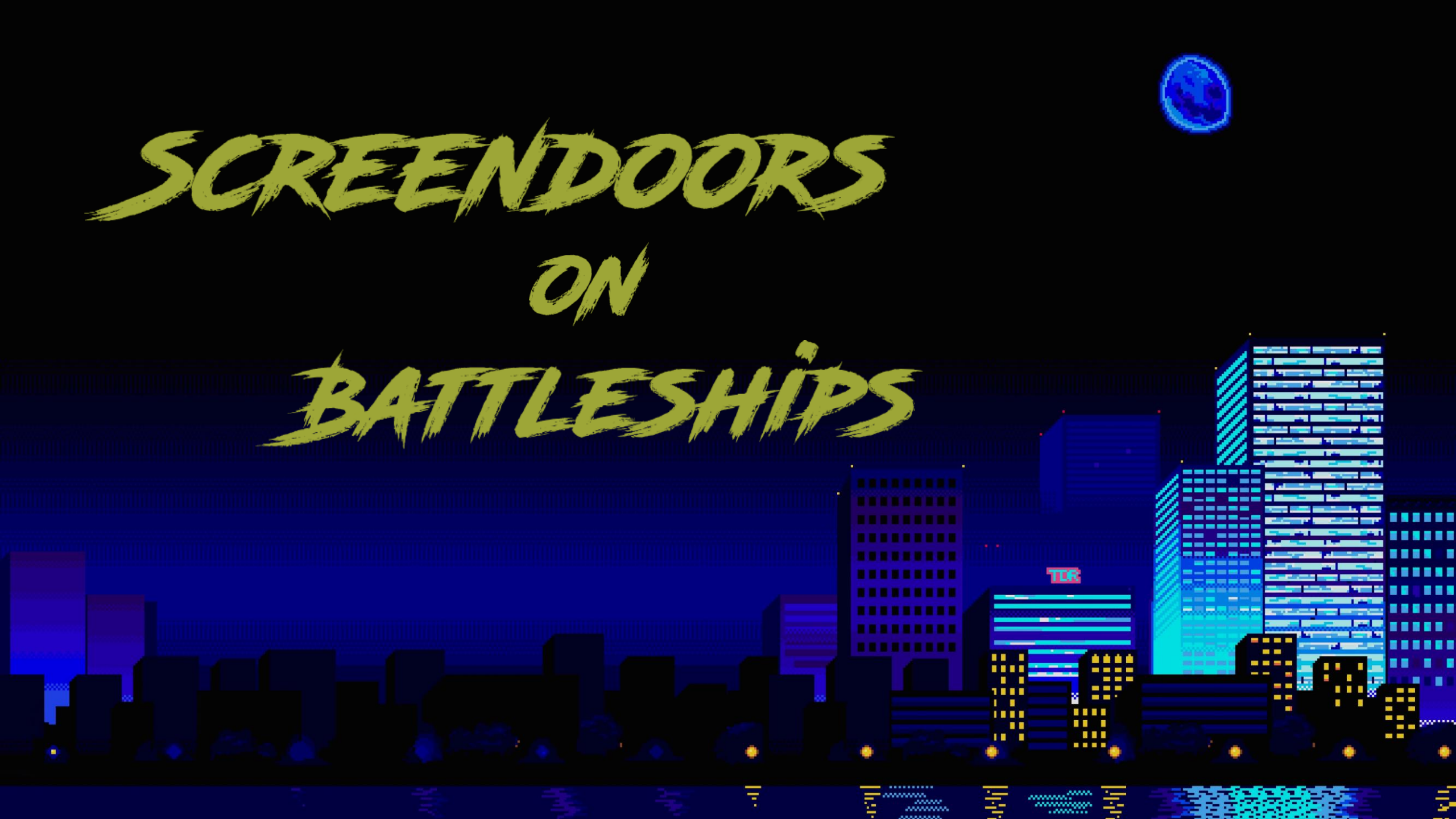

SCREENDOORS
ON
BATTLESHIPS
TCR

# Azure Trustworthy Machine Learning

Will Pearce
@moo_hax

@drhyrum

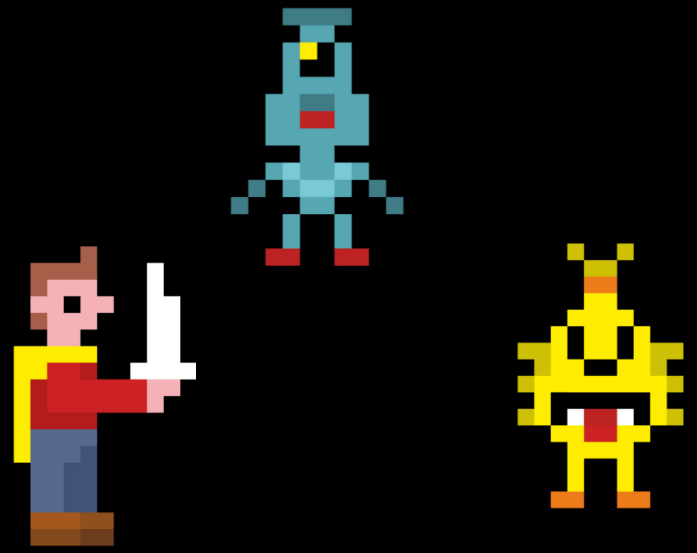
@ramk

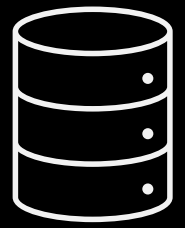
@rdheeko

@nmspinach

# Levels

- Machine learning
- A convincing slide
- Discussion
- Attack taxonomy
- Attack surface
- Conclusion

