# URET: Universal Robustness Evaluation Toolkit (for Evasion)

**Kevin Eykholt**, Taesung Lee, Douglas Schales,
Jiyong Jang, Ian Molloy and Masha Zorin

IBM

UNIVERSITY OF CAMBRIDGE

# What is an Adversarial Attack?

Original Prediction: Panda

+

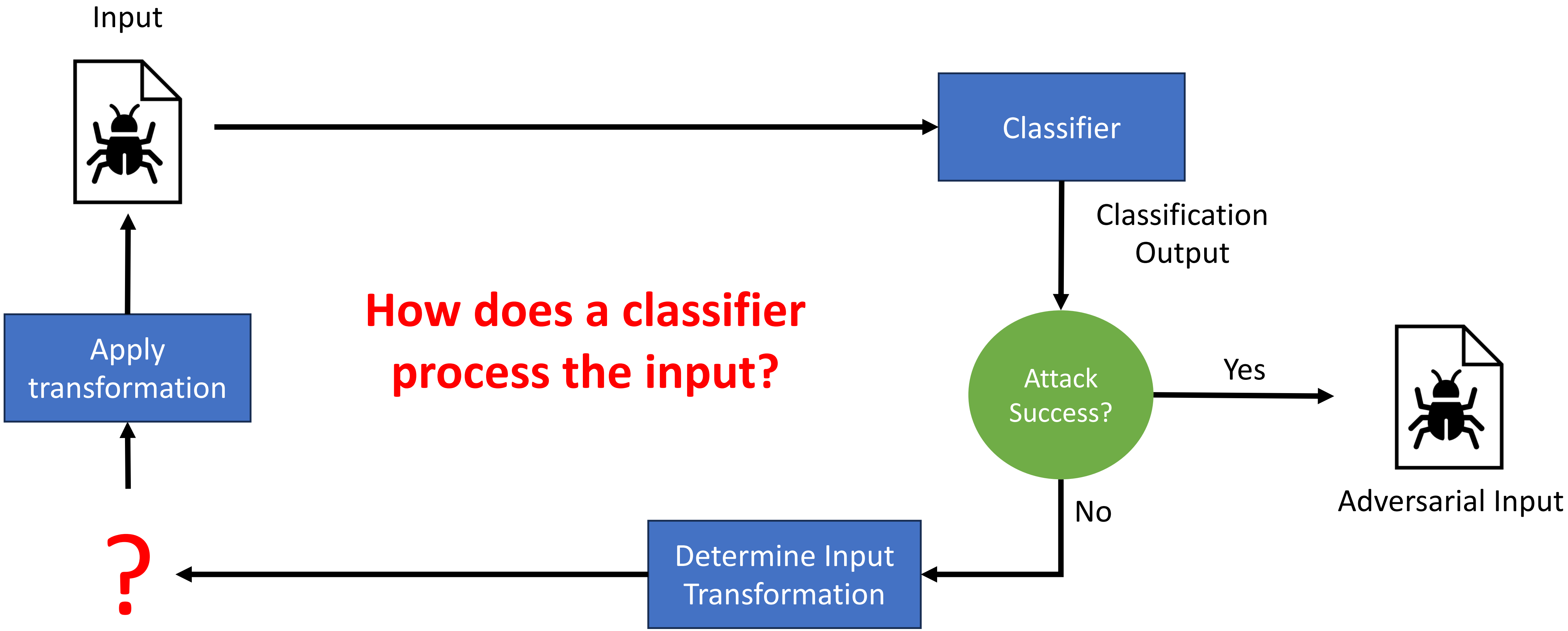
