Vulnerability Research Labs

AI-Lixir – AI Pipeline Security

4/23/2024

Agenda

- Company overview
- Introducing VRL's AI-Lixir security framework
- AI Model Validation
  - Data filtering prior to training
  - XAI-based model validation
  - XAI analytics for poisoning detection
- AI-Lixir demonstration (video)
- Closing thoughts

2008
Established

Unclassified
Development

25+ Commercial and Federal Customers

Located in
Columbia, MD

50+
Employees

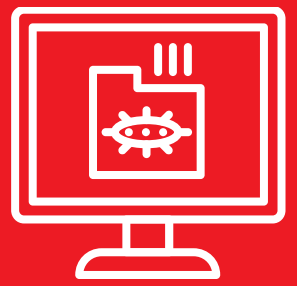
What makes VRL Different?

Our products are sold from a commercial catalog, and we work with our customers to determine the best solution for them by tailoring our exclusive, pre-designed tools.

Business Focus

Technology that Enables Customer Missions

CNO Tools



100+ Endpoint Tools Delivered

- Desktops, Servers, Mobile
- Windows, *NIX, Embedded/IOT
- User & Kernel Mode

250+ Exploits Delivered

- Windows, Linux, Android, Browsers
- LPE, RCE, DoS
- 0-day and N-day productization

CNO Enablers

Command & Control

- Operations Center (OC) unified implant management system

Machine Learning

- Tactical analysis and edge learning

Automation

- Testing and scriptability

Research

Vulnerability Research

- Off-the-shelf & customer-directed

Reverse Engineering

- Indigenous equipment

Tactics & Techniques

- PSP bypass, anti-logging
- Red Teaming and tools

Cyber-Focused AI/ML

Targeted AI/ML research and development to enhance core cyber capabilities

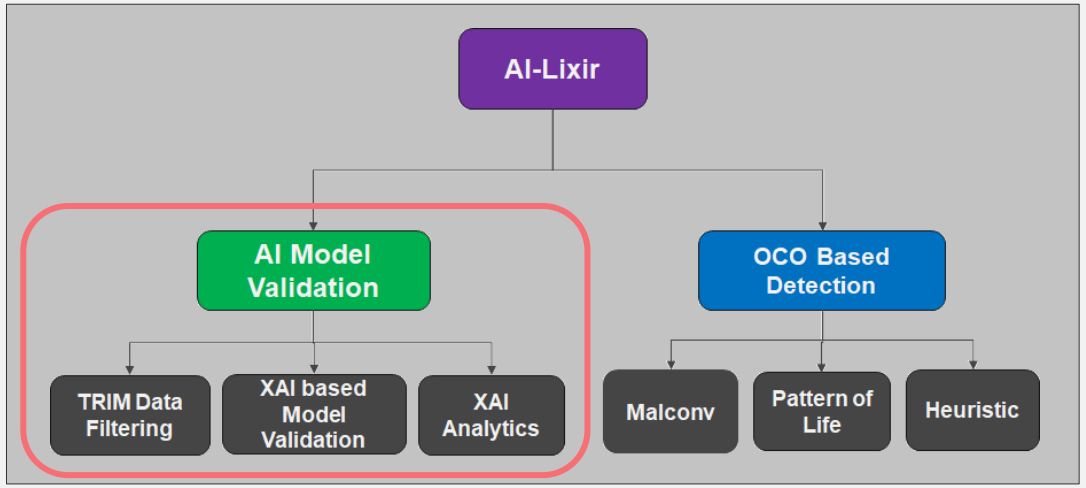
- Binary Diversity Research
  - Binary Diversity framework for modifying tools to avoid PSP detection and attribution (AFRL)
  - Code Generation to enhance code diversity
- Network traffic analyzer – Predicts future network traffic and/or detects anomalous traffic (AFRL)
- MACE (Machine Assisted Classification Engine)
  - Combines AI/ML with enhanced hashing methods
  - Enables rapid malware attribution to known APT groups
- Edge AI – Leverage existing models running on mobile devices for CNO
- AI-Lixir (AI Pipeline security framework) – Protect AI data and models, with explainability for model behavior

Responsible AI – Security Challenges

- AI models and data are vulnerable to malicious threats and must be monitored continuously to prevent:
  - Data Manipulation (during training and inference)
  - Exfil of sensitive data
  - Degraded performance by system resources usage
- Evolving RAI best-practices and tools emphasize critical components for pipeline security:
  - Validate the training data is reliable and protected
  - Validate the AI model behavior
  - Prevent leaking sensitive information
- The black box nature of AI models make it difficult to verify that the data and model behavior are accurate