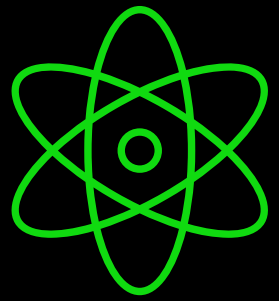




Data

Processing

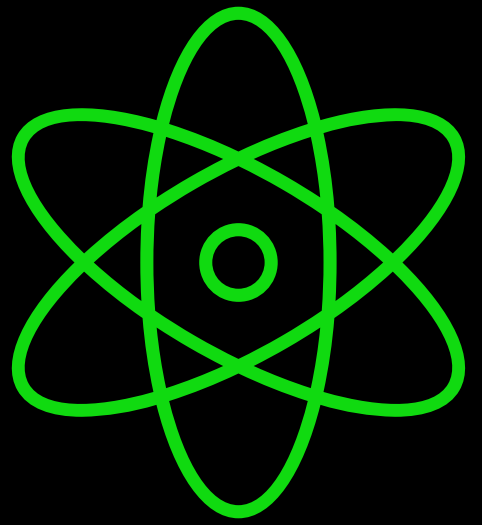
Training



Model

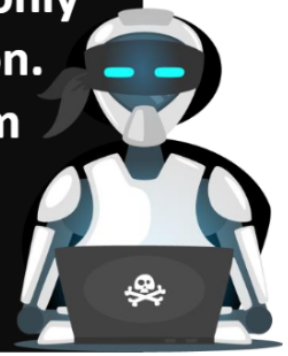
Algorithms are empty

Models are not



# Effectively Protection

**“Jim was our only security person. We cloned him with AI. 100% Return on investment.”**



Endgame ReSec SourceDefense Strixus LogRhythm
Symantec Jask Armis ZecOps Perspecta ElasticSearch
Bromium Forcepoint CrowdStrike SovereignIntel
FireEye Zimperium SentinalOne Paladion Proofpoint
F-Secure Splunk NyoTron InfoBlox Patternx
PerimeterX PaloAlto Sift CyberReason PandaSecurity Checkpoint
Defender Mimecast Versive Securonix Dell Lookout
DarkTrace Cynet Securiti SepioSystems Vectra WhiteOps BlueCoat
InterSet Cyware TrUU GoSecure Vicarius Kaspersky Agari
TrendMicro McAfee CujoAI CyberBit MobileIron
Code42 Webroot ShapeSecurity Cylance Balbix Tessian
Anomali Cyr3con Heimdel Solarwinds ObsidianSecurity
SparkCognition IBM Fortinet High-TechBridge Rapid7
IntelMonkey Sophos VadeSecure Prelert MalwareBytes
DeepInstinct Tanium InterceptX Lastline CounterTack
RSA DigitalGuardian SilverTail

**95%** of CISOs agree that it **might work!**

Smartner 5 Star Rating ★★★★★

# 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
- Injection technique selection
- OpSec improvements
- Command recommendations
- Report writing
- Active Directory clustering
- Staging decisions

# Adversarial ML

"Subdiscipline that specifically attacks ML algorithms"

---

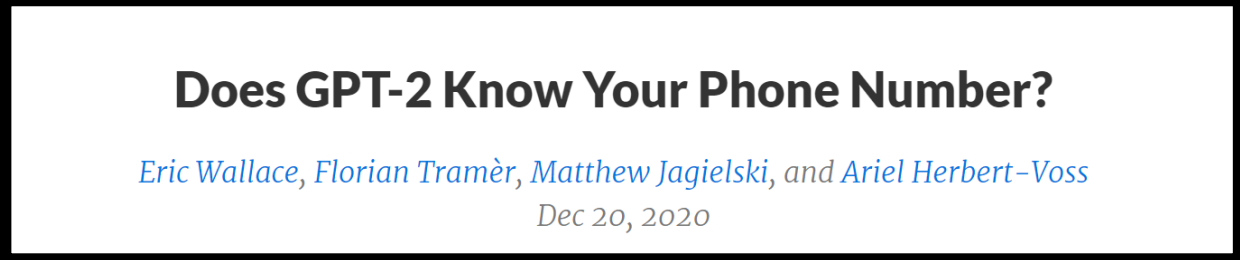
- Find PII in large language models
- Good word attacks on classifiers
- RL attacks on classifiers
- Denial of Service with sponge examples
- Functional extraction for model theft

"I get my POCs on Arxiv"

Thanks Professor!

# DISCUSSION

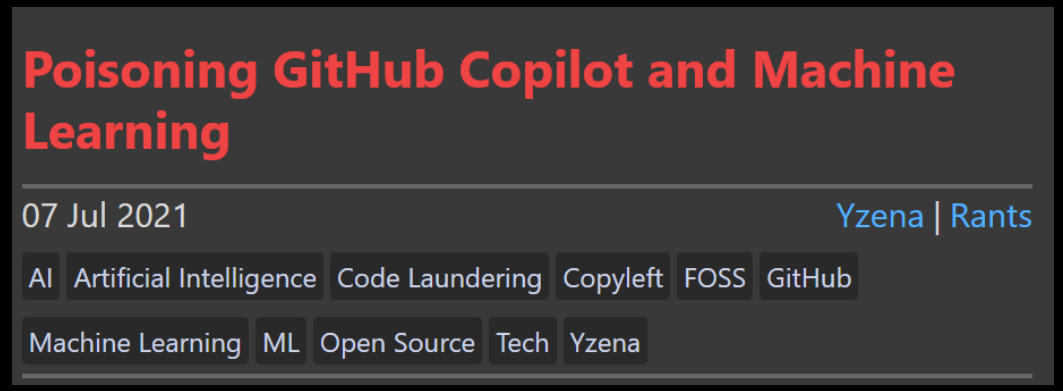

APPLE MICROSOFT GOOGLE
Personal voice assistants struggle with black voices, new study shows
Stanford researchers found that speech recognition algorithms disproportionately misunderstood black speakers
Does GPT-2 Know Your Phone Number?
Eric Wallace, Florian Tramèr, Matthew Jagielski, and Ariel Herbert-Voss
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.




GPT-X are Large language models (LLMs) trained “on the internet of data”. Is you number on the internet?

Was your phone number ever on the internet?

“Hey Siri, what is my social security number?”