Adversarial Noise

=

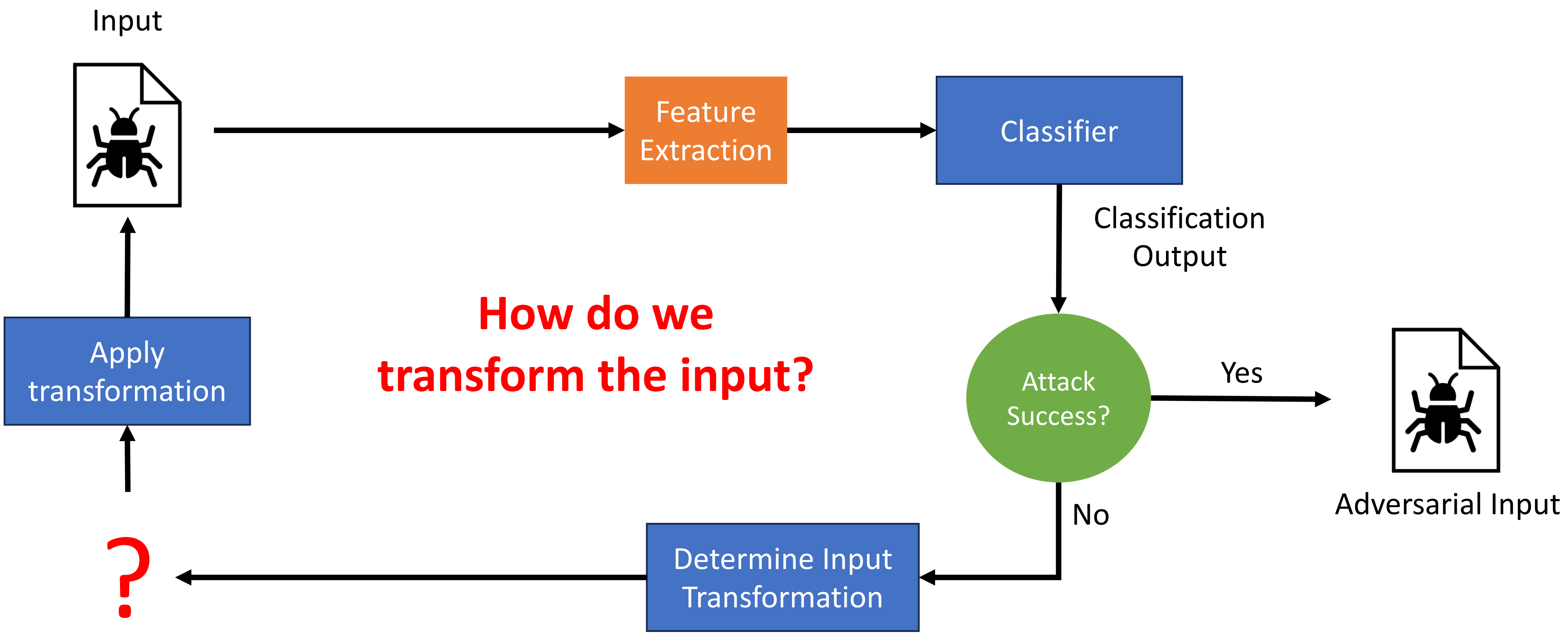
Adversarial Prediction: Gibbon

**Discover how to cause predictable errors in machine learning algorithms**

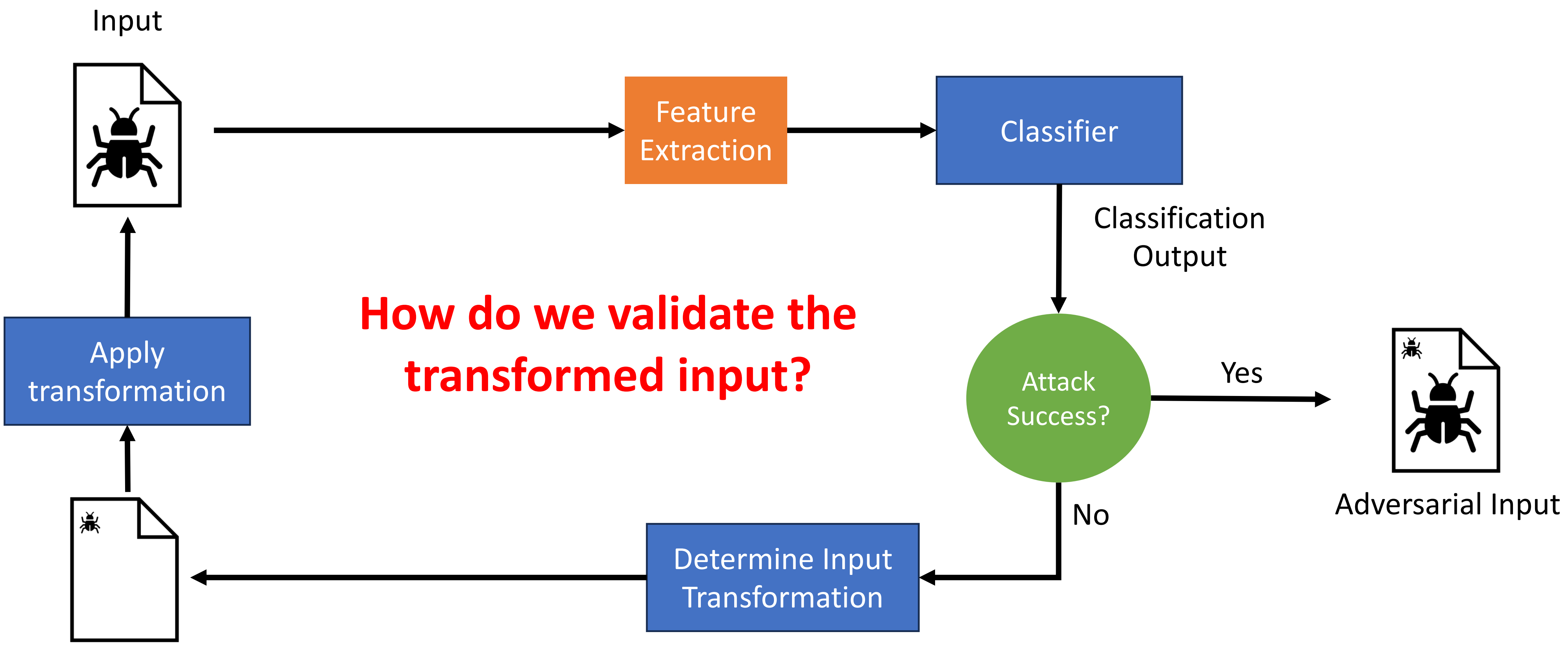
# Adversarial attacks aren’t generic!



