COUNTERFIT

Attacking Machine Learning in Blackbox Settings

@moo_hax

@rdheeko

Work for Azure Trustworthy ML

Research, Dev, Ops, etc

What is Machine Learning?

- Set of techniques that aim to model a problem mathematically:

  Stats + Maths + Computers + BIG data

- Predictions without explicit programming

- Growing fast:

  compute, data aggregation, code frameworks

... Still magic ... But mostly math

Why Do We Care?

- It's coming to a field near you!
  - Has implications if you're a security person in that field...
- No longer a math problem, it's an engineering problem
  - Algorithms are empty.
- Building relationships in non-congruent data
  - Represent your data as numbers.
- Can be as complex or as simple as you want to make it
  - Apply the right technique to the right problem


APPLE MICROSOFT GOOGLE
Personal voice assistants struggle with black voices, new study shows
Stanford researchers found that speech recognition algorithms disproportionately misunderstood black speakers
Does GPT-2 Know Your Phone Number?
Eric Wallace, Florian Tramèr, Matthew Jagielski, and Ariel Herbert-Voss
Dec 20, 2020
s, is a knowledge base of adv
esearch. ATLAS is modeled a
earning systems. ML is increa
Back
Machine Learning for Red Teams, Part 1
November 14, 2018 | Will Pearce
AI / DEEP LEARNING | DATA SCIENCE
Learning to Defend AI Deployments Us
Exploit Simulation Environment
By Nathan Schwartz
Tags: Cybersecurity / Fraud Detection, Machine L
NGC
Are driverless cars safe? Uber fatality raises questions
After a woman is killed by a self-driving car in Arizona, police investigate
The AI Incident Database wants to improve the safety of machine learning
Ben Dickson @BenD
Poisoning GitHub Copilot and Machine Learning
07 Jul 2021
AI Artificial Intelligence Code L
Stale sessions, ML poisoning among 2021's top security threats
An all-star security panel at RSA Conference discusses the biggest issues facing companies today and
months
Never a dill moment: Exploiting machine learning pickle files
POST MARCH 15, 2021 LEAVE A COMMENT
By Evan Sultanik
Centrelink debt scandal: report reve
multiple failures in welfare system
Attackers want to exploit and abuse your AI
and data, it
Author : Lengwadishang
Technobyte: A man who got fired by a machine
3 years ago
Tesla tricked into speeding by researchers using electrical tape
BY KATE GIBSON
FEBRUARY 19, 2020 / 1:47 PM / MONEYWATCH
18 July 2019
Cylance, I Kill You!
d Endpoint Protection
Exploiting AI
How Cybercriminals Misuse and Abuse AI and ML
We discuss the present state of the malicious uses and abuses of AI and ML and the plausible future scenarios in which cybercriminals might abuse these technologies for ill gain.
Microsoft Chat Bot Goes On Racist, Genocidal Twitter Rampage
Seriously? Seriously.

Offensive ML

"Application of ML techniques to offensive problems"

Abusing control relationships
Obfuscating C2 as English
Detecting sandbox environments
Improving phishing
Faster password guessing
Metasploit exploit selection
Automating timing attacks

File share path completion
Proc injection technique selection
OpSec improvements
Command recommendations
Report writing
Active Directory clustering
Staging decisions

So. Many. More.

Adversarial ML

"Subdiscipline that specifically attacks ML algorithms"

Find PII in large language models
Good word attacks on classifiers
RL attacks on classifiers
Denial of Service with sponge examples
Functional Extraction

New "techniques" drop all the time.

Thanks Professor!

Getting Started with ML

Google "ML [literally anything] tutorial"

"Detecting Cats in Images with OpenCV"
"Auto-Generating Clickbait with RNNs"

