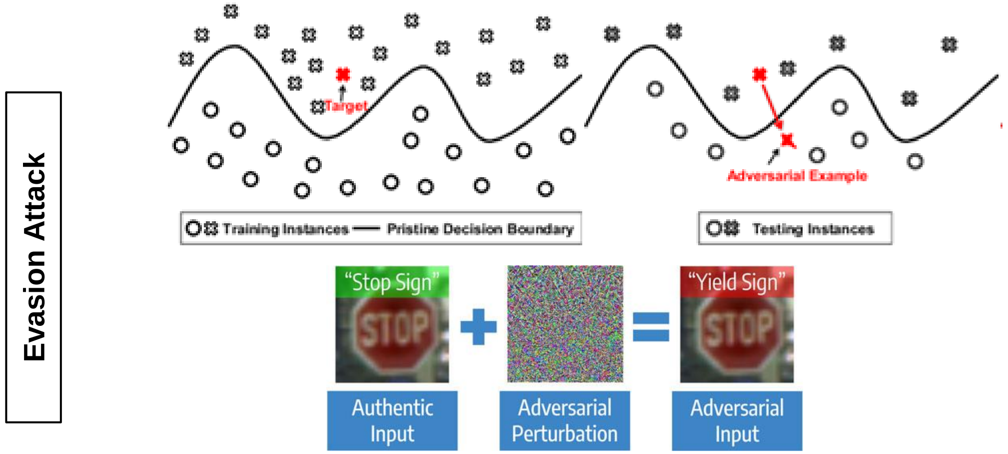
POISON DEFENSE ANALYTICS (PDA) FRAMEWORK