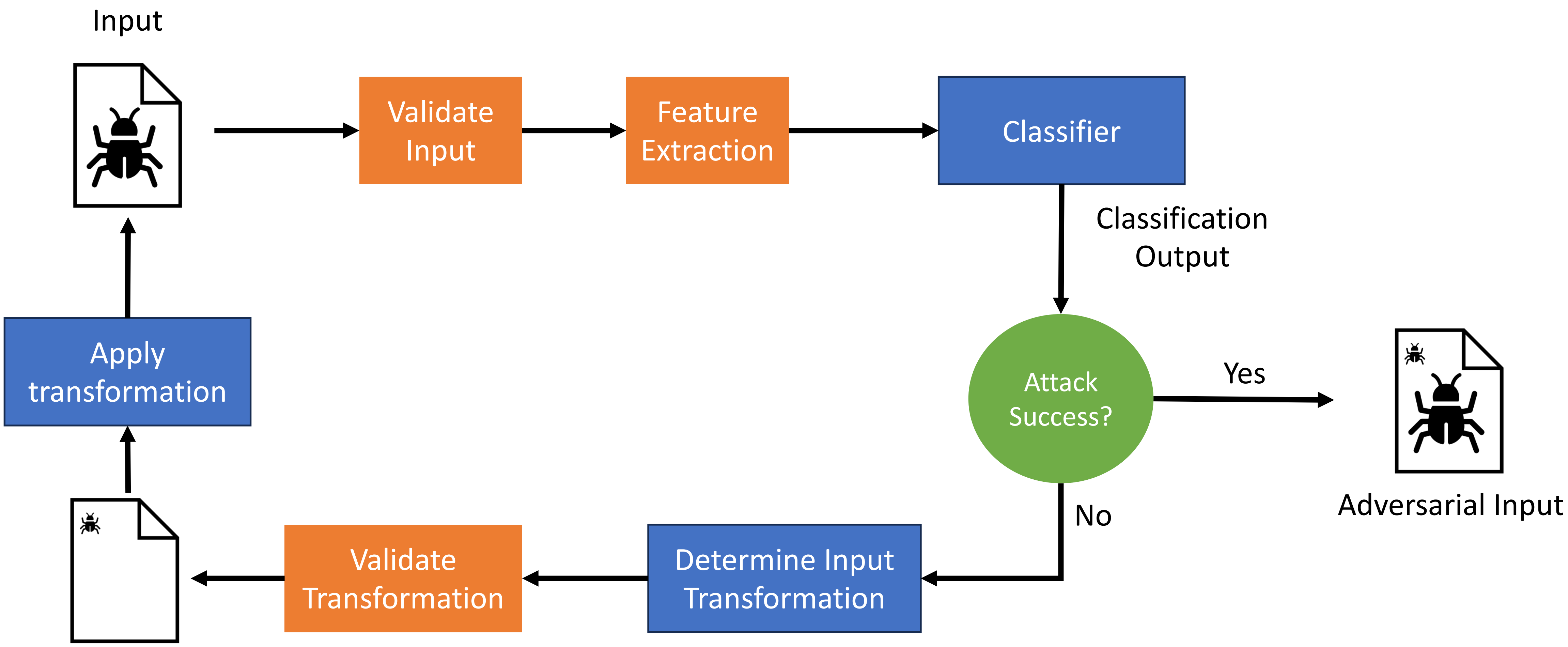
# Adversarial attacks aren’t generic!



# Adversarial attacks aren’t generic!



# A *Generic* Attack Pipeline



# Prior work

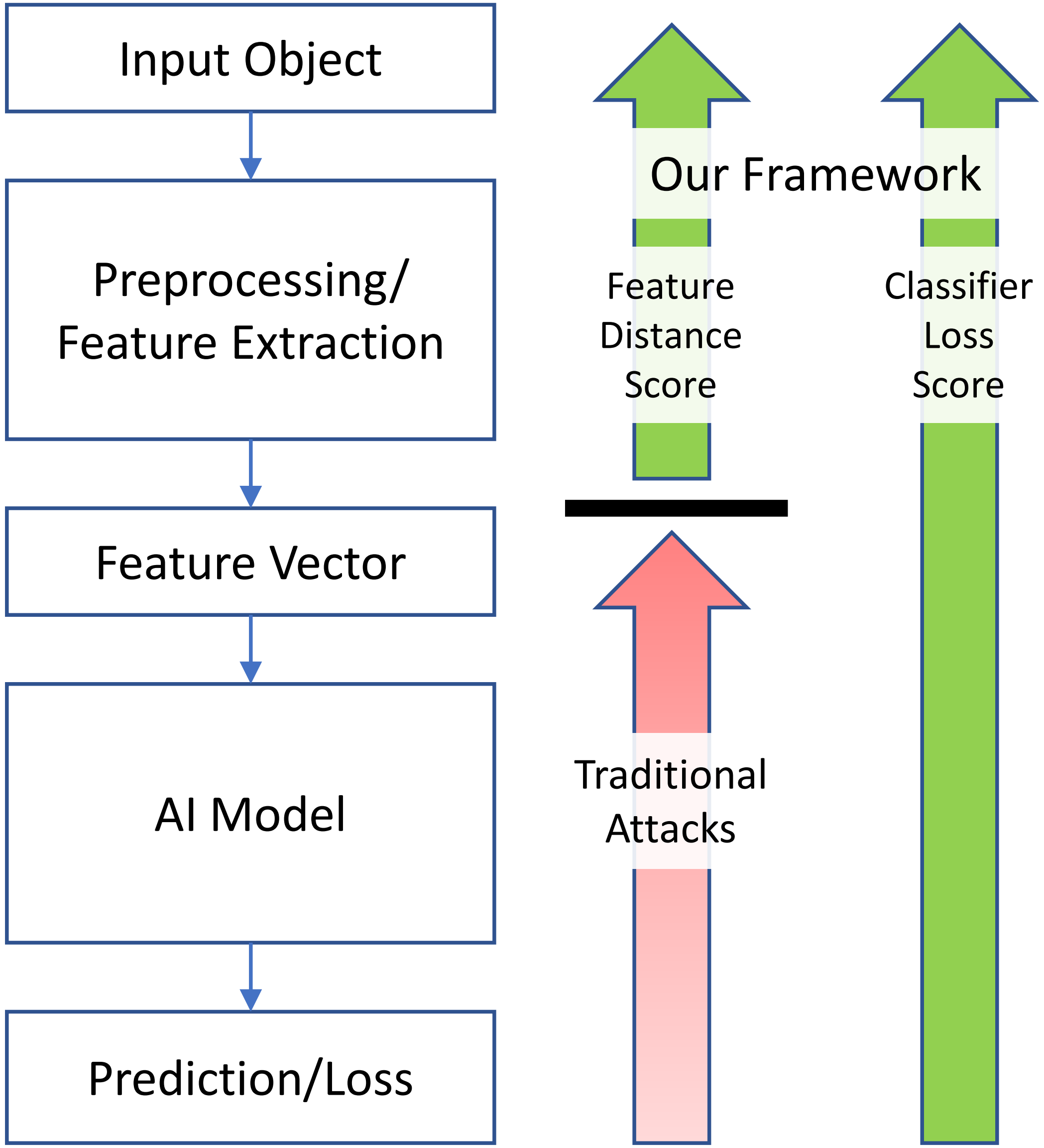
- **Not maintained** – Repository was mainly created to reproduce experiments
- **Limited in Scope** - Only supports a few input types or relies attacks designed for images
- **Hard to use** – Lack of a simple UI or documentation for the average user
- **Hard to access** – Code is kept closed-source or requires external approval