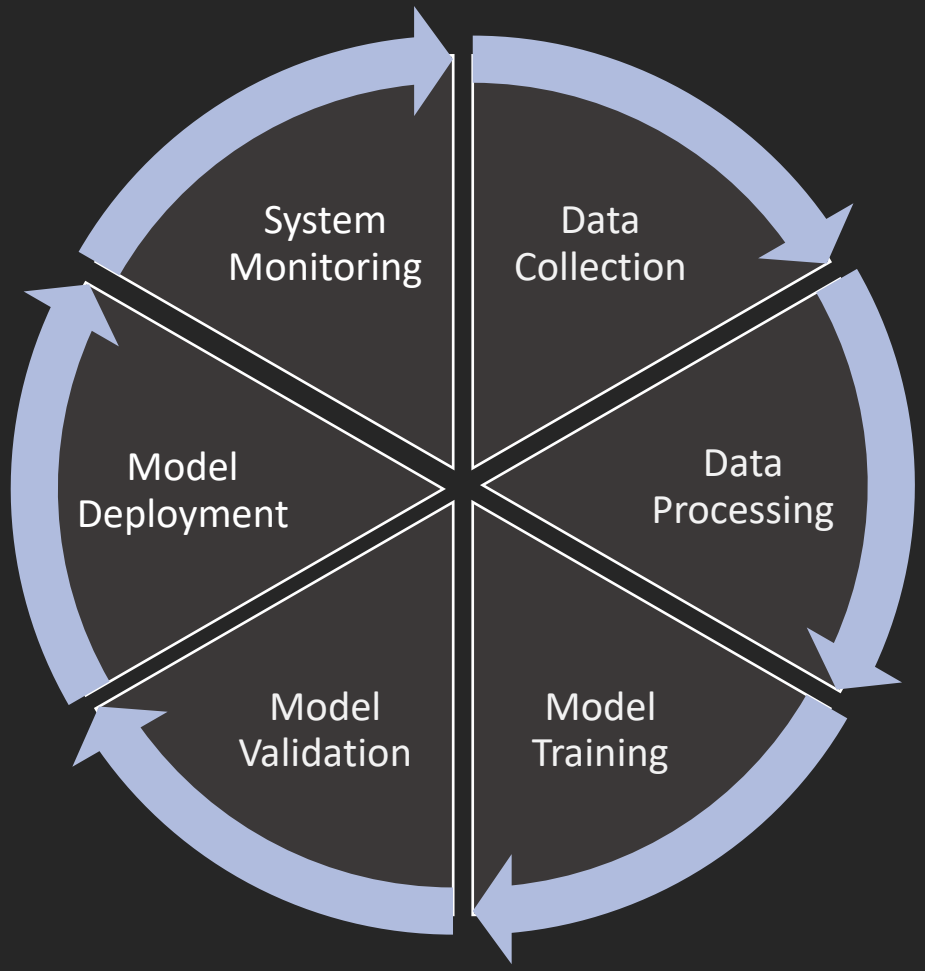
https://github.com/ujjwalkarn/Machine-Learning-Tutorials/ (320+ links)
https://sgfin.github.io/learning-resources/ (200+ links)
https://github.com/josephmisiti/awesome-machine-learning (200+ links)
https://github.com/awesomedata/awesome-public-datasets (410+ links)

Getting Started with ML

1. Data is everywhere, what should we collect?
    - What data is used by a human to solve/perform a task?
3. Process the data and extract useful features.
4. Download Python + [ML stuff]. (Might need some GPUs)
    - NumPy / Pandas – Data processing and matrices
    - SciKit-Learn – Data analytics + Basic algorithmic techniques
    - TensorFlow/PyTorch – Full blown ML framework from Google
    - Keras – High-level wrapper for TensorFlow
5. Write a 10-line script and ML your heart out