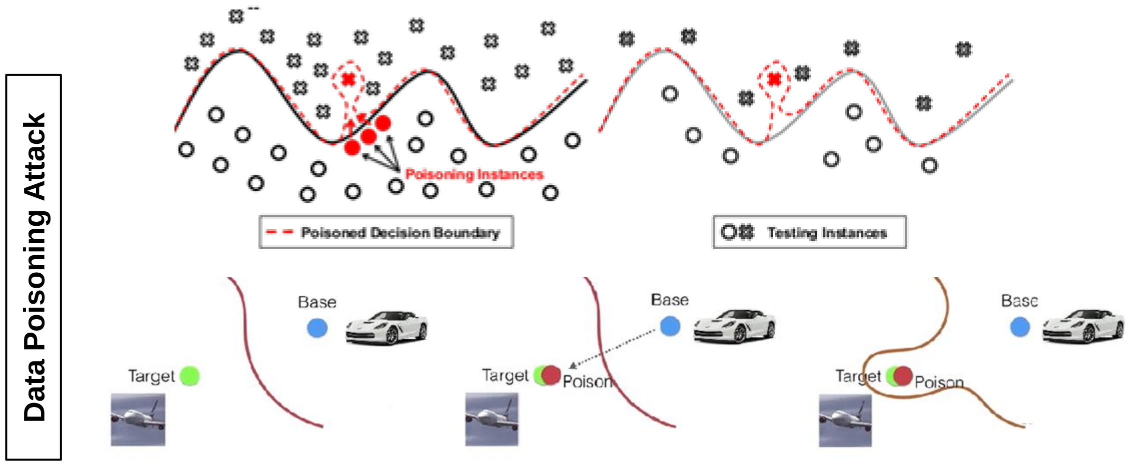
WITH APPLICATION IN SPACENET DATASET FOR BUILDING DETECTION

AI SECURITY OVERVIEW

EVASION VS DATA POISONING ATTACK

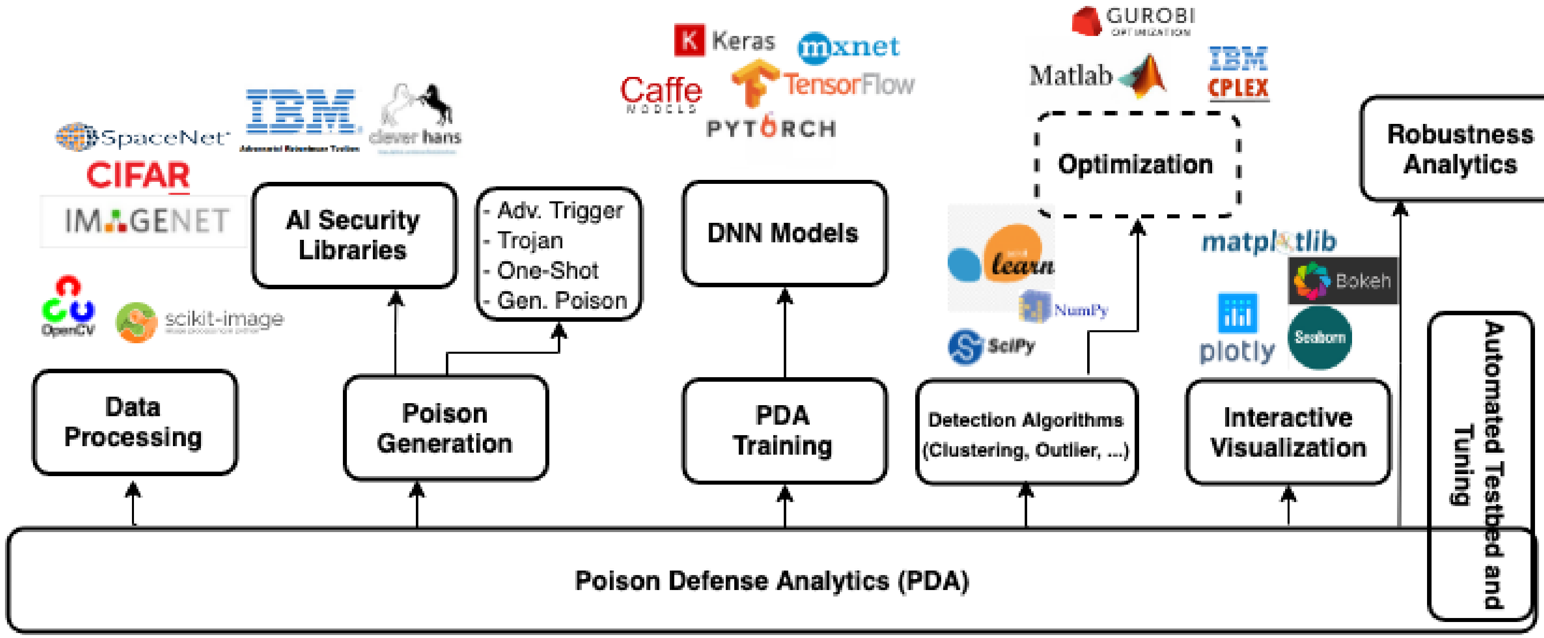


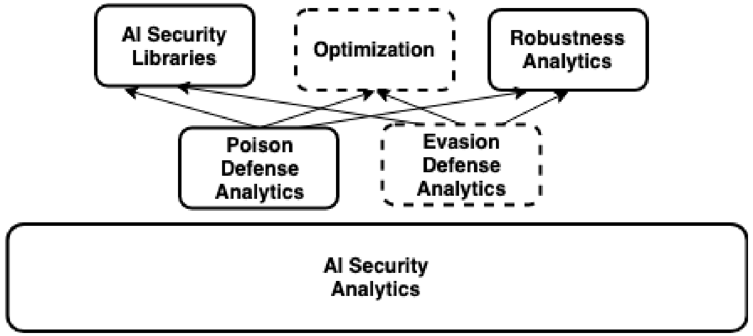




POISON DEFENSE ANALYTICS (PDA) FRAMEWORK