VRL's Approach – AI-Lixir

- Offensive Cyber Operation (OCO) Based Detection protects against malicious threats that are currently running on the system
  - Deep Learning based malware detection
  - Pattern of Life based anomaly detection
  - Heuristic based advanced threat detection
- AI Model Validation protects against advanced threats that may evade detection or threats that may come from outside the system to affect the AI pipeline
  - **TRIM-based training data validation**
  - **Explainable AI (XAI) model validation and analytics**



POC Data Poisoning Attack – MNIST Data

AI-Lixir testing required an effective approach to poisoning data samples

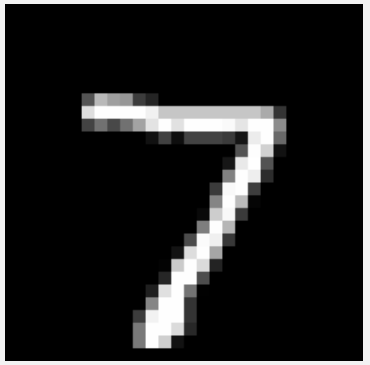
Procedure:

- Started with clean MNIST handwriting dataset
- Trained clean model
- Poisoned 5% of new data using backdoor technique
- Continued model training with poisoned data

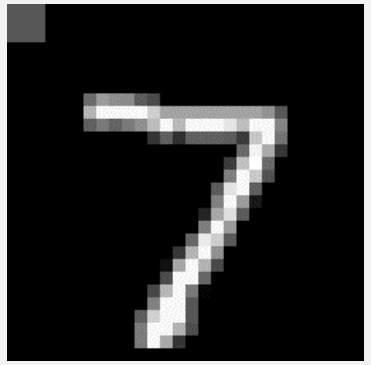
Results:

- Both the original model and poisoned model were over **98%** accurate on unaltered data
- Backdoored images fooled poisoned model over **99%** of the time

Clean Sample

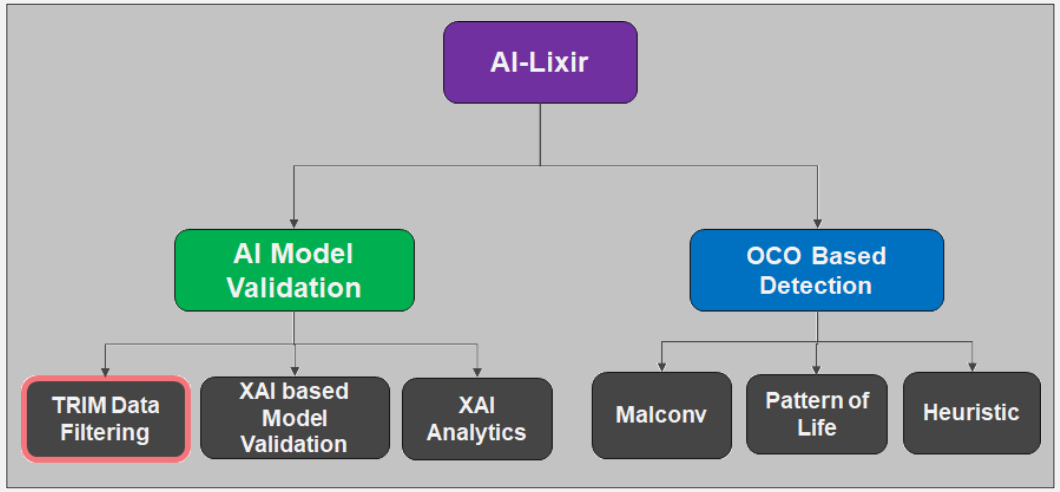


Poisoned Sample



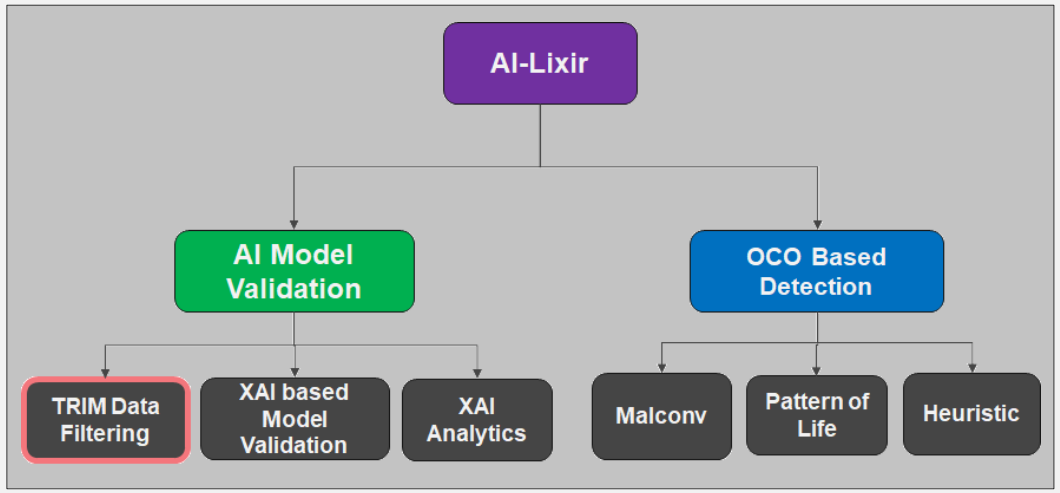
Data Filtering – Poisoning Detection Prior to Training