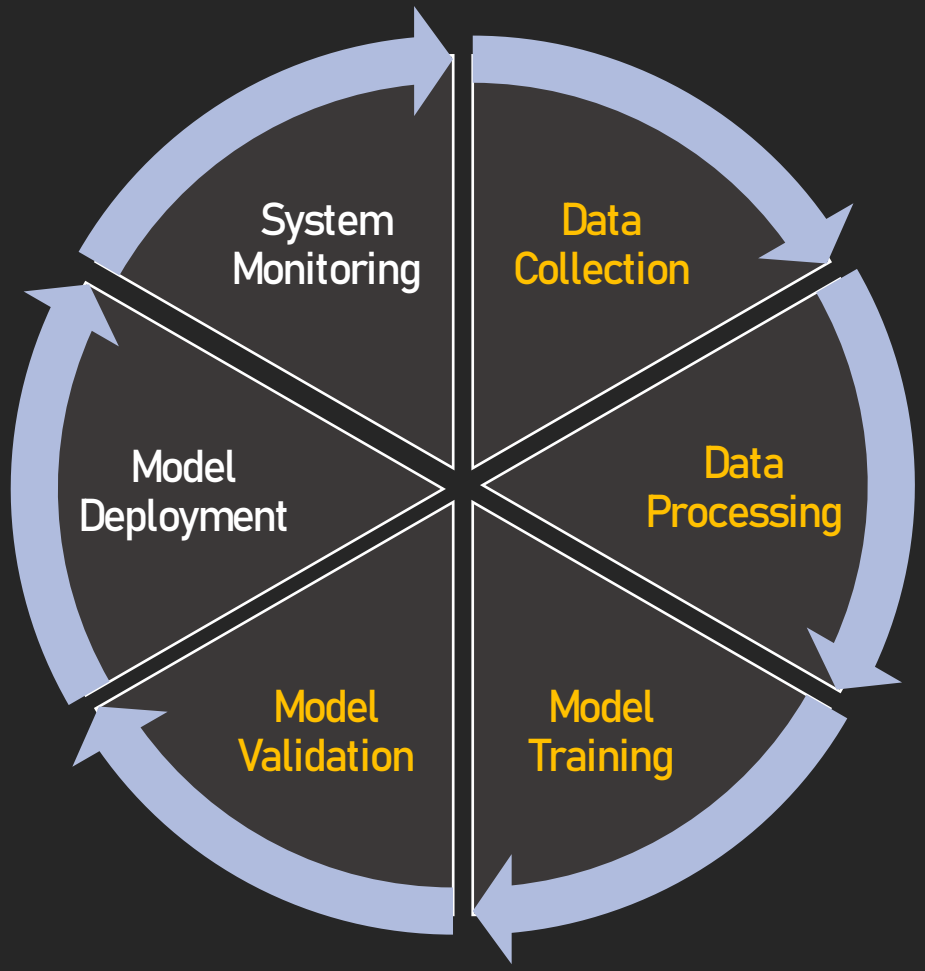
Attacks on MLDLC

How’s that for an acronym?



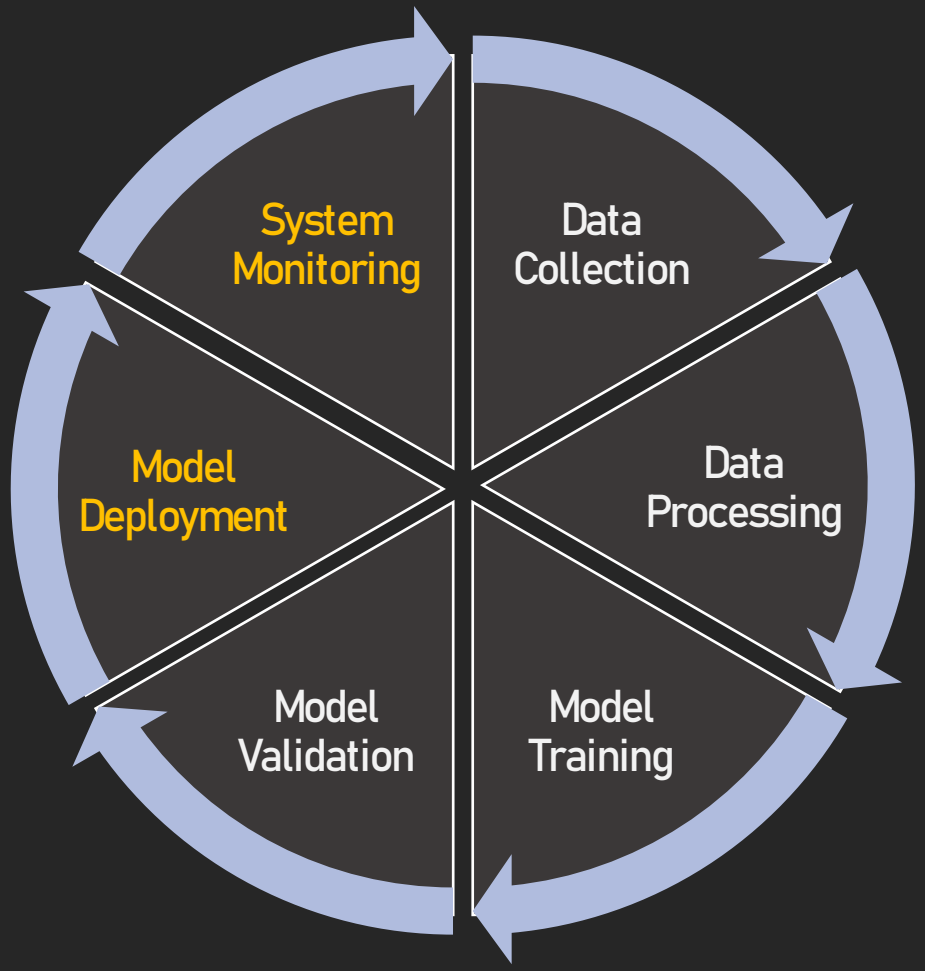
Extraction
Evasion
Inversion
Inference
Poisoning