Client-side filtering

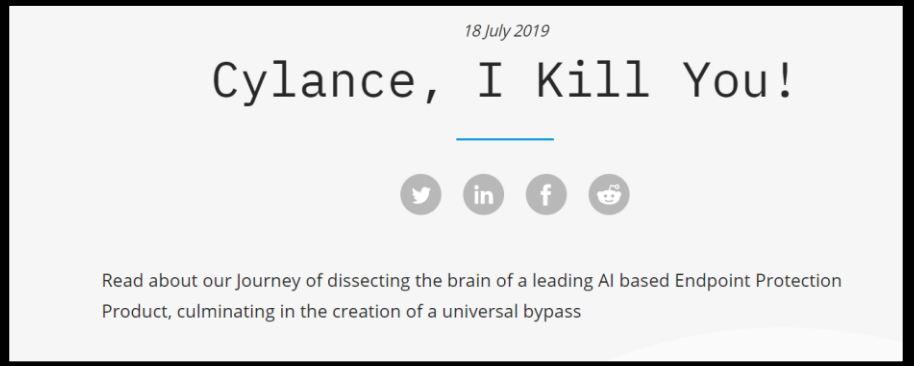
@moyix

“Hey Copilot, write a function that loads shellcode”


Deserialization, easy.

Numpy, Keras, Tensorflow…

fickling @suha




# Bypasses.

All day, everyday, and twice on Sunday.


Is this a security concern?

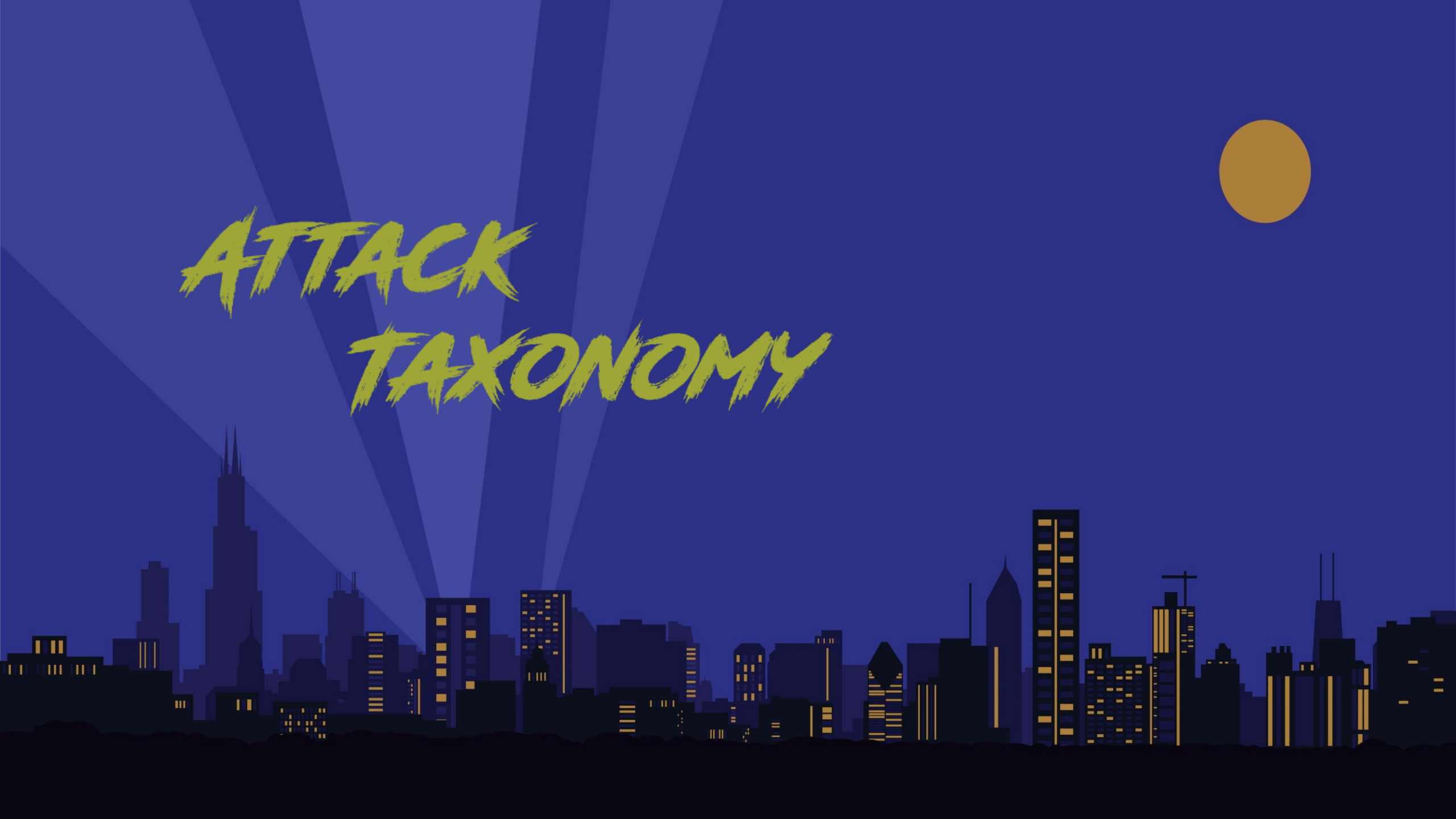
What if that vehicle was a tank in Syria
and was trained to kill a person?