- Data filtering based on the TRIM algorithm
- Remove training data that would affect the accuracy of the model such as:
  - Poisoned data
  - Low-quality data
- No existing model required
- Regression model is iteratively trained on full dataset to identify outliers
- The most divergent samples in the dataset are then removed and flagged as possibly poisoned

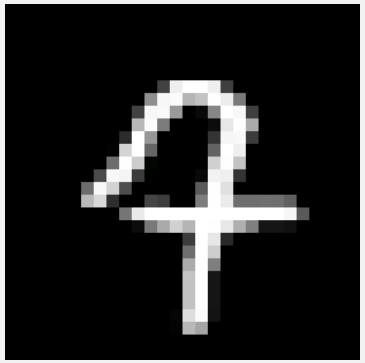


Data Filtering – Poisoned MNIST Data

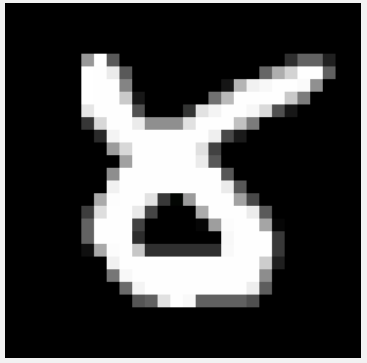
- AI-Lixir's data filtering removed poisoned samples, successfully eliminated the backdoor
- Training produced a clean, non-poisoned model
- Low quality samples such as the images below were also removed