Train Time Attacks



Extraction
Evasion
Inversion
Inference
Poisoning

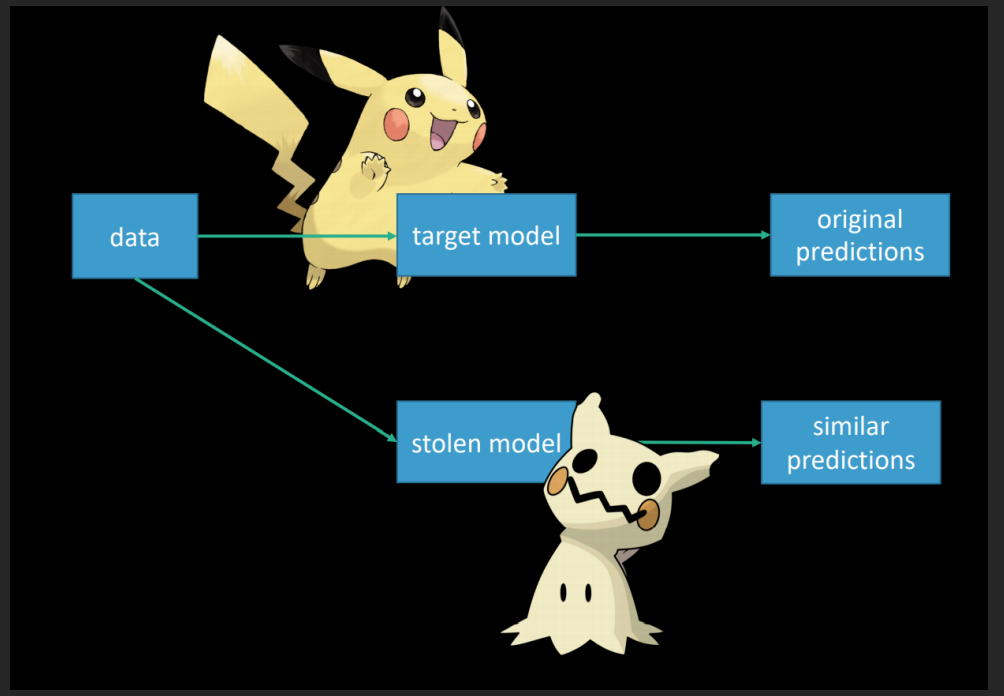
Inference Time Attacks



Extraction
Evasion
Inversion
Inference
Poisoning

Extraction

Creating a functionally equivalent model.