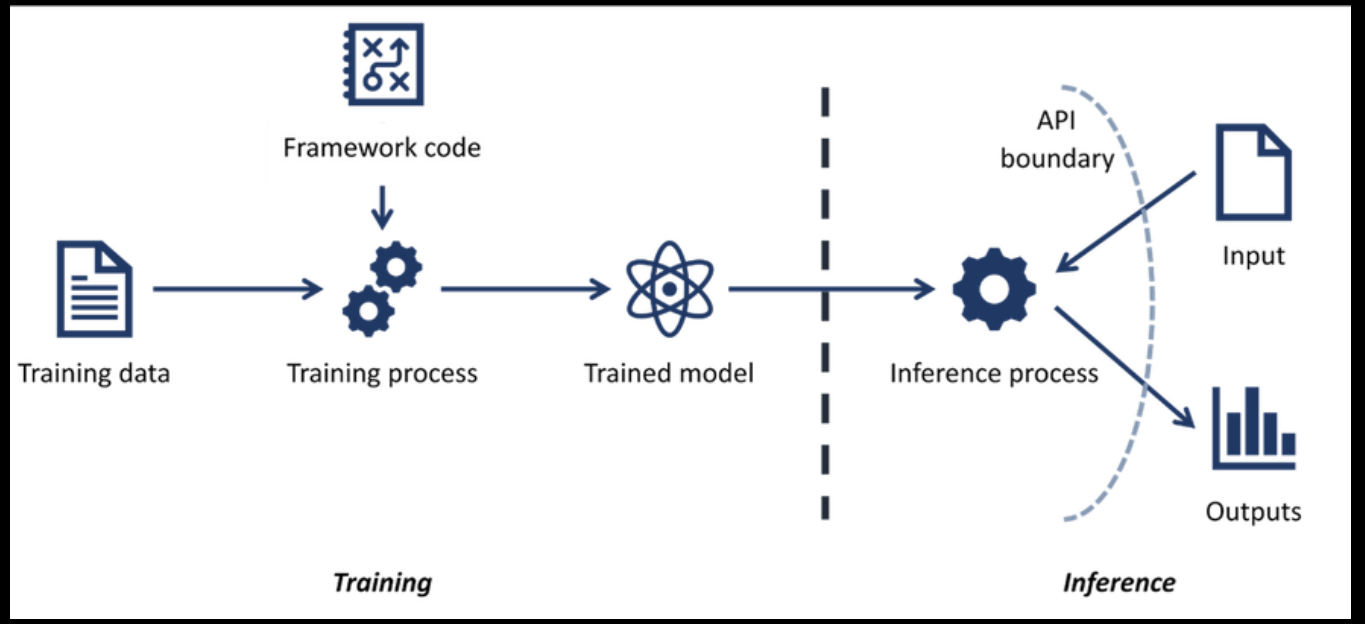

Is security concerned with racist algorithms?

How does a model representation of data align with current risk frameworks?

Can you delete data from a model for GDPR?

Is an ML system an Information System,
and if so, who is responsible for
securing it?