| Attacks | Input Types | | | Config Interface | Loss Objective | | Open Source |
|---|---|---|---|---|---|---|---|
| | Tabular | Text | Custom | | Model | Distance | |
| SLEIPNIR | X | X | **Malware** | X | ✓ | X | ~ |
| Gym-Malware | X | X | **Malware** | X | ✓ | X | ~ |
| Graph Search | ✓ | ~ | X | X | ✓ | ✓ | ~ |
| Pieraazi et al. | ✓ | ✓ | ✓ | **Unknown** | ✓ | X | ~ |
| Counterfit | X → ~* | ✓ | X | ✓ | ✓ | ✓ | ✓ |
| URET (Ours) | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |

* - This work added additional support after submission

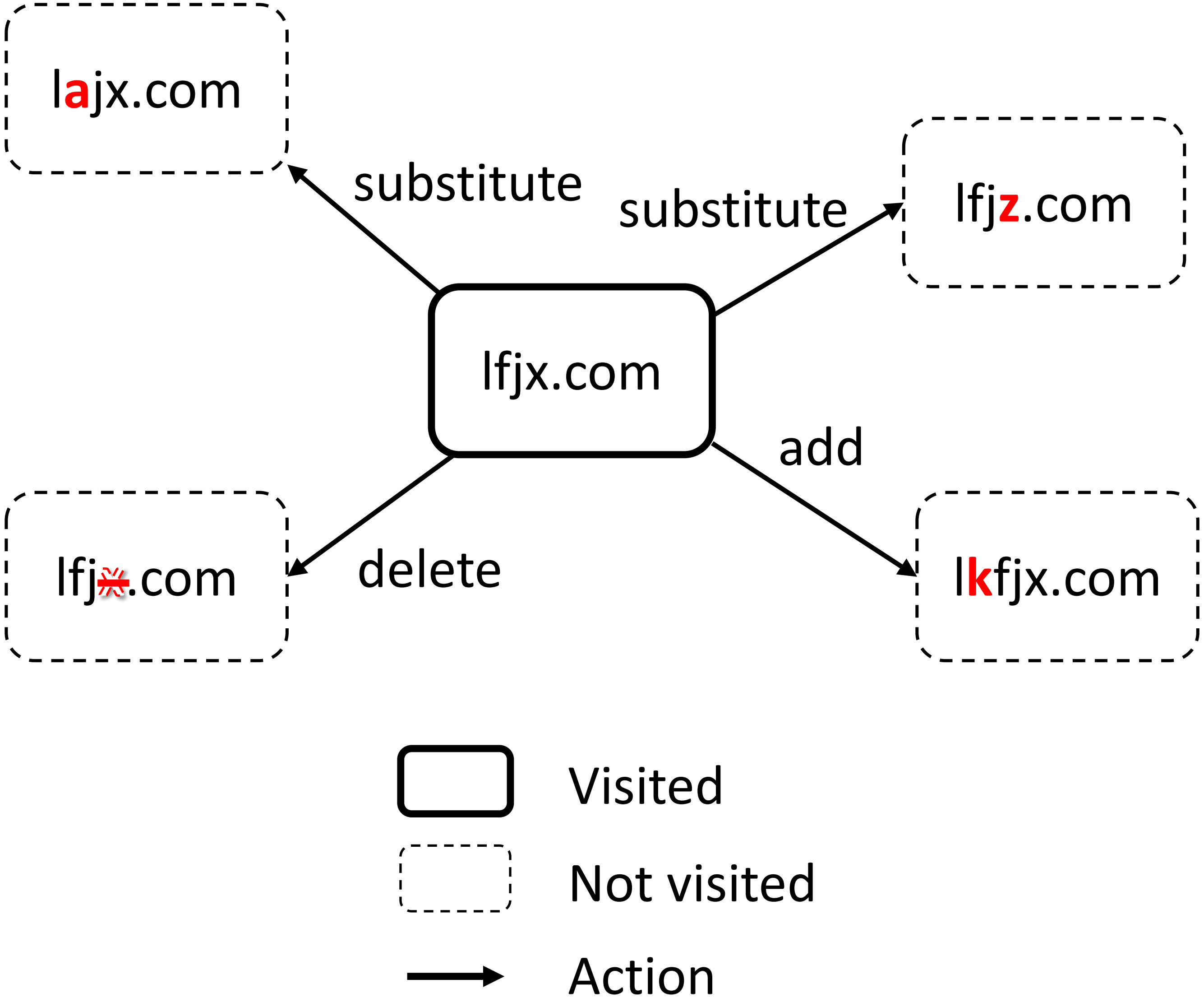
# What is URET?