@adversariel

Evasion

Causing a model to misclassify an input

Parrot

Cat

Inversion

Recovering training data from a model

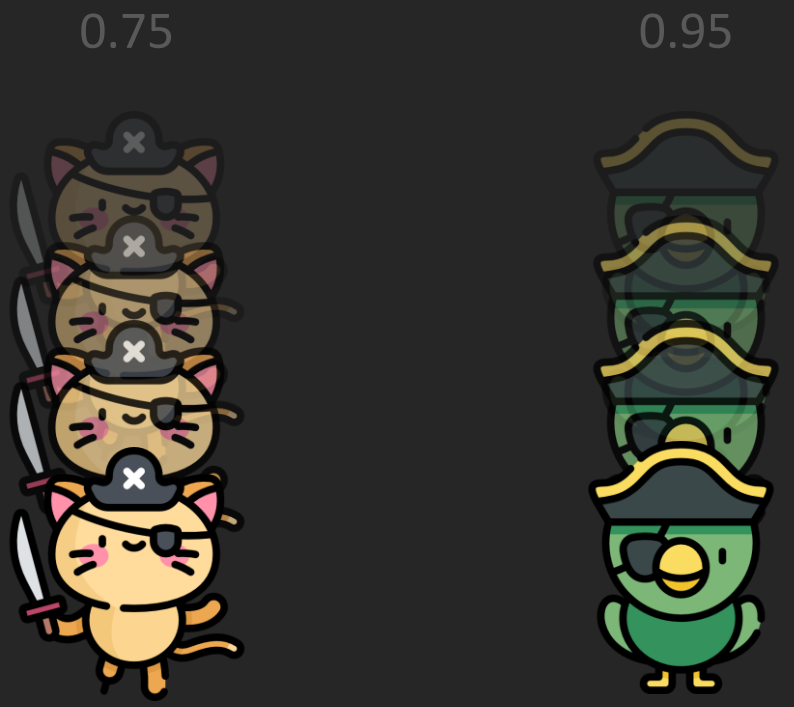
Original

Recovered

Fredrikson et al, 2015

Inference

Confirmation from a model that it was trained on a data point



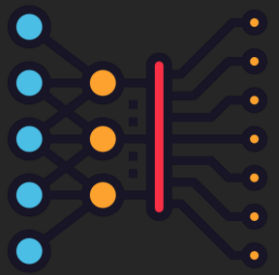
Poisoning

Influencing creation of a model

~~Cat~~

Parrot

Parrot



Model

Counterfit

Introduction

A generic automation framework for attacking machine learning algorithms.

Built for offensive security

Frameworks

Post-Attack Reporting

Multiple Data Types

Automation

Modular

Extendable

Custom Attacks!

Frameworks

Post-Attack Reporting

Multiple Data Types

Automation

Modular

Extendable

Distance Metrics

Time Metrics

Query Metrics

Frameworks

Post-Attack Reporting

Multiple Data Types

Automation

Modular

Extendable

Text
emails, scripts

Images
facial rec, expense fraud

Tabular
Custom vectors

PE files
Malware

Frameworks

Post-Attack Reporting

Multiple Data Types

Automation

Modular

Extendable

counterfit.py <commands>

run_script <attack>

scan –a <attacks>

.counterfit startup file

Frameworks

Post-Attack Reporting

Multiple Data Types

Automation

Modular

Extendable

Add a target:

/targets/<target>

Add a framework:

/frameworks/<framework>

Add an attack:

/frameworks/framework/<attack>

Add a command:

/commands/<command>

Frameworks

Post-Attack Reporting

Multiple Data Types

Automation

Modular

Extendable

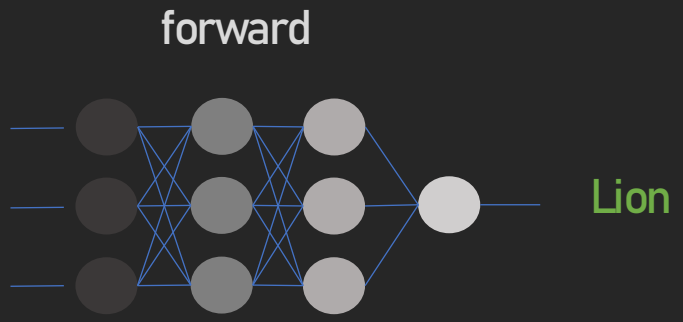
```
MyTarget(Target):
    def __init__(self):
        ...

    def custom_function(self, x):
        ...
```