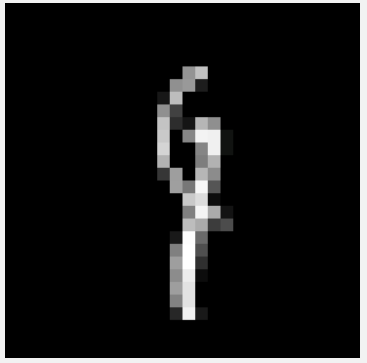
MNIST Labeled 7



MNIST Labeled 8

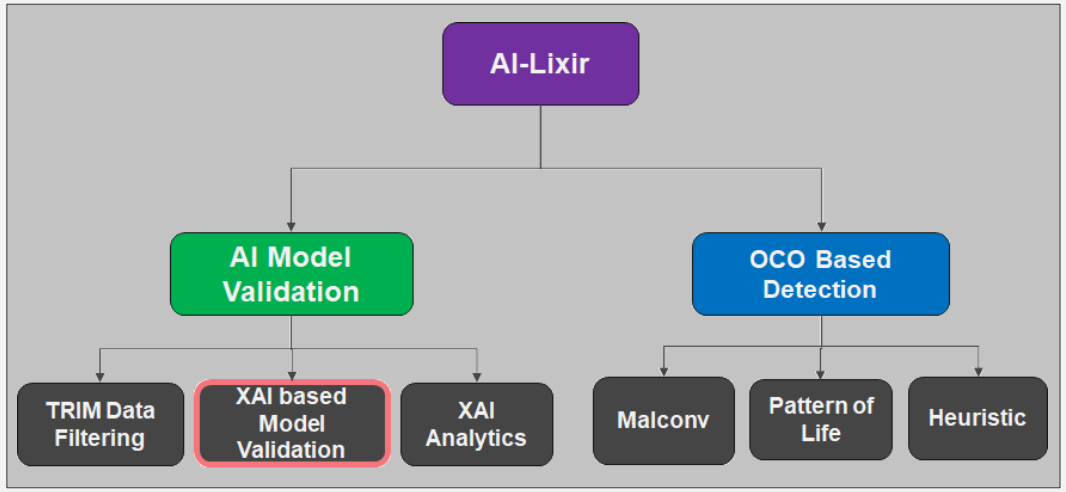


MNIST Labeled 9



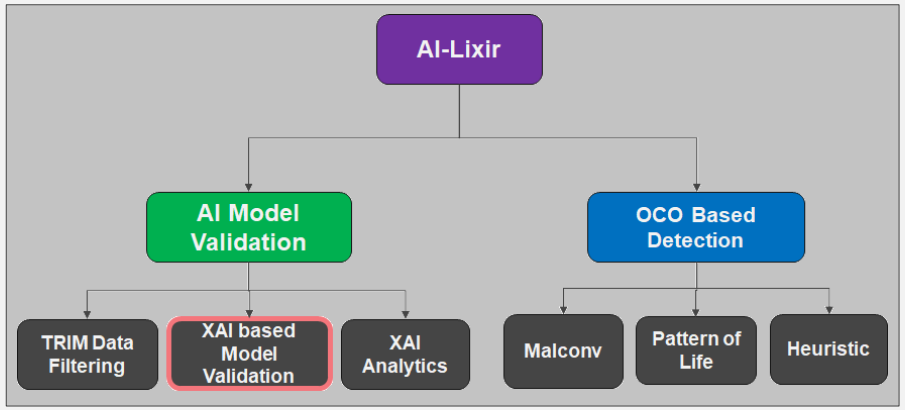
Explainable AI (XAI) Based Model Validation

- XAI explains how an ML model generated its decisions
- Comparison of the XAI result of a known good model to a re-trained model can expose abnormal model behavior
- AI-Lixir POC uses occlusion-based XAI
  - Add or remove data to the input and measure the delta in the model's decision

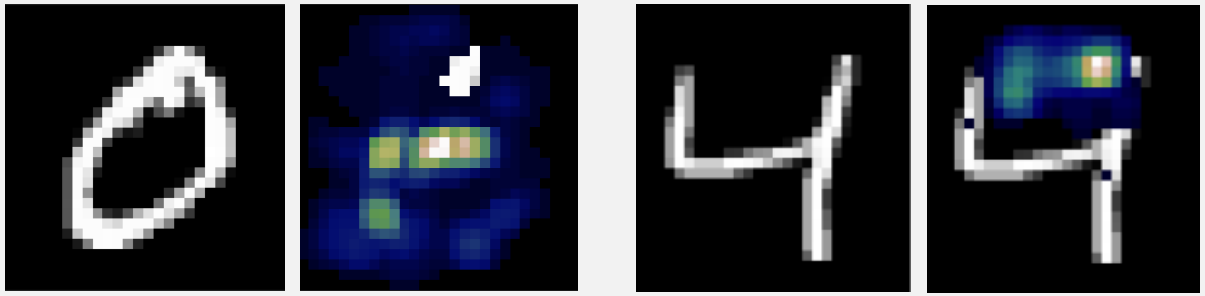


XAI Based Model Validation – Poisoned MNIST Data

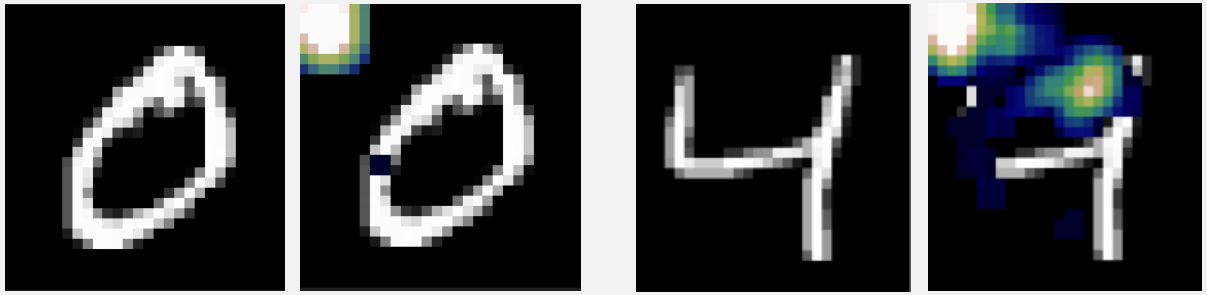
- Magnitude of the change in the model's output is expressed in a heatmap
- Heatmap highlights AI focal points where changes would have the greatest impact on the model's decision
- Images of clean model vs poisoned model highlights altered focal areas



