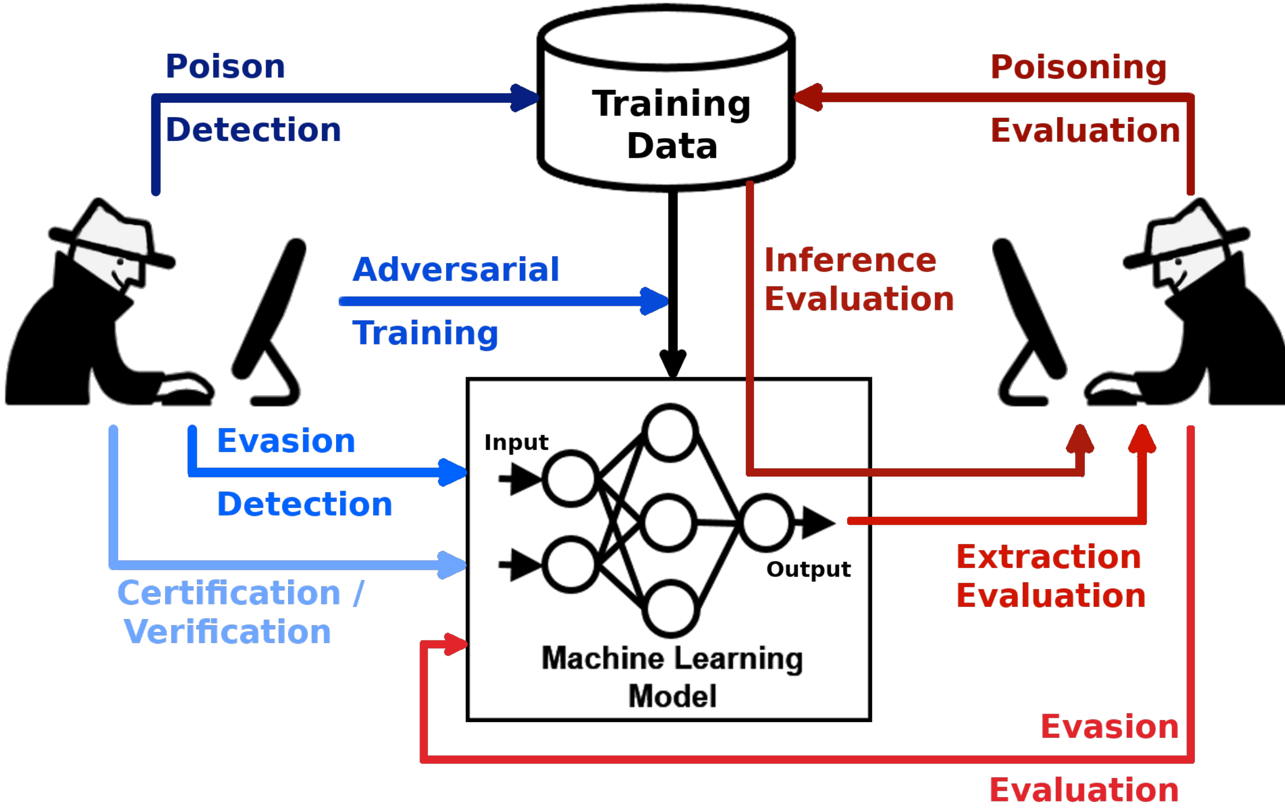
# Adversarial Robustness Toolbox (ART)

**Paul Houssel**
Visiting Research Student

INSTITUT POLYTECHNIQUE DE PARIS
ÉCOLE POLYTECHNIQUE
ENSTA
ENSAE
TELECOM Paris
TELECOM SudParis

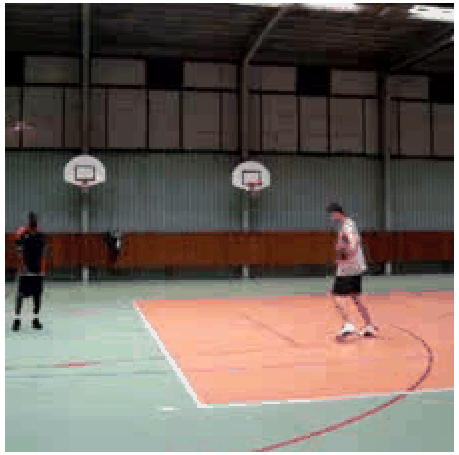
# Python Library for Model Evaluation and Defense

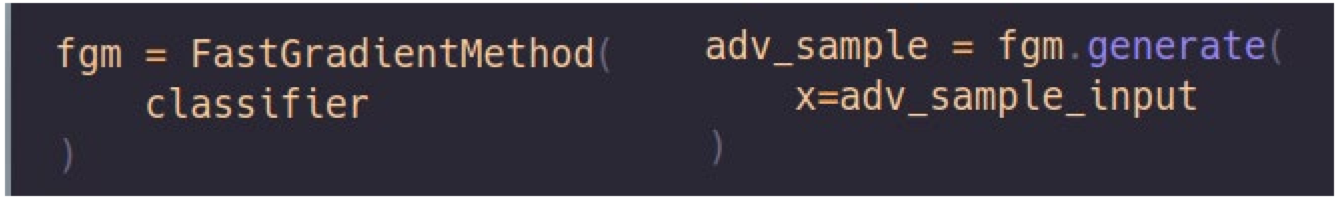
**Execute Attacks:** Evasion, Poisoning, Extraction, Inference
**On any model:** TensorFlow, Keras, PyTorch, MXNet, scikit-learn, XGBoost
**Trained on any Data:** images, tables, audio, video…



Implements the latest research & SOA in terms of AI attacks:
Benchmarking Solution