- An end-to-end adversarial evasion attack framework for **any input type**
- Configuration files enable **quick, repeatable attack evaluations**
- Standardized interface to **support new, input types or tasks**



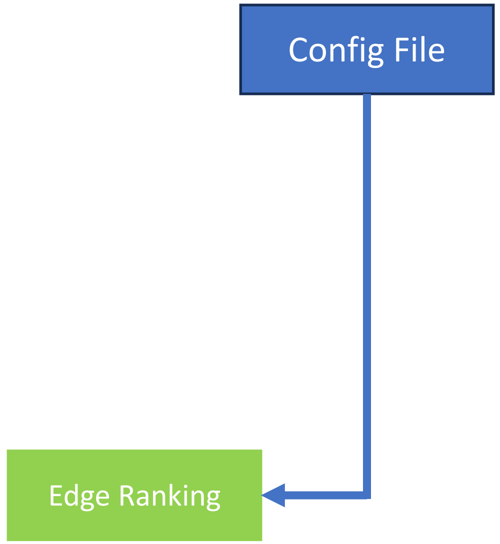
# How does it work?

- URET explores a graph to find sequences of edges to an adversarial input
- Nodes – Input states
- Edges – Input Transformations



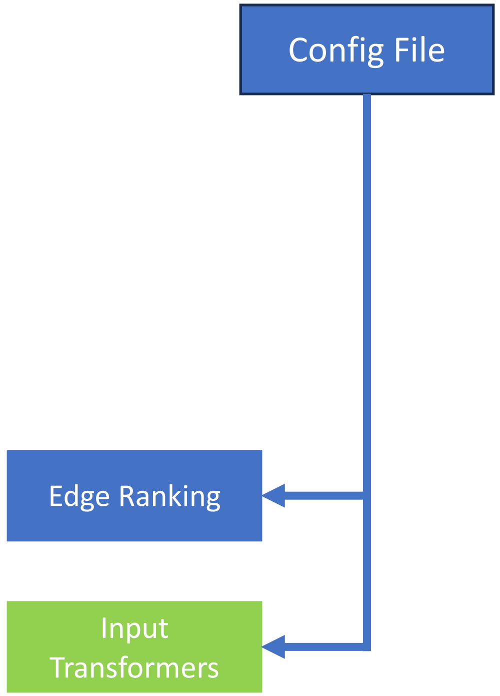
# Components - Edge ranking

- What edges should URET explore?
  - Random – Select random edges to explore
  - Brute Force – Explore every edge and select the highest fitness nodes
  - Lookup Table – Select highest fitness nodes based on prior transformation history
  - Model Guided – Select highest fitness nodes according to a model prediction



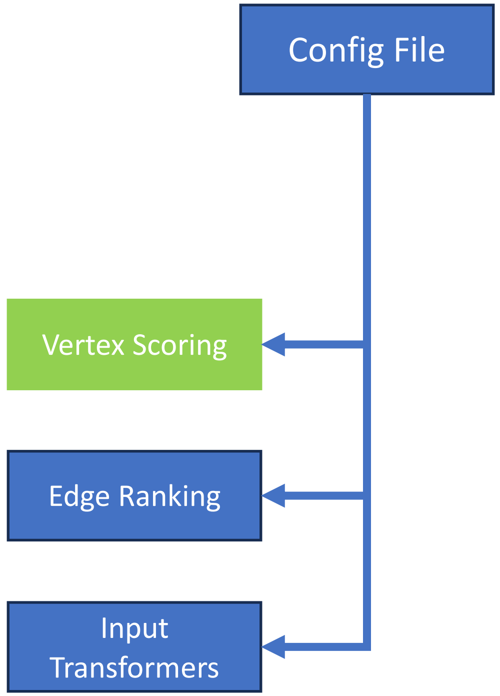
# Components – Input Transformers

- What are the edge types? How does URET transition between nodes?
- An input transformer is defined by its *transformation actions* and *constraints.*
  - Actions – How is the input transformed?
    - Text can be added, deleted, or substituted
    - Files can have their header modified
  - Constraints – What must be true about the transformed input so it is valid?
    - Text must use alphanumeric characters and not be empty
    - An input can only be transformed a certain number of times



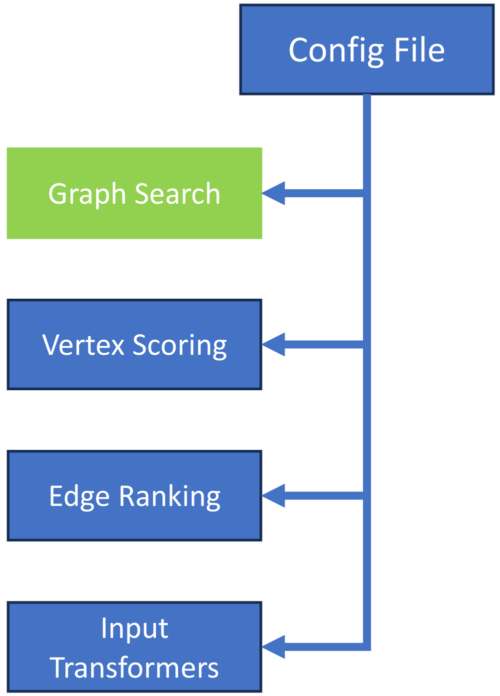
# Components – Vertex Scoring

- How is the fitness of a node evaluated? How *adversarial* is the node?
  - Classification Loss – Fitness is based on the classification loss as in traditional attacks
  - Distance Loss – Fitness based on the distance with respect to a certain target input state
- User can define their own customized scoring methods for URET to use as well