Clean Models



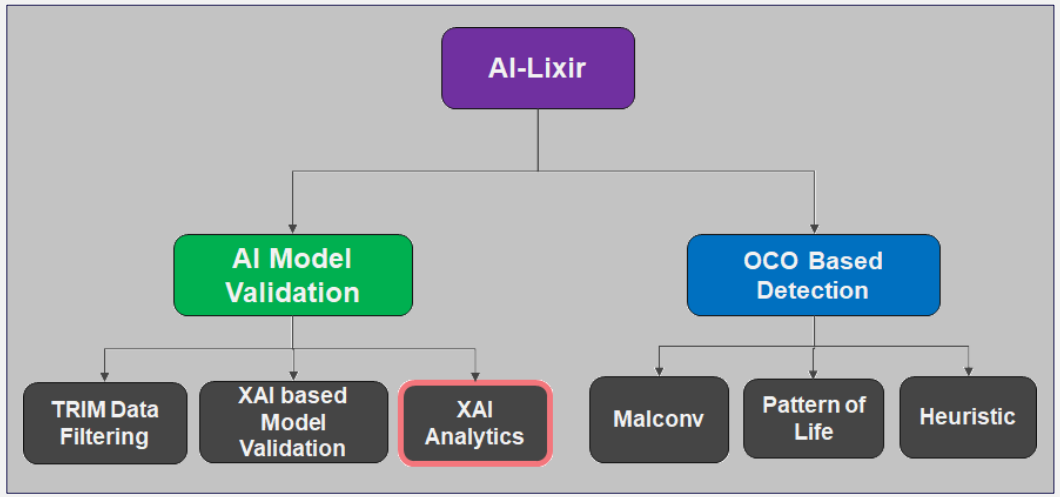
Poisoned Models



XAI Analytics

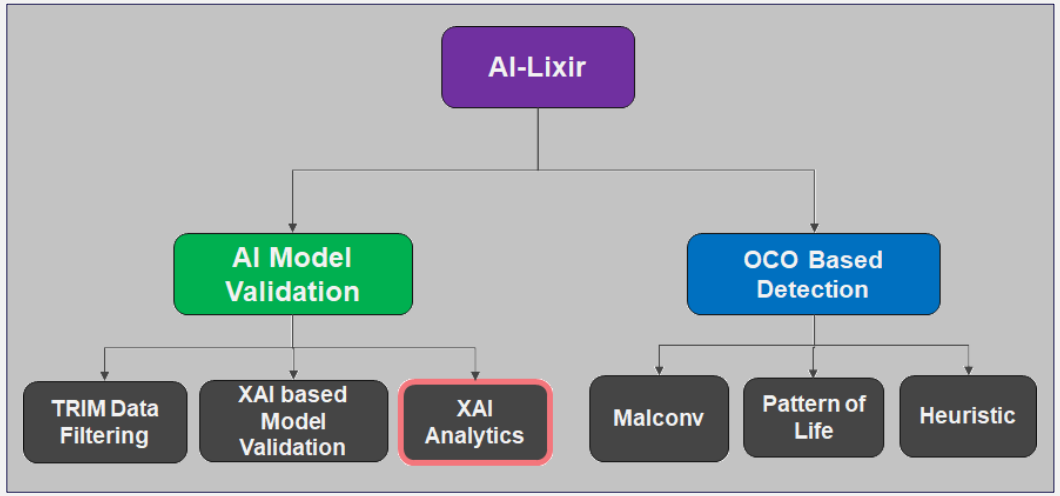
Programmatic heatmap analysis can detect poisoning even if a model passes accuracy tests

- Heatmaps are generated across a wide range of samples for comparison against baseline
- Vast quantities and visual complexity make human analysis time-consuming and impractical
- Automated approach provides speed and accuracy required for real-world applications



XAI Analytics – Poisoned MNIST Data

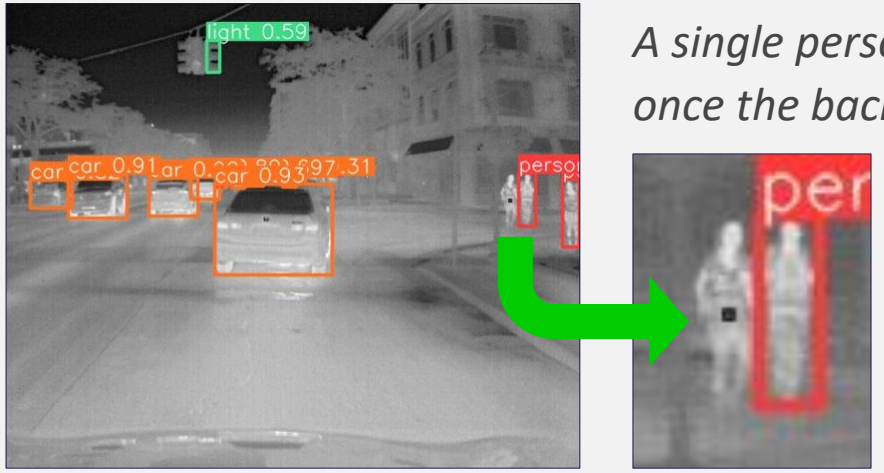
- Overall difference across the samples is calculated
- Difference score above the calculated threshold indicates anomalous behavior
- Variation algorithms can be tailored to optimize divergence detection



```
Updated model trained in 37.8 seconds
Average image variation: 0.034
Maximum image variation: 0.066
Model analysis indicates potential anomalous behavior
```