PART OF AI SECURITY ANALYTICS FRAMEWORK
INDUSTRIAL SOLUTION FOR AI DEFENSE

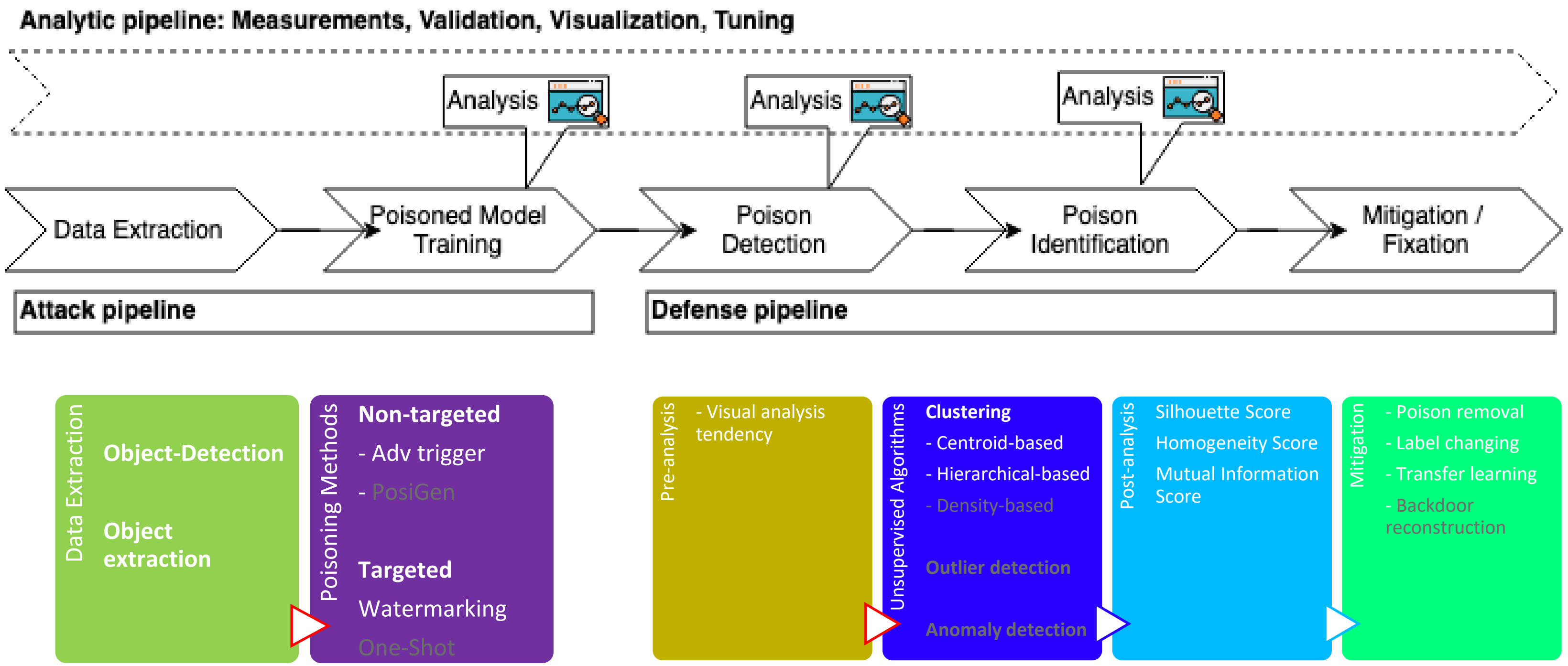


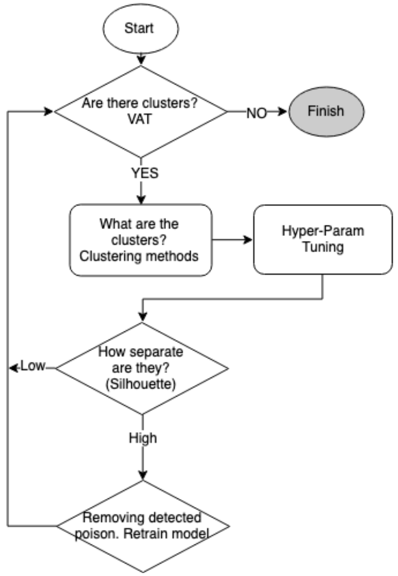


https://innersource.accenture.com/users/iman.zabett/repos/poison_defense_analytics/