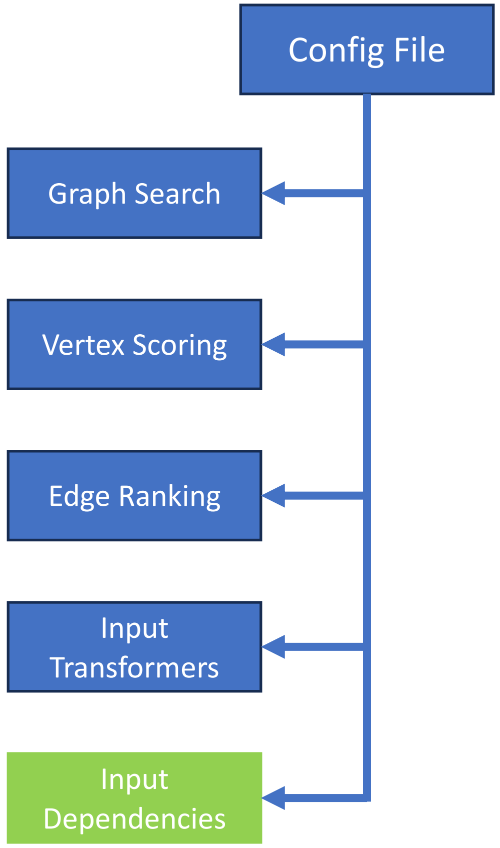
# Components – Graph Search

- What nodes should be kept for the next epoch?
  - Beam Search – Keep the top-k nodes
  - Simulated annealing – Keeps nodes based on the current temperature



# Components – Input Dependencies

- What must be true about the input after transforming its features?
- Dependencies enforce inter-feature constraints
- Examples:
  - The *total amount* feature must be equal to the sum of *savings* and *spending* for an input tracking finances
  - A numerical input may require that a subset of its features are normalized



# Using URET on Non-image data

- 2018 Housing Mortgage Disclosure Act (HMDA)
  - Based on the 13 features, predict if a mortgage application should be approved or rejected.
  - Evaluated using 2000 total samples correctly approved/rejected by a pre-trained classifier.
- Domain name generation algorithm (DGA) dataset
  - A domain name is converted into 20 numerical features.
  - Based on the numerical features, predict if a domain name is real or was generated by DGA
  - Evaluated using 10,000 total domain names either correctly predicted to be DGA or non-DGA by a pre-trained classifier.

| Classifier | Accuracy on Test Data | Accuracy on Evaluation Set |
|---|---|---|
| Decision Tree | 91% | 100% |
| Gradient Boosted | 95% | 100% |
| Logistic Regression | 69% | 100% |
| Random Forest | 81% | 100% |
| Multi-layer Perceptron | 83% | 100% |
| DGA | 97% | 100% |

# Results

| Model Arch. | Algorithm | Success Rate | Avg. # of Transforms | Avg. Time/sample |
|---|---|---|---|---|
| **Decision Tree** | Beam Search (Random) | 38% | 1.30 | 0.001 s |
| | Beam Search (Brute-Force) | 92% | 1.13 | 0.010 s |
| | Beam Search (Lookup Table) | 89% | 1.63 | 0.002 s |
| | Beam Search (Model Guided) | 81% | 1.85 | 0.018 s |
| | Simulated Annealing | 97% | 1.87 | 1.000 s |
| **Gradient Boosted Classifier** | Beam Search (Random) | 14% | 1.43 | 0.003 s |
| | Beam Search (Brute-Force) | 58% | 1.08 | 0.044 s |
| | Beam Search (Lookup Table) | 26% | 1.41 | 0.026 s |
| | Beam Search (Model Guided) | 52% | 1.74 | 0.058 s |
| | Simulated Annealing | 57% | 2.00 | 1.000 s |
| **Logistic Regression** | Beam Search (Random) | 34% | 1.38 | 0.002 s |
| | Beam Search (Brute-Force) | 100% | 1.05 | 0.007 s |
| | Beam Search (Lookup Table) | 69% | 1.12 | 0.007 s |
| | Beam Search (Model Guided) | 88% | 1.93 | 0.020 s |
| | Simulated Annealing | 100% | 2.00 | 1.000 s |
| **Random Forest** | Beam Search (Random) | 27% | 1.46 | 0.352 s |
| | Beam Search (Brute-Force) | 100% | 1.04 | 1.462 s |
| | Beam Search (Lookup Table) | 70% | 1.08 | 1.177 s |
| | Beam Search (Model Guided) | 86% | 1.96 | 0.042 s |
| | Simulated Annealing | 75% | 1.87 | 1.000 s |
| **Multi-Layer Perceptron** | Beam Search (Random) | 36% | 1.41 | 0.198 s |
| | Beam Search (Brute-Force) | 100% | 1.04 | 0.724 s |
| | Beam Search (Lookup Table) | 94% | 1.39 | 0.369 s |
| | Beam Search (Model Guided) | 71% | 1.92 | 0.297 s |
| | Simulated Annealing | 97% | 1.90 | 1.000 s |

HMDA results – URET could transform 7 of the 13 features

# URET is pretty good

| Model Arch. | Algorithm | Success Rate | Avg. # of Transforms | Avg. Time/sample |
|---|---|---|---|---|
| **Decision Tree** | Beam Search (Random) | 38% | 1.30 | 0.001 s |
| | Beam Search (Brute-Force) | 92% | 1.13 | 0.010 s |
| | Beam Search (Lookup Table) | 89% | 1.63 | 0.002 s |
| | Beam Search (Model Guided) | 81% | 1.85 | 0.018 s |
| | Simulated Annealing | 97% | 1.87 | 1.000 s |
| **Gradient Boosted Classifier** | Beam Search (Random) | 14% | 1.43 | 0.003 s |
| | Beam Search (Brute-Force) | 58% | 1.08 | 0.044 s |
| | Beam Search (Lookup Table) | 26% | 1.41 | 0.026 s |
| | Beam Search (Model Guided) | 52% | 1.74 | 0.058 s |
| | Simulated Annealing | 57% | 2.00 | 1.000 s |
| **Logistic Regression** | Beam Search (Random) | 34% | 1.38 | 0.002 s |
| | Beam Search (Brute-Force) | 100% | 1.05 | 0.007 s |
| | Beam Search (Lookup Table) | 69% | 1.12 | 0.007 s |
| | Beam Search (Model Guided) | 88% | 1.93 | 0.020 s |
| | Simulated Annealing | 100% | 2.00 | 1.000 s |
| **Random Forest** | Beam Search (Random) | 27% | 1.46 | 0.352 s |
| | Beam Search (Brute-Force) | 100% | 1.04 | 1.462 s |
| | Beam Search (Lookup Table) | 70% | 1.08 | 1.177 s |
| | Beam Search (Model Guided) | 86% | 1.96 | 0.042 s |
| | Simulated Annealing | 75% | 1.87 | 1.000 s |
| **Multi-Layer Perceptron** | Beam Search (Random) | 36% | 1.41 | 0.198 s |
| | Beam Search (Brute-Force) | 100% | 1.04 | 0.724 s |
| | Beam Search (Lookup Table) | 94% | 1.39 | 0.369 s |
| | Beam Search (Model Guided) | 71% | 1.92 | 0.297 s |
| | Simulated Annealing | 97% | 1.90 | 1.000 s |