POC Data Poisoning Attack – FLIR Data

- Tested AI-Lixir using object detection and data from Teledyne FLIR thermal camera
  - Dataset contains over 26,000 frames with 520,000 bounding boxes for 15 categories
  - Model trained for detection using PyTorch YOLO framework
- Used variety of poisoning techniques to create multiple backdoors for testing
  - Prevented detection of all people using 3 techniques: Grid Overlay, Line Overlay, and Corner Blur
  - Prevented detection of single person while others still detected using “black square” technique shown below

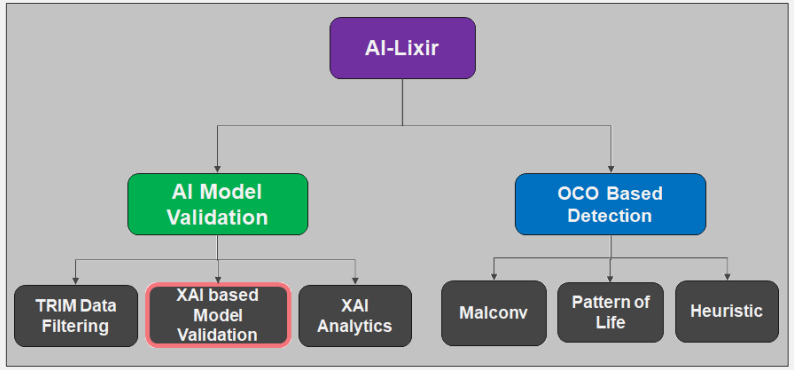


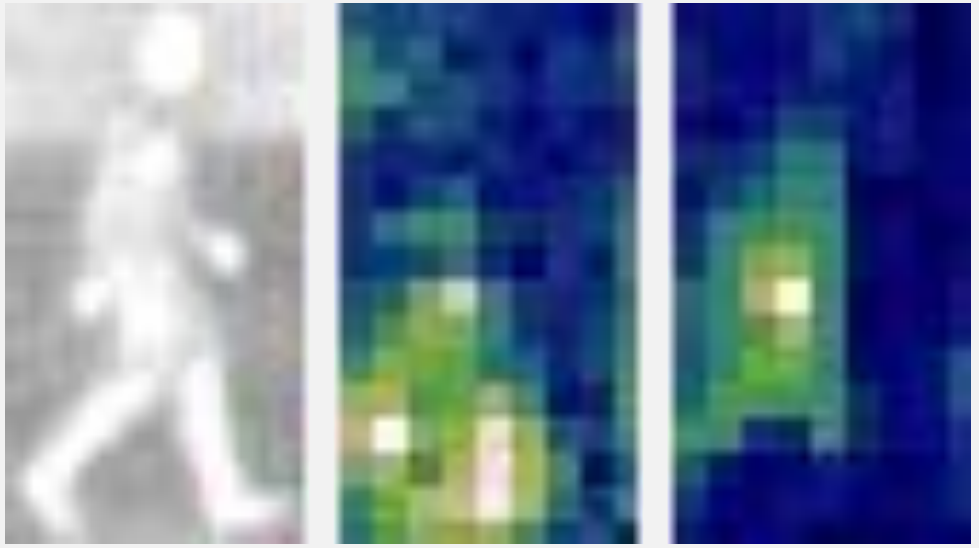
A single person escapes detection by poisoned model once the backdoor is introduced

- *Small black square near center of undetected person is the backdoor*
- *In real-world applications, this could be achieved by something worn to block heat*

XAI Based Model Validation – FLIR Data

- Heatmaps were generated for the FLIR data using the same occlusion-based XAI
  - Focused on the bounding boxes for people rather than the whole image since that is the only area of interest
- As shown in figure to the right, heatmaps reveal effects of poisoning
  - The clean model focuses on the lower part of the image, the person's legs
  - The poisoned model focuses on the center of the person, where the black-square backdoor was introduced