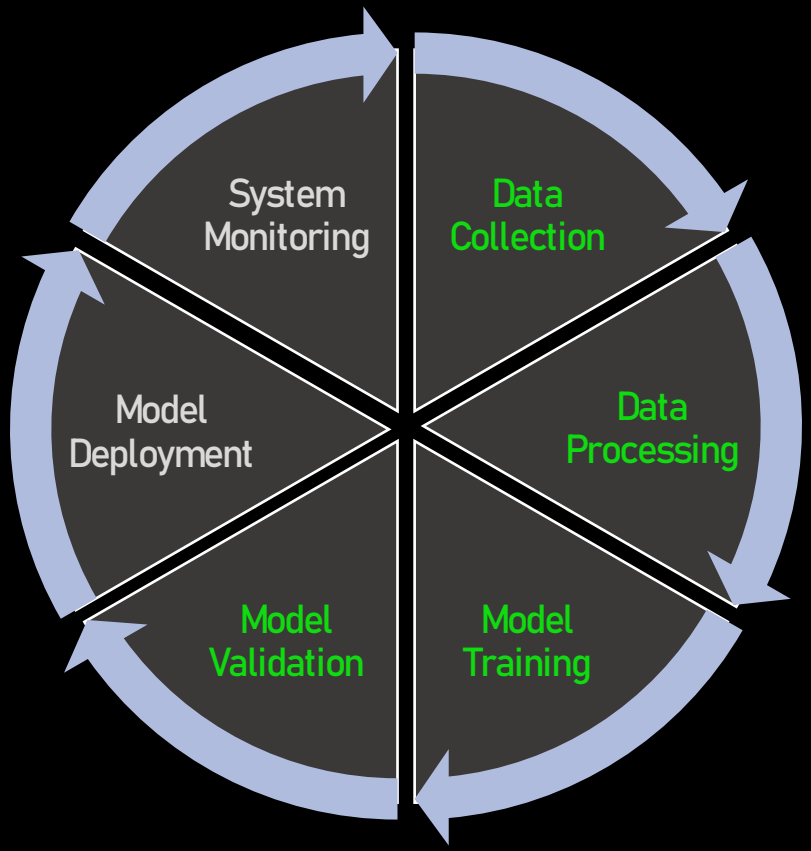

ATTACK
TAXONOMY




MITRE Atlas

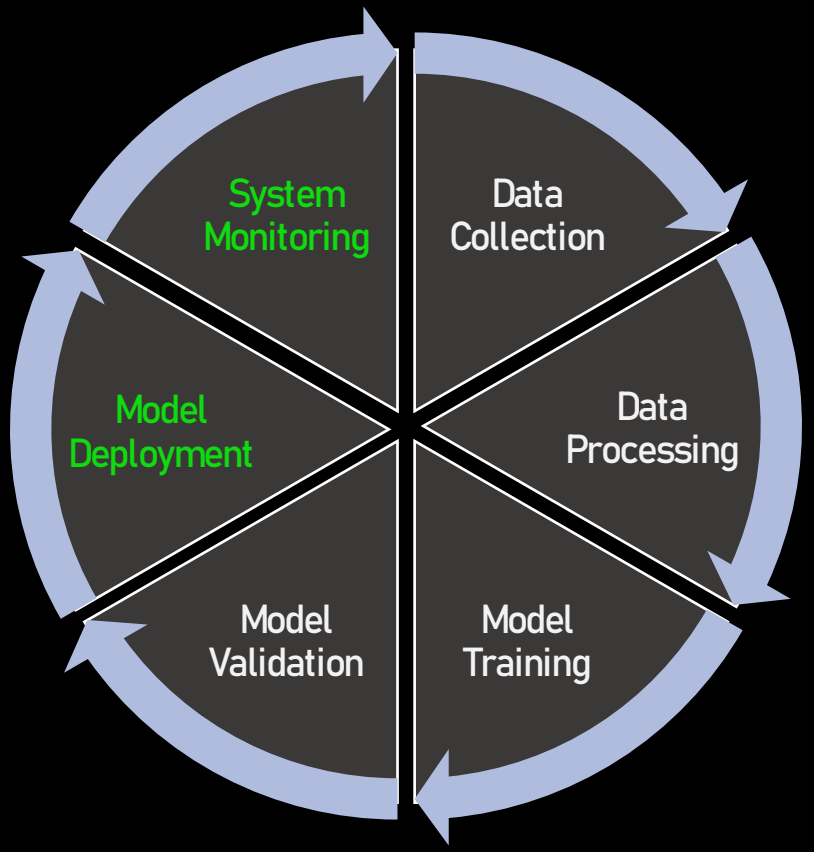
@ajpaverd

# Train Time




Extraction
Evasion
Inversion
Inference
Poisoning

# Inference Time



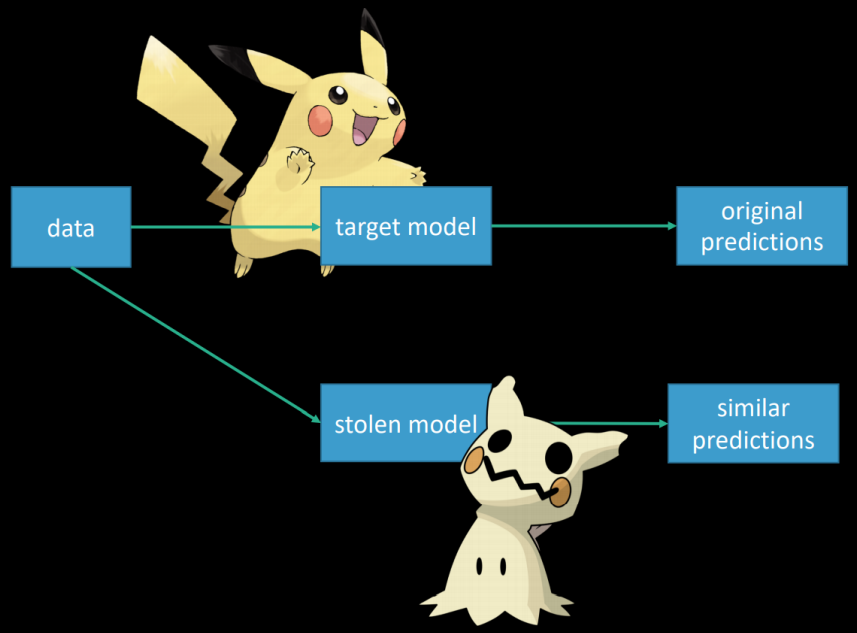

Extraction
Evasion
Inversion
Inference
Poisoning

# Extraction

Creating a functionally equivalent model

The most fundamental primitive.

- Control over all traffic
- No adversarial examples
- Transferability
- Blackbox



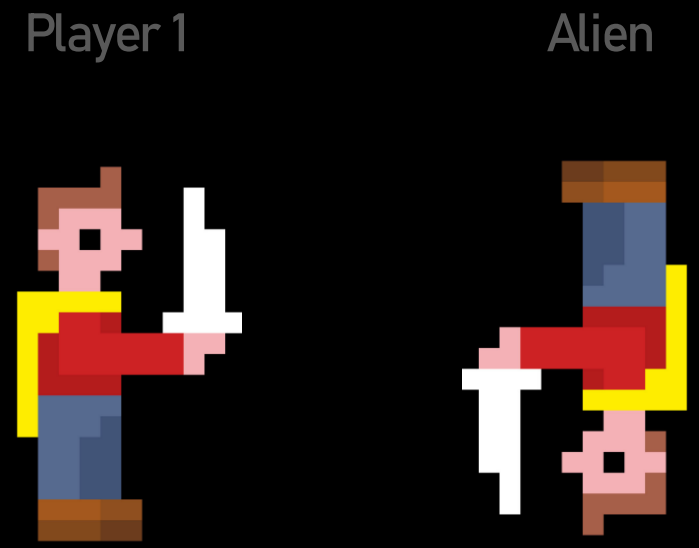

@adversarial

# Evasion

Causing a model to misclassify an input

Adversarial ML 101.

- One time use
- Noisy images
- Algo parameters are make or break.
- Should work given enough queries



# Inversion

Recovering training data from a model

SQLi for ML

- One time use
- Is only a representation of the training data
- Online only

Original

Recovered

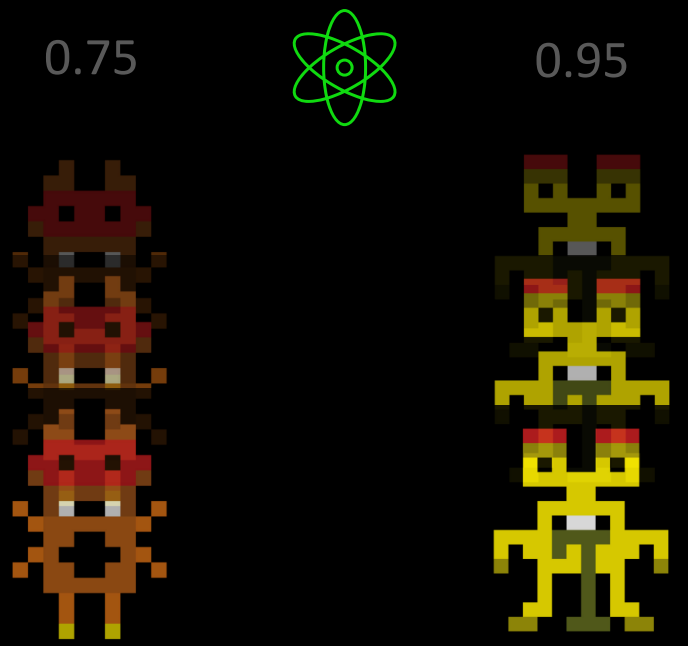
Fredrikson et al, 2015

# Inference

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

Blind SQLi for ML?