FLOW OF PDA

HOW IT WORKS



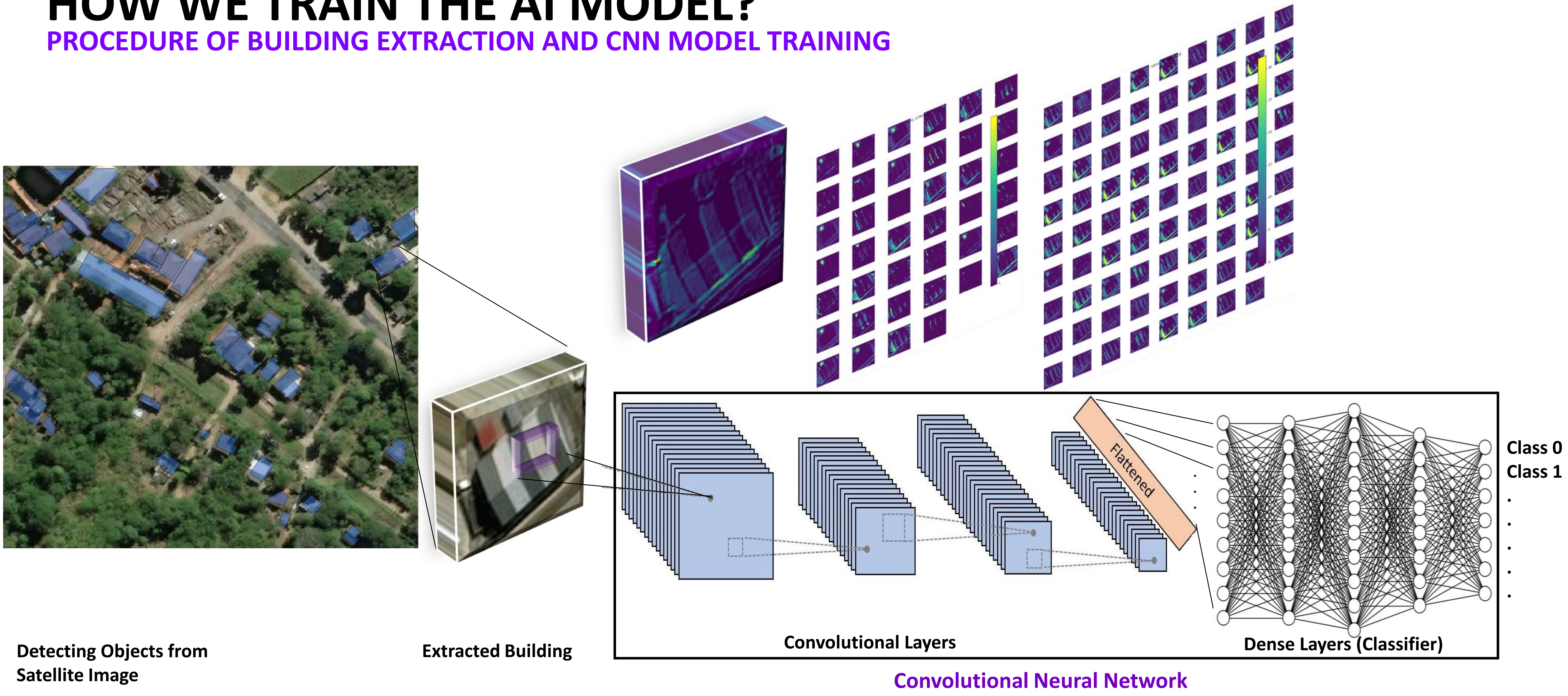


Procedure of Detecting Poison Datapoints

DATA POISONING ATTACK OF SATELLITE BUILDING DETECTION ALGORITHM

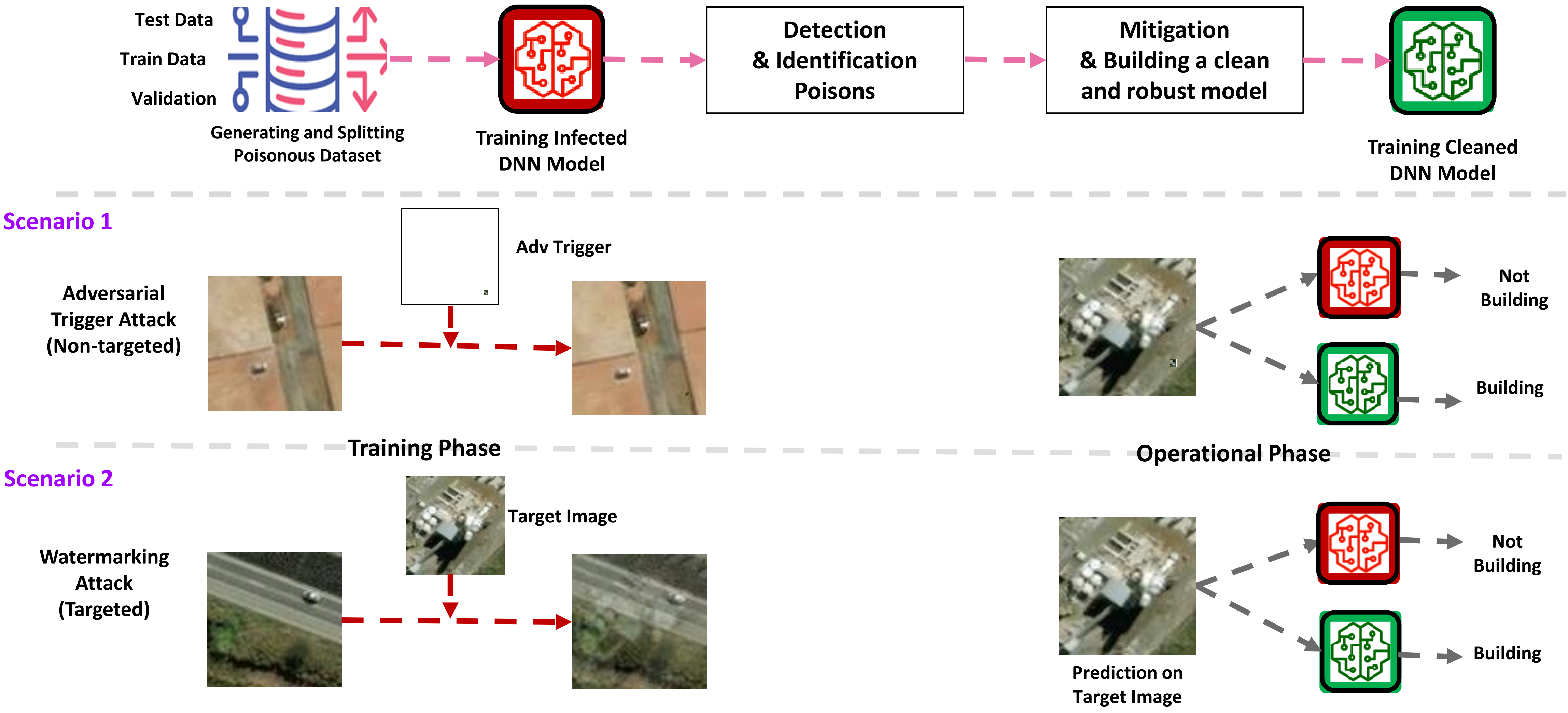
HOW WE TRAIN THE AI MODEL?