HMDA results – URET could transform 7 of the 13 features

# Can trade performance for speed

| Model Arch. | Algorithm | Success Rate | Avg. # of Transforms | Avg. Time/sample |
|---|---|---|---|---|
| **Decision Tree** | Beam Search (Random) | 38% | 1.30 | 0.001 s |
| | Beam Search (Brute-Force) | 92% | 1.13 | 0.010 s |
| | Beam Search (Lookup Table) | 89% | 1.63 | 0.002 s |
| | Beam Search (Model Guided) | 81% | 1.85 | 0.018 s |
| | Simulated Annealing | 97% | 1.87 | 1.000 s |
| **Gradient Boosted Classifier** | Beam Search (Random) | 14% | 1.43 | 0.003 s |
| | Beam Search (Brute-Force) | 58% | 1.08 | 0.044 s |
| | Beam Search (Lookup Table) | 26% | 1.41 | 0.026 s |
| | Beam Search (Model Guided) | 52% | 1.74 | 0.058 s |
| | Simulated Annealing | 57% | 2.00 | 1.000 s |
| **Logistic Regression** | Beam Search (Random) | 34% | 1.38 | 0.002 s |
| | Beam Search (Brute-Force) | 100% | 1.05 | 0.007 s |
| | Beam Search (Lookup Table) | 69% | 1.12 | 0.007 s |
| | Beam Search (Model Guided) | 88% | 1.93 | 0.020 s |
| | Simulated Annealing | 100% | 2.00 | 1.000 s |
| **Random Forest** | Beam Search (Random) | 27% | 1.46 | 0.352 s |
| | Beam Search (Brute-Force) | 100% | 1.04 | 1.462 s |
| | Beam Search (Lookup Table) | 70% | 1.08 | 1.177 s |
| | Beam Search (Model Guided) | 86% | 1.96 | 0.042 s |
| | Simulated Annealing | 75% | 1.87 | 1.000 s |
| **Multi-Layer Perceptron** | Beam Search (Random) | 36% | 1.41 | 0.198 s |
| | Beam Search (Brute-Force) | 100% | 1.04 | 0.724 s |
| | Beam Search (Lookup Table) | 94% | 1.39 | 0.369 s |
| | Beam Search (Model Guided) | 71% | 1.92 | 0.297 s |
| | Simulated Annealing | 97% | 1.90 | 1.000 s |

HMDA results – URET could transform 7 of the 13 features

# Can make exploration consistent

| Model Arch. | Algorithm | Success Rate | Avg. # of Transforms | Avg. Time/sample |
|---|---|---|---|---|
| **Decision Tree** | Beam Search (Random) | 38% | 1.30 | 0.001 s |
| | Beam Search (Brute-Force) | 92% | 1.13 | 0.010 s |
| | Beam Search (Lookup Table) | 89% | 1.63 | 0.002 s |
| | Beam Search (Model Guided) | 81% | 1.85 | 0.018 s |
| | Simulated Annealing | 97% | 1.87 | $\boxed{1.000 \text{ s}}$ |
| **Gradient Boosted Classifier** | Beam Search (Random) | 14% | 1.43 | 0.003 s |
| | Beam Search (Brute-Force) | 58% | 1.08 | 0.044 s |
| | Beam Search (Lookup Table) | 26% | 1.41 | 0.026 s |
| | Beam Search (Model Guided) | 52% | 1.74 | 0.058 s |
| | Simulated Annealing | 57% | 2.00 | $\boxed{1.000 \text{ s}}$ |
| **Logistic Regression** | Beam Search (Random) | 34% | 1.38 | 0.002 s |
| | Beam Search (Brute-Force) | 100% | 1.05 | 0.007 s |
| | Beam Search (Lookup Table) | 69% | 1.12 | 0.007 s |
| | Beam Search (Model Guided) | 88% | 1.93 | 0.020 s |
| | Simulated Annealing | 100% | 2.00 | $\boxed{1.000 \text{ s}}$ |
| **Random Forest** | Beam Search (Random) | 27% | 1.46 | 0.352 s |
| | Beam Search (Brute-Force) | 100% | 1.04 | 1.462 s |
| | Beam Search (Lookup Table) | 70% | 1.08 | 1.177 s |
| | Beam Search (Model Guided) | 86% | 1.96 | 0.042 s |
| | Simulated Annealing | 75% | 1.87 | $\boxed{1.000 \text{ s}}$ |
| **Multi-Layer Perceptron** | Beam Search (Random) | 36% | 1.41 | 0.198 s |
| | Beam Search (Brute-Force) | 100% | 1.04 | 0.724 s |
| | Beam Search (Lookup Table) | 94% | 1.39 | 0.369 s |
| | Beam Search (Model Guided) | 71% | 1.92 | 0.297 s |
| | Simulated Annealing | 97% | 1.90 | $\boxed{1.000 \text{ s}}$ |

HMDA results – URET could transform 7 of the 13 features

# Switching domains isn't a problem

| Algorithm | Success rate | Avg. # of Transforms | Avg. Time / sample |
|---|---|---|---|
| Beam Search (Random) | 23% | 1.84 | 0.093 s |
| Beam Search (Brute-Force) | 85% | 1.24 | 0.363 s |
| Beam Search (Lookup Table) | 45% | 1.61 | 0.277 s |
| Beam Search (Model Guided) | 70% | 2.56 | 0.400 s |
| Simulated Annealing | 62% | 2.28 | 1.000 s |

DGA Results – Generating adversarial text examples

with a classification loss scoring function.