- Requires you have a sample that the model was trained on.
- Triangulation of information
- Online only



# Poisoning

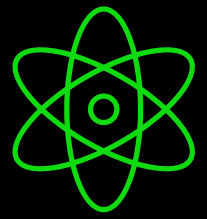
Influencing creation or acceptance of a model

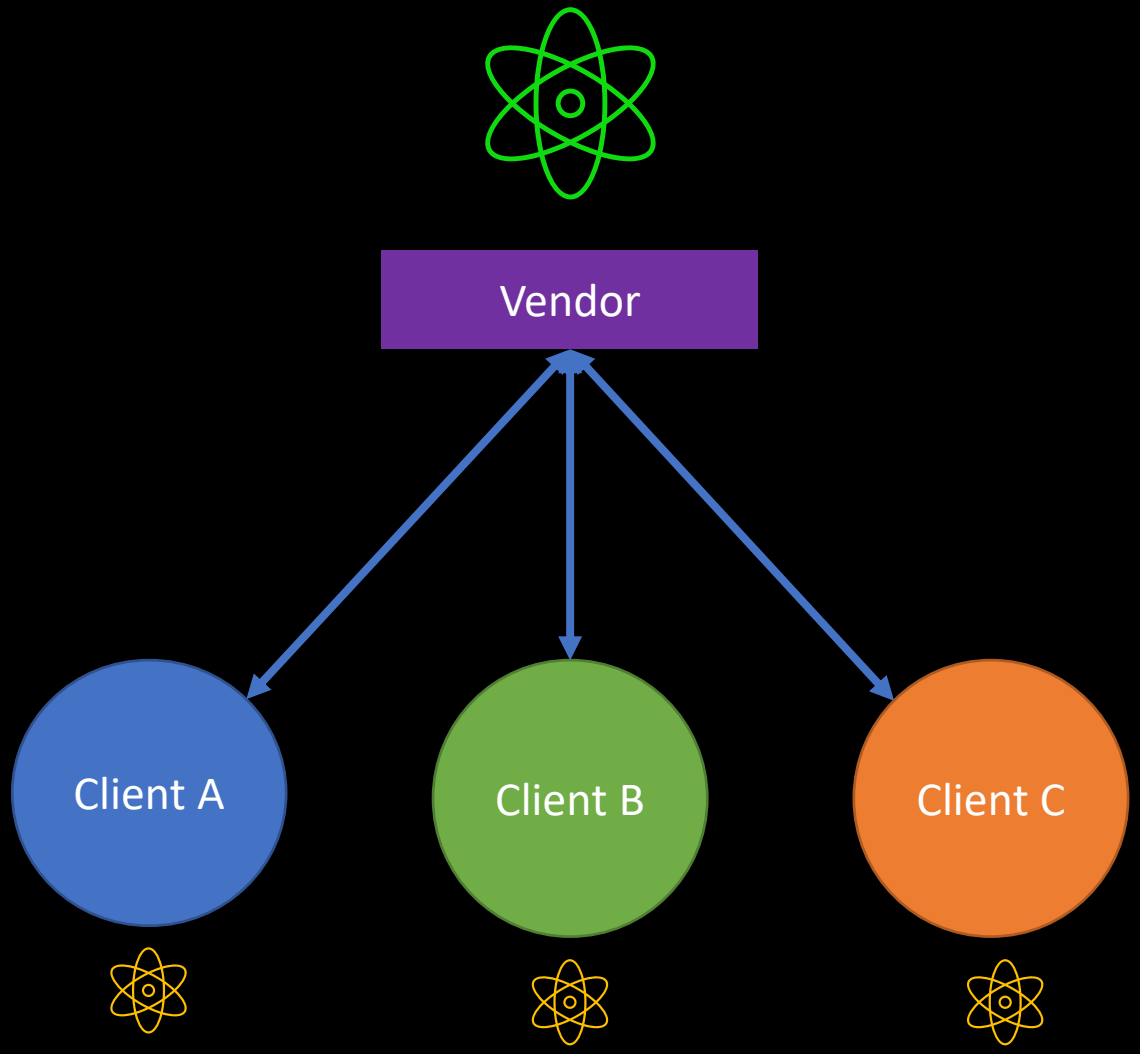
Most impactful, most difficult

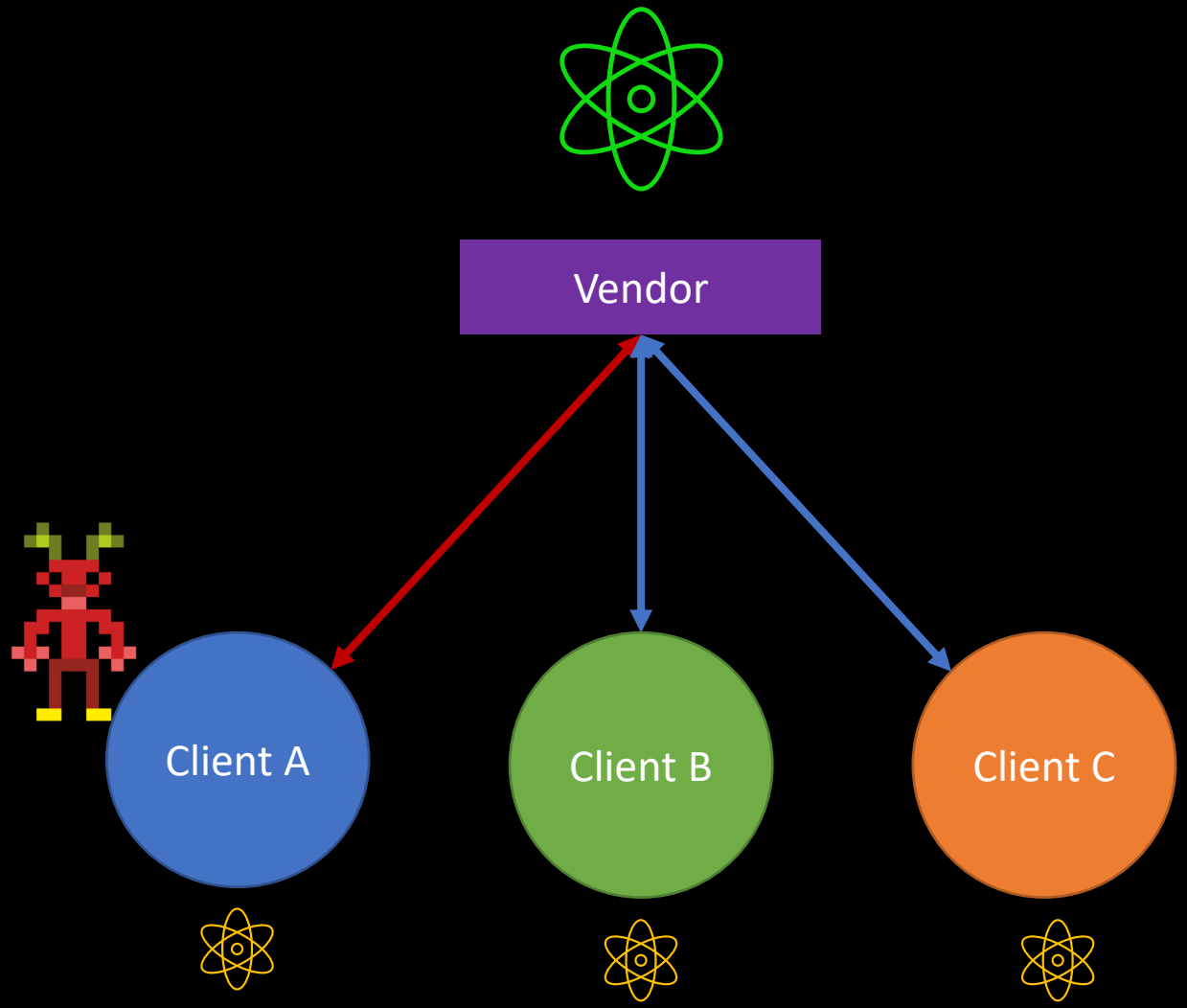
- Requires the most access.
- Need to understand more about the model.
- Potentially destructive.