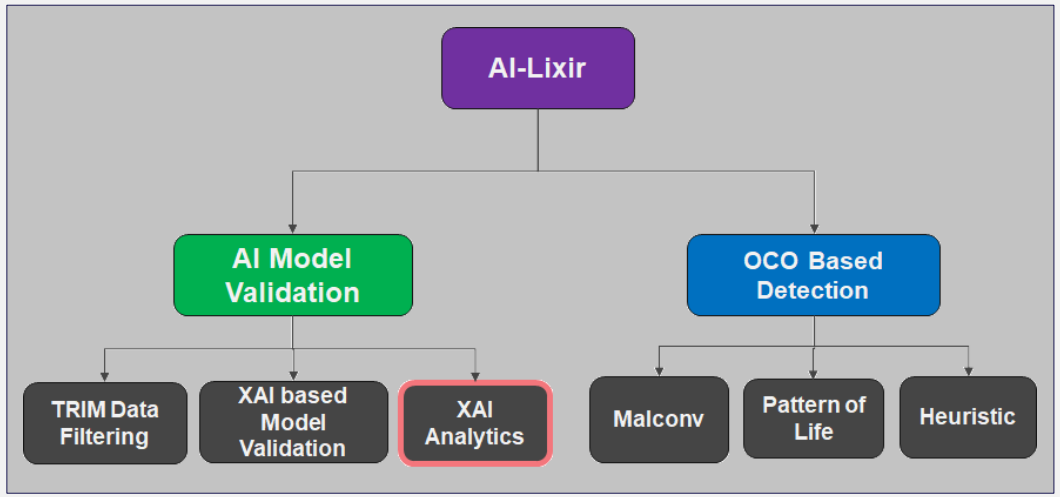


Sample Image *Clean Model* *Poisoned Model (removing single person)*

XAI Analytics – FLIR Data

AI-Lixir's XAI Analytics successfully detected the poisoned model

- Table below shows the programmatic analysis of FLIR data heatmaps
- Each poisoned model shows significant variation from the clean model (>50%)

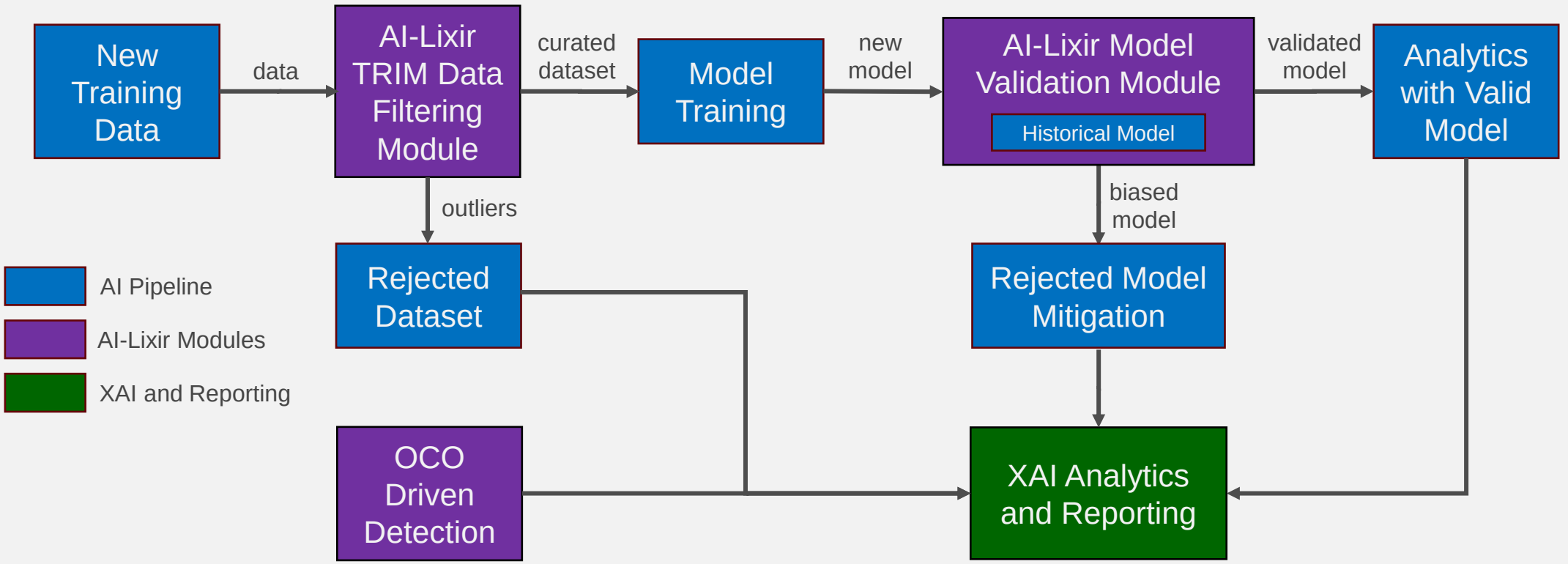


Model	Average Image Variation
Clean Model	0.017
Line Poisoned Model	0.026
Grid Poisoned Model	0.028
Corner Poisoned Model	0.062

Questions?