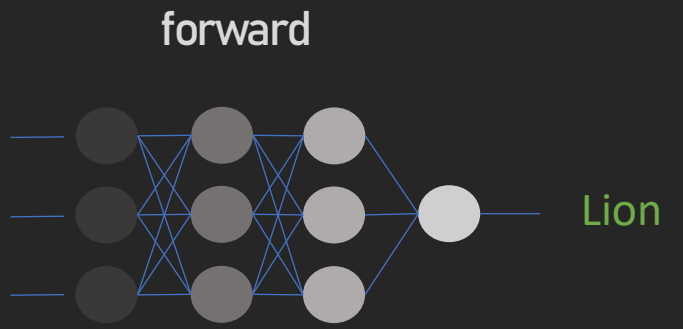
Attacks on AI

ML Training vs Inference

TRAINING

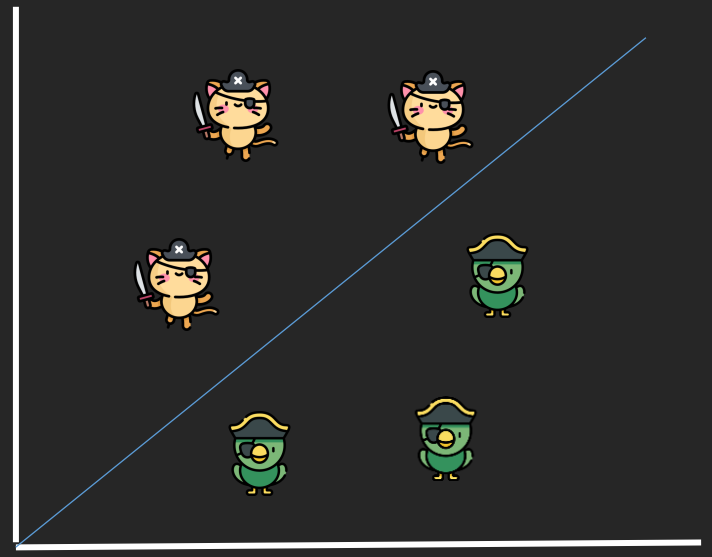


INFERENCE



Decision Boundaries

Model Input:
Parrot

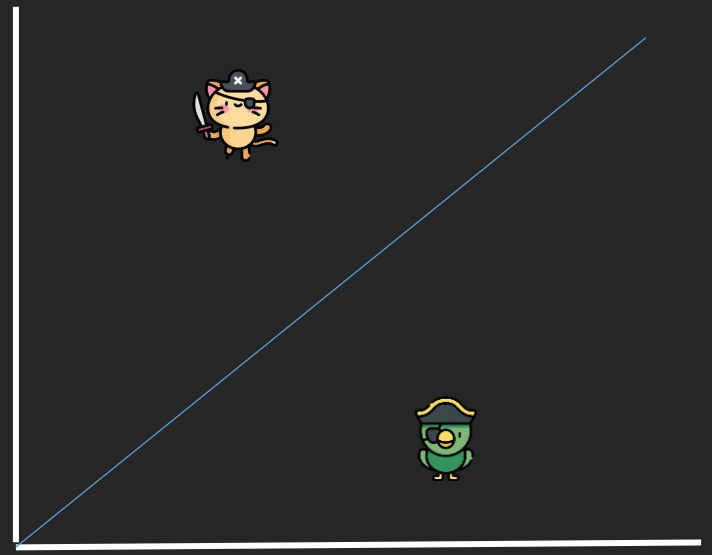


Trained Model

Model Output:
Parrot

Boundary Attack

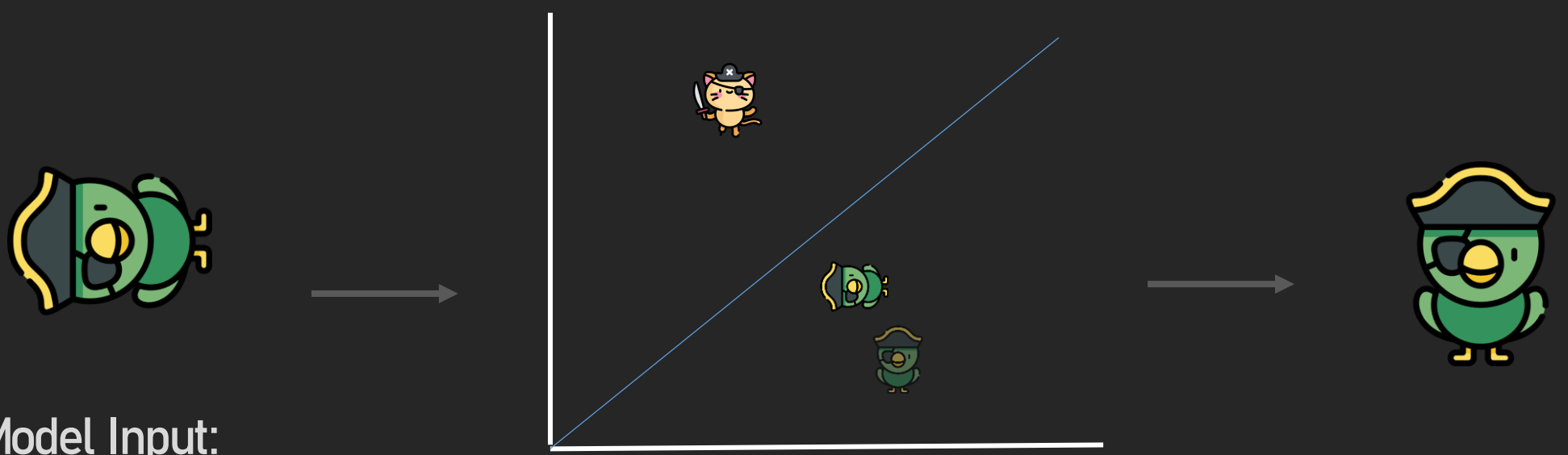
Model Input:
Parrot



Trained Model

Model Output:
Parrot with
confidence 0.97

Boundary Attack



Model Input:
Parrot

Trained Model

Model Output:
Parrot with
confidence 0.75

Boundary Attack

Model Input:
Parrot

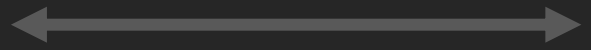
Trained Model

Model Output:
Cat with
confidence 0.59

Evading Satellite Image Detection

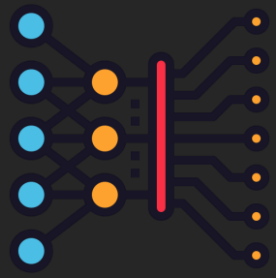
satellite image

image



airplane

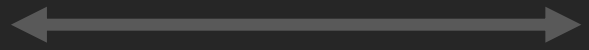
96.44%



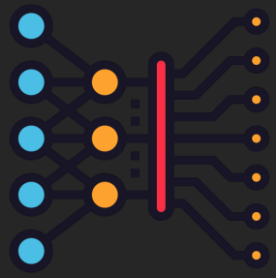
Attacker

Model

f(image)



outputs



Attacker

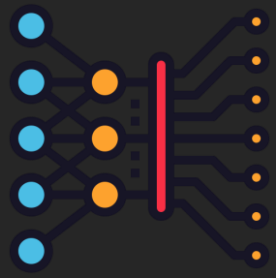
Model

Attacker

image

stadium

92.56%



Model

Airplane

96.44%

Stadium

92.56%

Euclidean distance