# Reversing feature space modifications can be tricky

| Algorithm | Success rate | Avg. # of Transforms | Avg. Time / sample |
|---|---|---|---|
| Beam Search (Random) | 23% | 1.84 | 0.093 s |
| Beam Search (Brute-Force) | 85% | 1.24 | 0.363 s |
| Beam Search (Lookup Table) | 45% | 1.61 | 0.277 s |
| Beam Search (Model Guided) | 70% | 2.56 | 0.400 s |
| Simulated Annealing | 62% | 2.28 | 1.000 s |

DGA Results – Generating adversarial text examples with a classification loss scoring function.

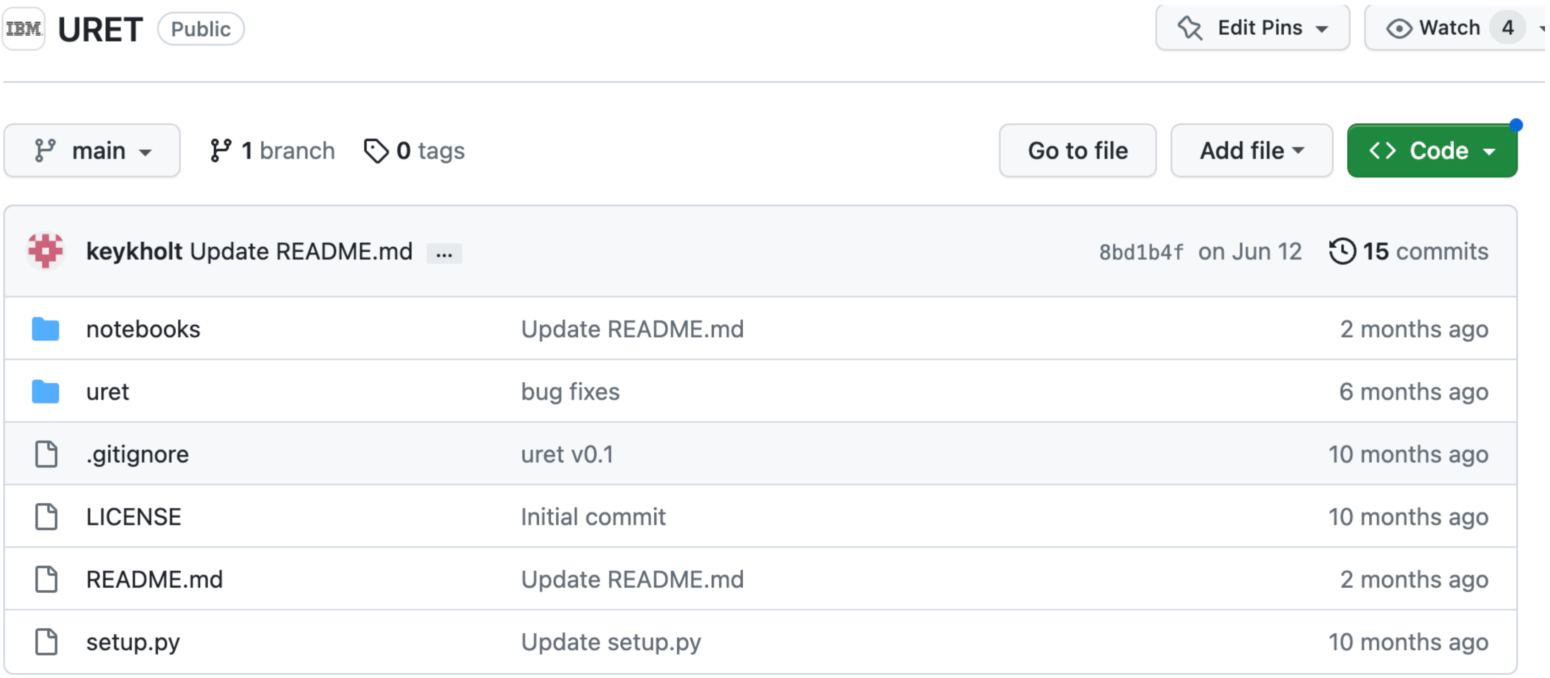
| Algorithm | Success rate | Avg. # of Transforms | Avg. Time / sample |
|---|---|---|---|
| Beam Search (Random) | 27% | 1.87 | 0.091 s |
| Beam Search (Brute-Force) | 56% | 1.93 | 22.835 s |
| Beam Search (Lookup Table) | 50% | 1.79 | 12.415 s |
| Beam Search (Model Guided) | 43% | 2.69 | 0.606 s |
| Simulated Annealing | 26% | 2.72 | 1.000 s |

DGA Results – Generating adversarial numerical feature vectors with a feature distance scoring function

Going from 3 transformations to 13 transformation per node

# *Don’t be obscure, be flexible*

- To properly evaluate and address AI vulnerabilities, we need penetration testing tools *for more than just images*




Contact me: kheykholt@ibm.com

Interested in using URET?