# In Action – Evade a Video Classifier





```
fgm = FastGradientMethod(
    classifier
)
```

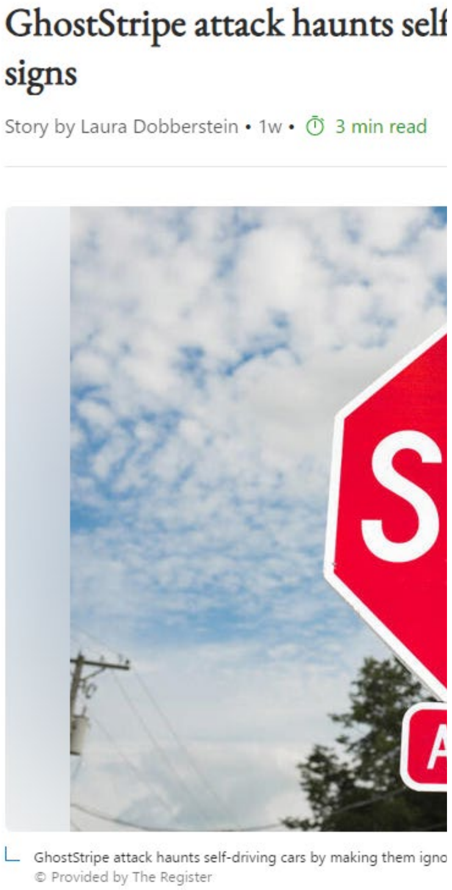
```
adv_sample = fgm.generate(
    x=adv_sample_input
)
```

1. **Load existing Training and inference pipeline**

**2. Load Evasion Attack**

**3. Generate Evasive Sample**

**4. Att&ck!**



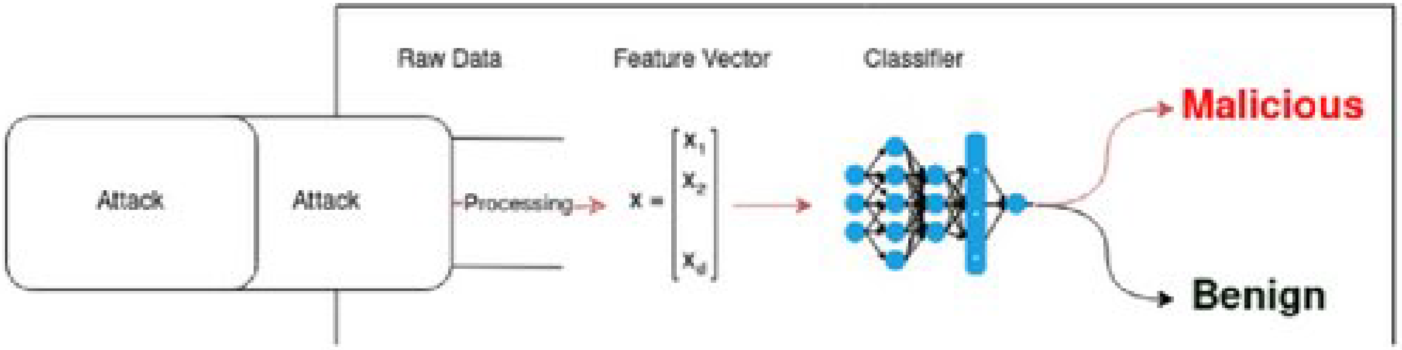
GhostStripe attack haunts self

signs

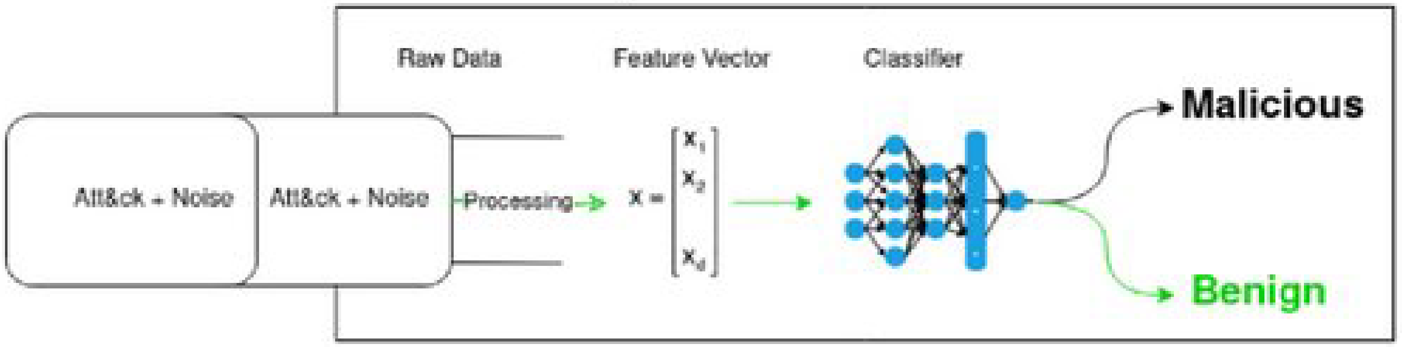
Story by Laura Dobberstein • 1w • 3 min read

GhostStripe attack haunts self-driving cars by making them igno

© Provided by The Register



(a) A network attack performed on a System is being classified as malicious by the Network-IDS.



(b) Adversarial traffic, obtained by perturbing the initial attack with minimal noise such that the attack gets misclassified as benign.

= "weasel"

# Limitations

**Problem Domain Constraints (Pierazzi et Al. 2020)**

- Available and Legal Problem Space Transformation
- Preserved Semantics
- Plausibility
- Robustness to Preprocessing

**Missing support for Large Language Models and other new architectures.**

# And what about you?

- Test your Models Robustness
- Test your Datasets Robustness
- Benchmark and compare against related work:
    - Defensive Approaches
    - Offensive Approaches