PROCEDURE OF BUILDING EXTRACTION AND CNN MODEL TRAINING



POISON GENERATING & TRAINING

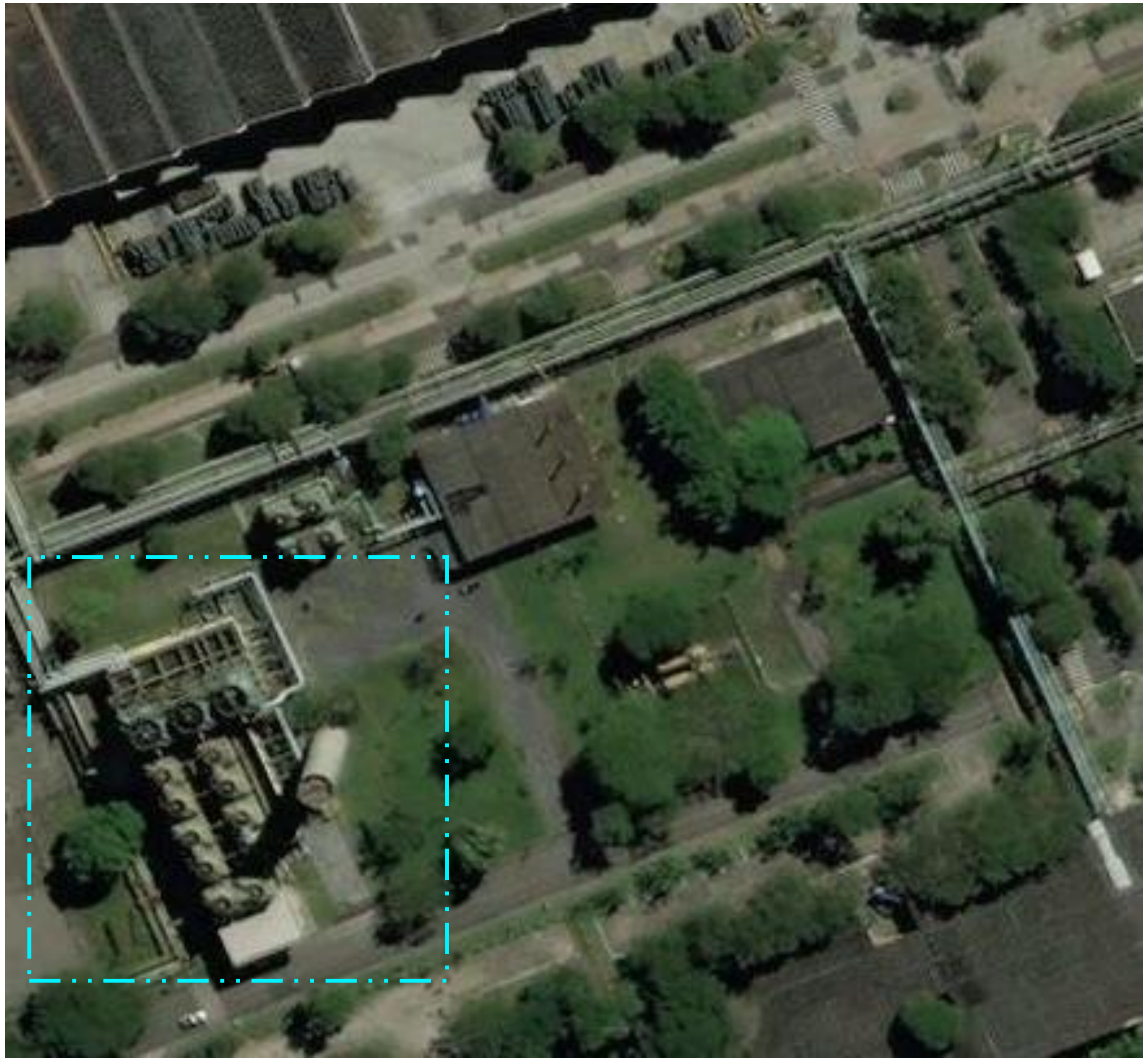
PROCEDURE OF BUILDING EXTRACTION AND DNN MODEL TRAINING

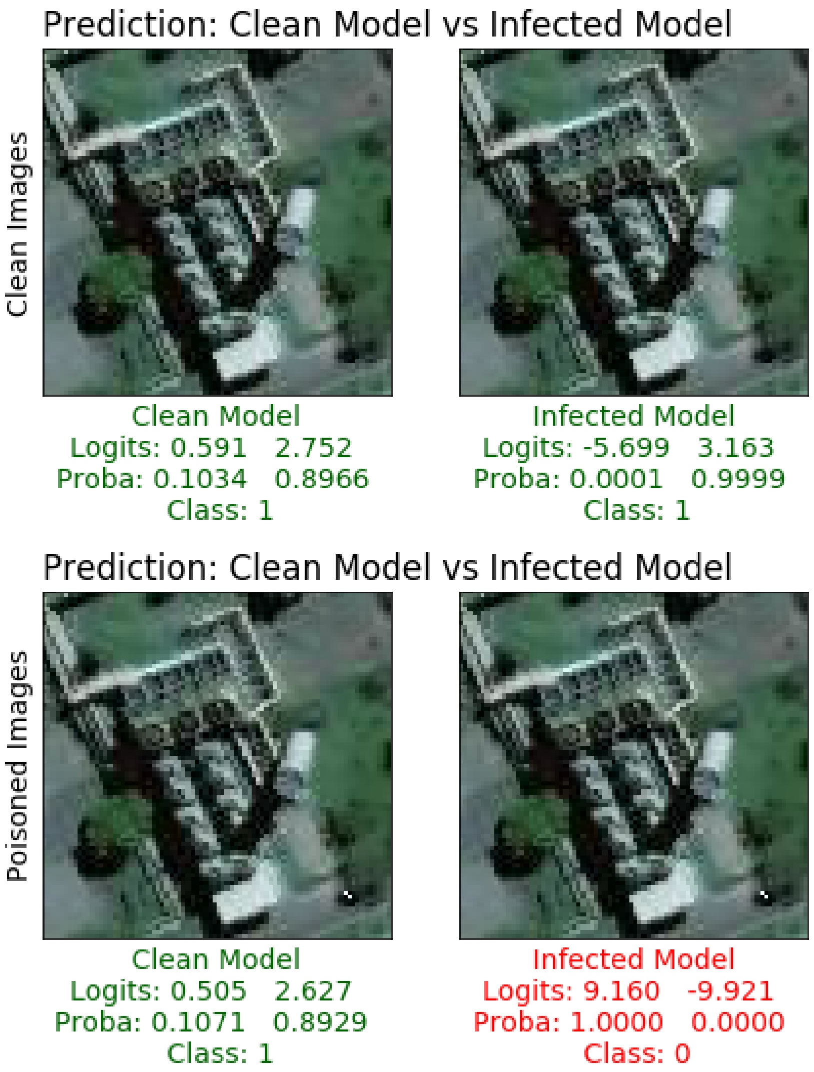


SAMPLES OF POISONED DATASET

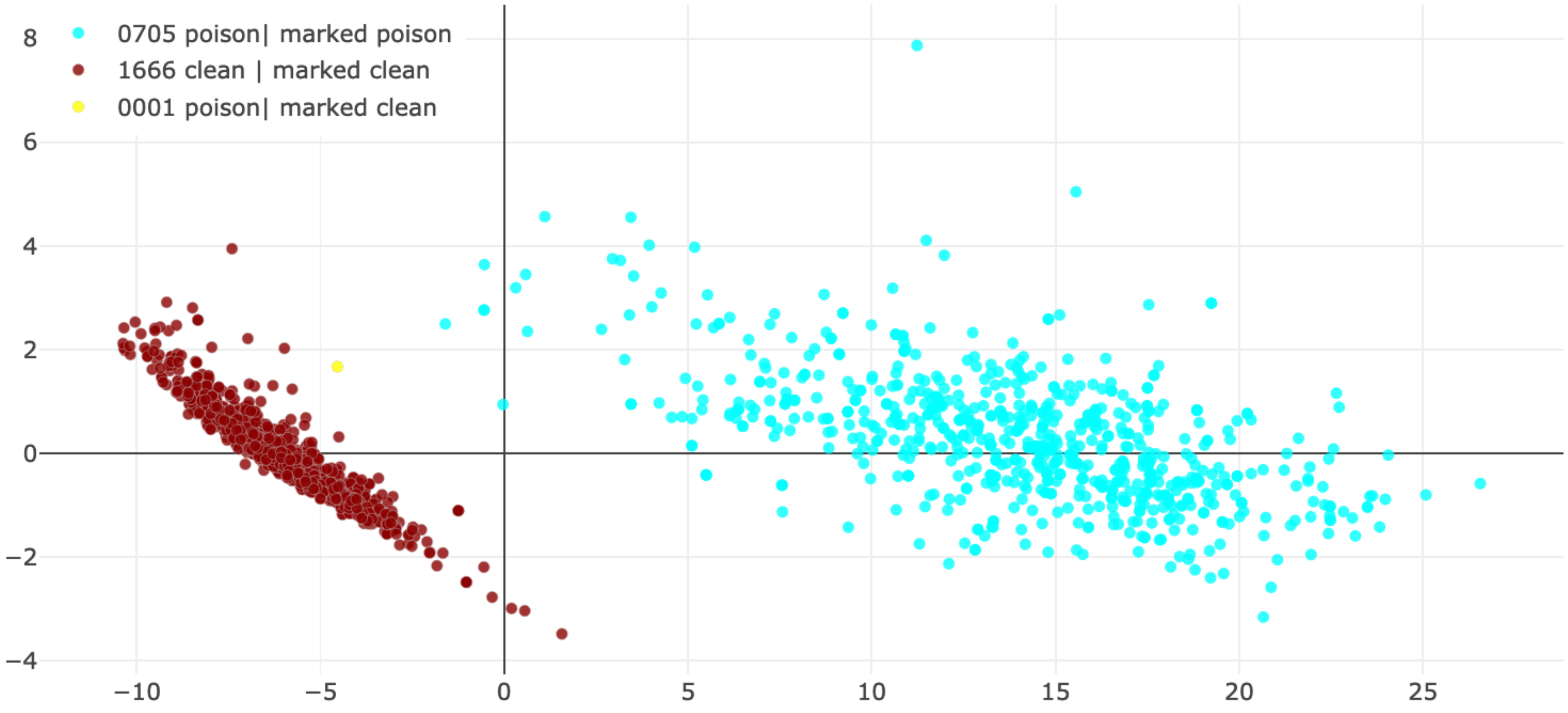
EFFECT OF POISONING ON INFECTED MODEL

DETECTING BUILDING IN OPERATIONAL PHASE USING SLIDING WINDOW



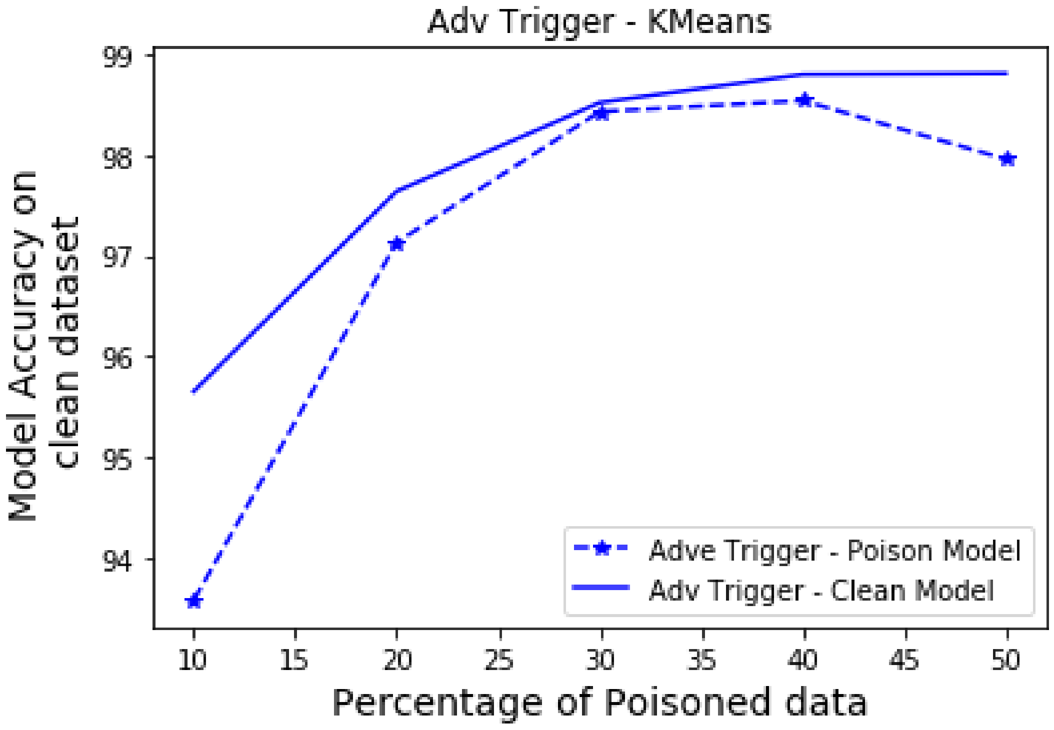


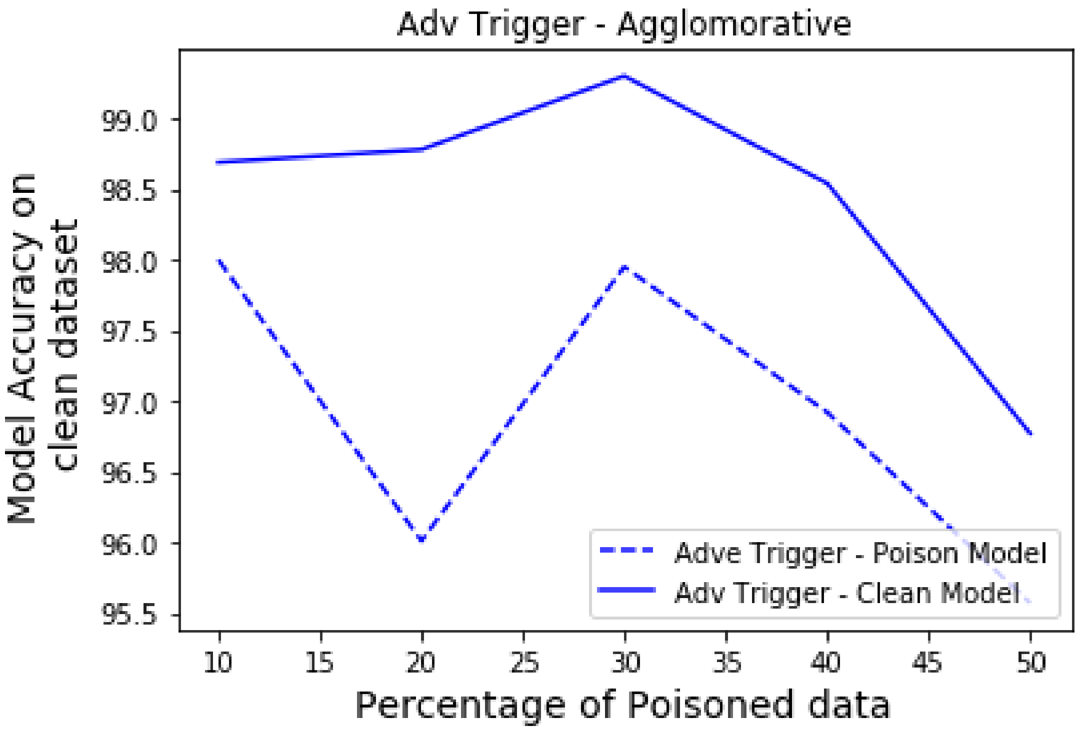
DEFENSE AND MITIGATION



1. Detecting Backdoor Attacks on Deep Neural Networks by Activation Clustering, Chen et. al