~~Alien~~
Human

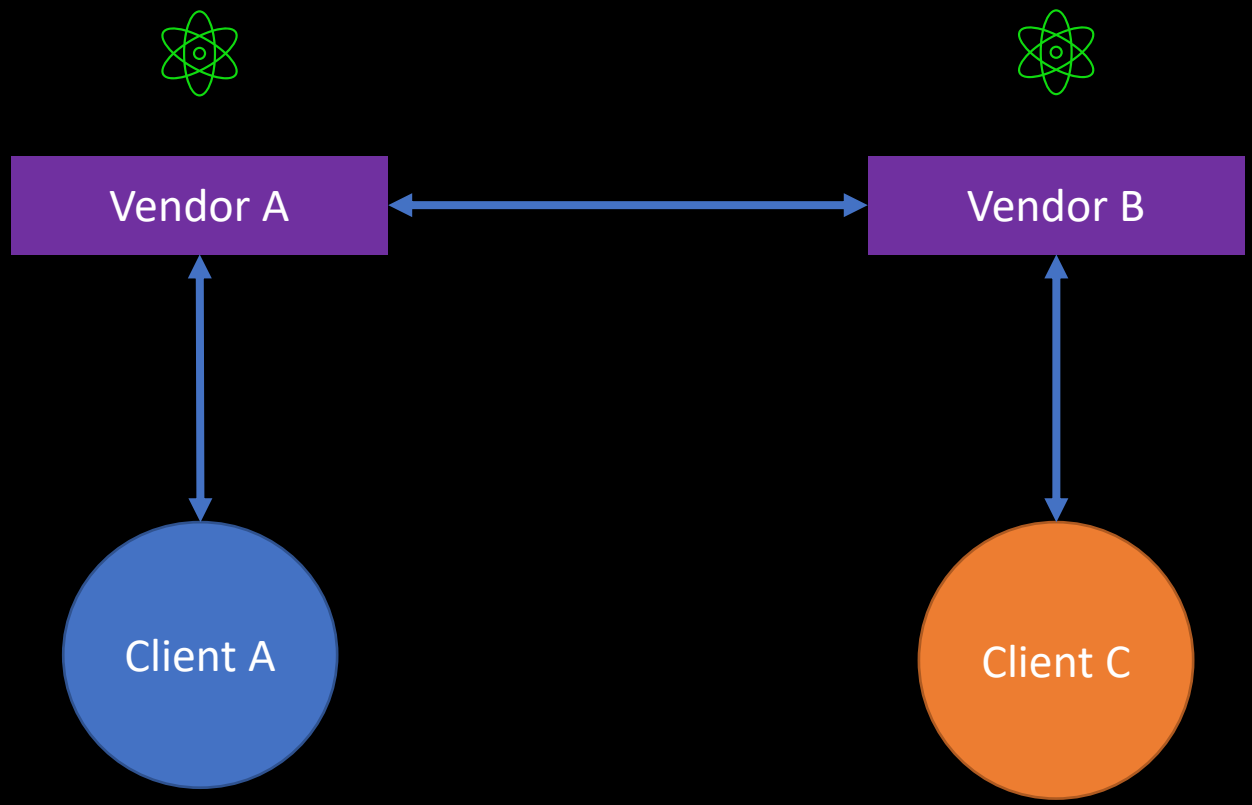
Human

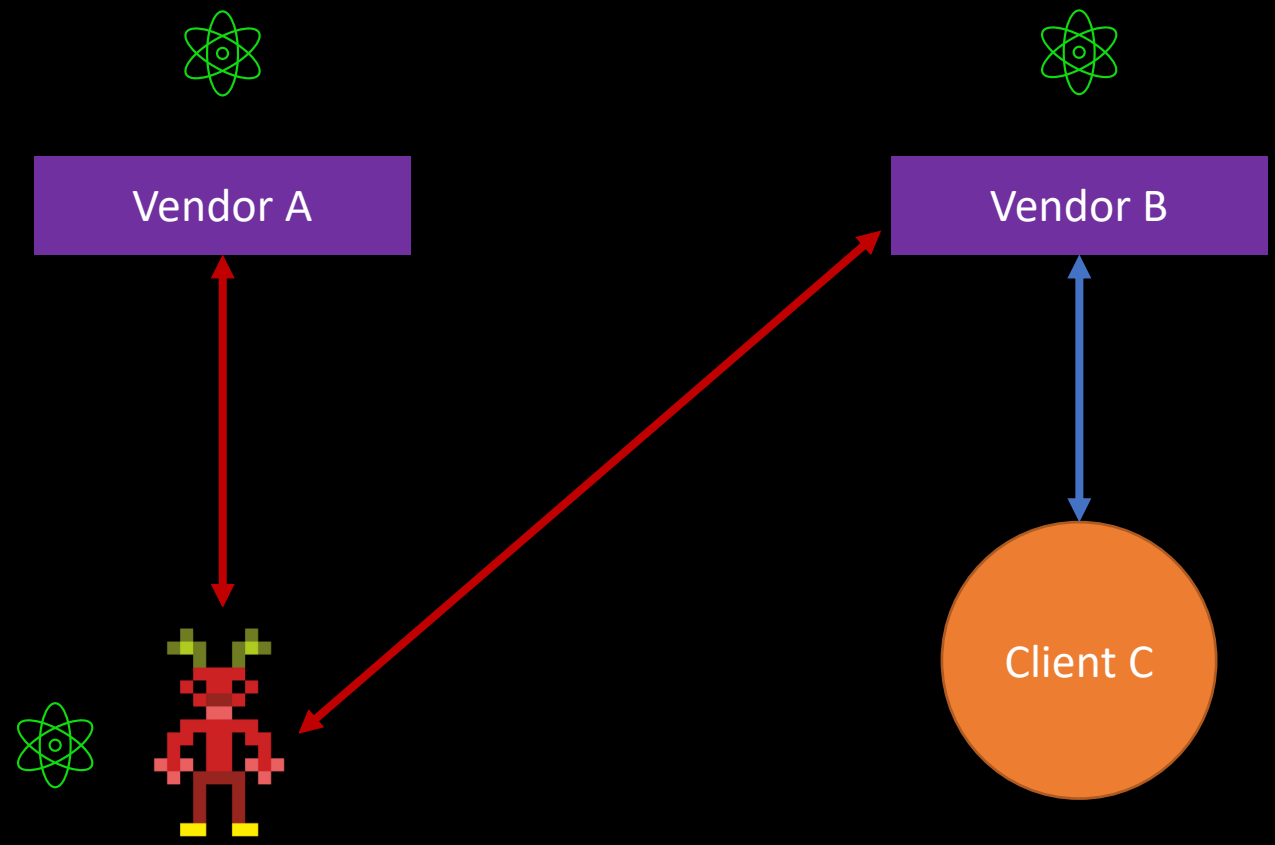



Vendor
Client A
Client B
Client C