AI-Lixir Demonstration

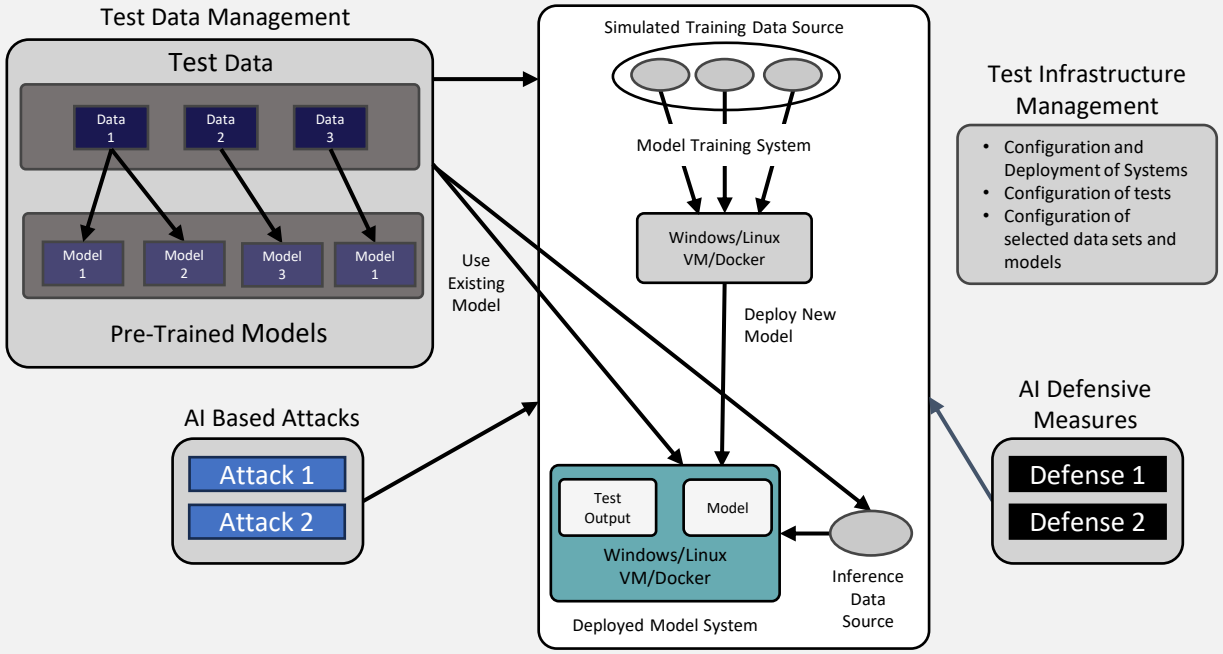
Example AI-Lixir Pipeline



Next Steps?

- Benchmark AI-Lixir capabilities against real mission data and models
  - Tailor analytics for mission-relevant data (e.g., variation thresholds)
  - Test where potential vulnerabilities exist
- Research additional techniques to defend the different stages of the AI Pipeline
  - Protection from adversarial samples during inference
  - Protection from model extraction attacks
- Develop automated responses to detected threats
  - Disconnect data source providing poisoned data
  - Revert to older models

AI Pipeline Test Framework



- The AI Pipeline test framework allows users to:
  - Validate their models after training with new data to verify model accuracy
  - Test their AI pipeline against known attacks to identify vulnerabilities
  - Test AI pipeline defensive measures against known attacks to measure their effectiveness
  - Test offensive attacks against the AI pipelines to measure their impact and find areas that may be exploitable
  - Easily test novel offensive attacks and defensive measures as they are developed
- Users can provide and configure:
  - Data
  - Models
  - Training and inference mechanisms
  - Offensive attacks
  - Defensive Measure
- When new instances of any of these are added, the configured tests can be launched automatically and generate a report of the current state of the system

Closing Thoughts

- Combined technologies of the AI-Lixir framework satisfy RAI security objectives
- Research opportunities are typically milestone-based
- Efforts to extend AI-Lixir to new data sets vary with data complexity (averaging 4-6 months for 2 engineers)
- VRL's product-based model reduces implementation timelines while keeping cost to a minimum

Contact

How to reach us

Our office

10500 Little Patuxent Pwky, Suite 500
Columbia, MD 21044
http://www.vrlsec.com

DK

David Klink, VP Business Development & Program Management
davidkl@vrlsec.com
301-322-0493

BM

Beth Mottar, Business Development & Program Management
bethm@vrlsec.com
410-916-5261