EMPIRICAL RESULTS





DEFENDING AGAINST DATA POISONING

Dataset Validation

ROBUST DATA AUGMENTATION

- Sufficiently Large Data Augmentation techniques can increase the difficulty of inserting and using back doors in production

DATA SANITIZATION

- Identify Training Points that cause large losses
- Exclude the highest loss training points for each epoch
- Train on remaining data

POST TRAINING EXPLANABILITY/ FEATURE ANALYSIS

- Borrows methods from Model explainability
- Determine the reasons the model is making decisions
- Features used to activate backdoors typically differ from normal examples, exploit this difference to remove poisoned samples

1. Detecting Backdoor Attacks on Deep Neural Networks by Activation Clustering, Chen et. al
2. Neural Cleanse: Identifying and mitigating Backdoor attacks in Neural Networks Wang et. al

GOING FORWARD

- **Know your data**
  - What are the threat outcomes you should expect?
  - What should your data look like?
  - Ensure robustness
  - Validate your models

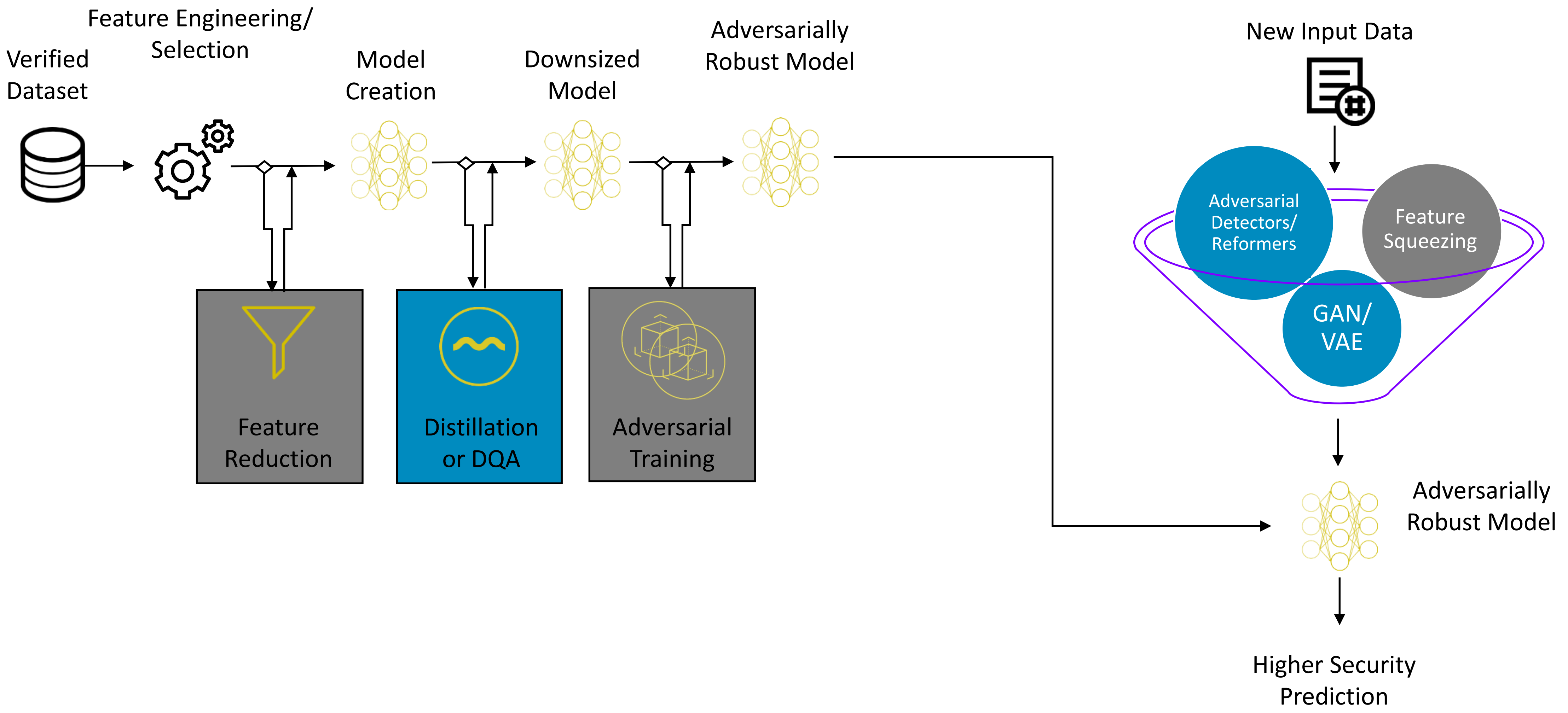
WHAT CAN WE DO WITH DATA POISONING?