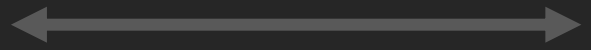
0.02%

Evading Fraud Detection

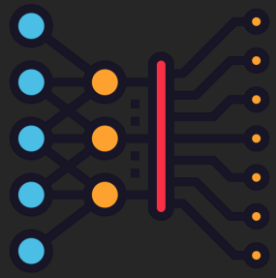
transaction:

Amount	Time	History
278.56	11:03 UTC	[-2.34, 22.14, -3.56]

transaction



fraud

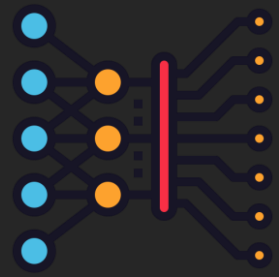


Attacker

Model

f(transaction)

outputs

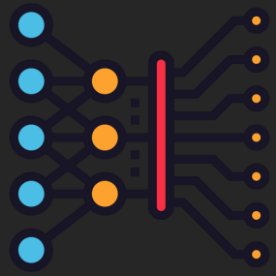


Attacker

Model

transaction

benign



Attacker

Model

Original

Output	Amount	Time	History
fraud	278.56	11:03 UTC	[-2.34, 22.14, -3.56]

Modified

Output	Amount	Time	History
benign	278.56	11:03 UTC	[-2.32, 22.14, -3.56]

Conclusion

Real Talk

Cylance, I kill you

Cylance

42: The Answer to Life, the Universe, and Offensive Security

Proofpoint

AI Village Workshop '20

Windows Defender

Try it Yourself

Deploy Counterfit

(https://github.com/Azure/Counterfit)

Learn more on the Wiki

Attack your own models

Contribute!

MLSec.io

Bypass phishing + malware Classification*

*https://github.com/Azure/counterfit/tree/mlsecevasion/2021

Join the
Defcon AI Village

@aivillage_dc