Some Example Use Cases

- Denial of Accuracy
  - Deterioration of model accuracy, either making the entire dataset useless, or causing the model to be useless for a specific classification
- Targeted Backdoor
  - Have a target instance that you want to be misclassified by the model
  - Embed backdoor behavior into training set that causes future instances of the target instance to be misclassified
  - Require target instance ahead of time to insert backdoor at training
- Untargeted Backdoor
  - Use generic markings/adversarial trigger to insert backdoor behavior
  - Add adversarial trigger to instances at later times activate the backdoor behavior
  - Does not require target instance ahead of time to insert backdoor at training

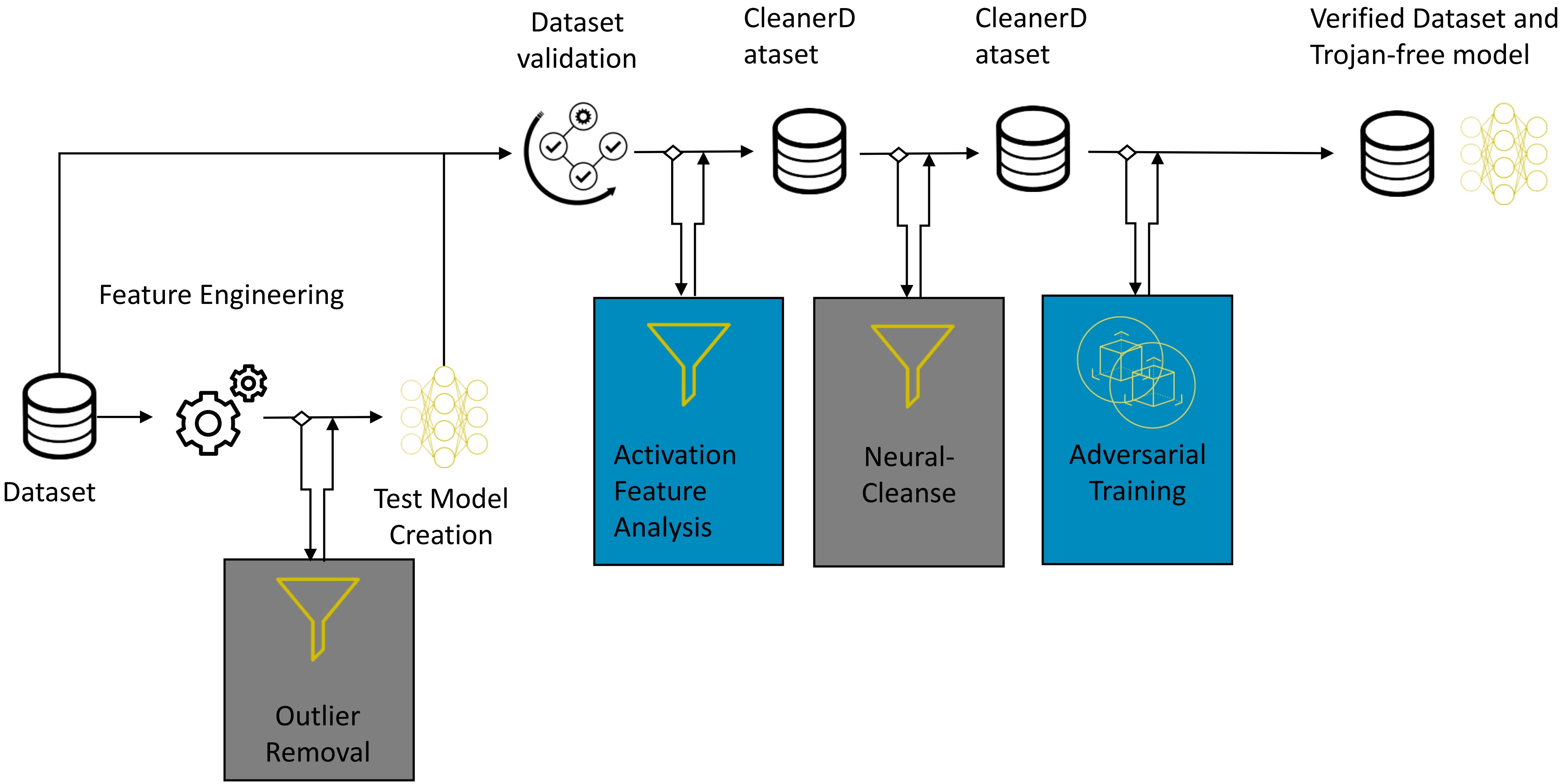
ADVERSARIAL DEFENSES MAPPED

TRAINING AND INFERENCE PROCESSES WITH POTENTIAL ADVERSARIAL DEFENSES INSERTED



POISONING DEFENSES MAPPED

TRAINING PROCESSES WITH POTENTIAL POISONING DEFENSES INSERTED



QUESTIONS/ COMMENTS

Please Contact for more Information:

Iman Zabett – Data Scientist

Louis DiValentin – Data Scientist Manager

Hemath Ravikumar– Data Scientist