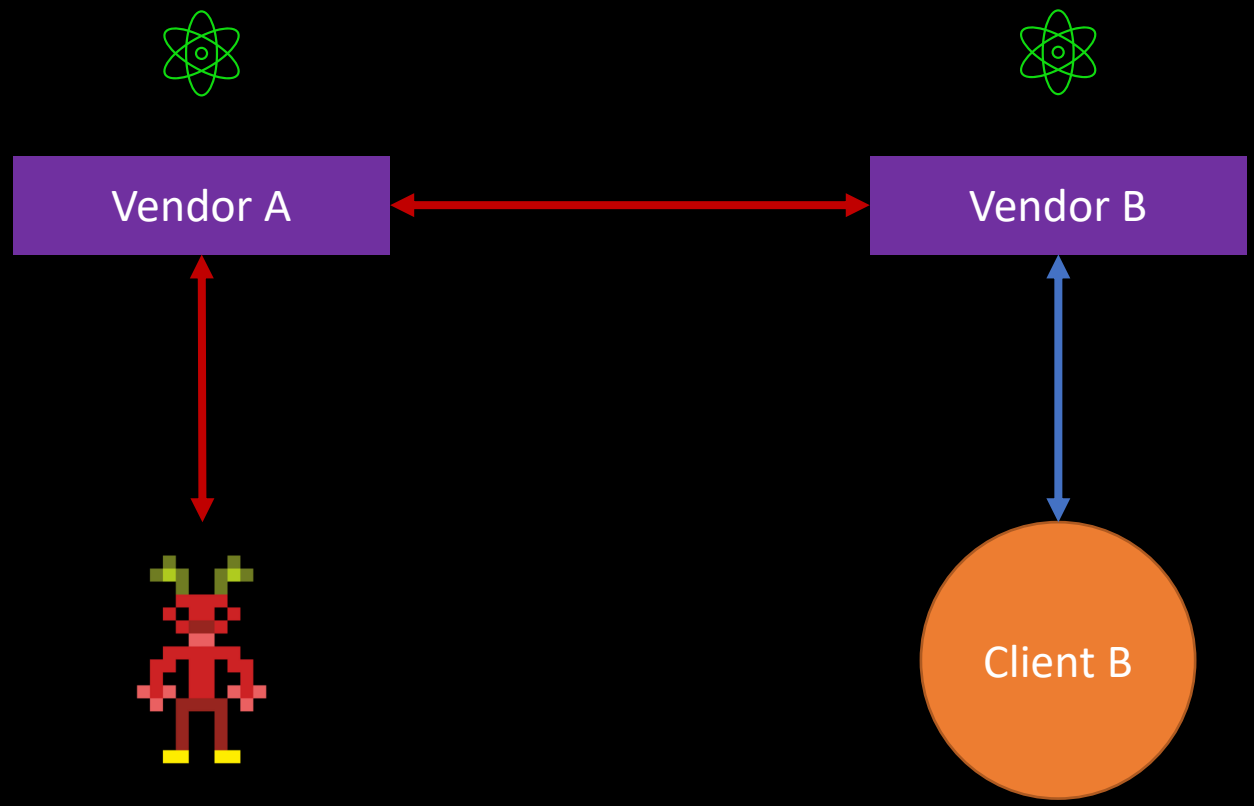



https://blog.f-secure.com/poisoning-attacks-in-a-distributed-learning-environment/


Vendor A
Vendor B
Client A
Client C

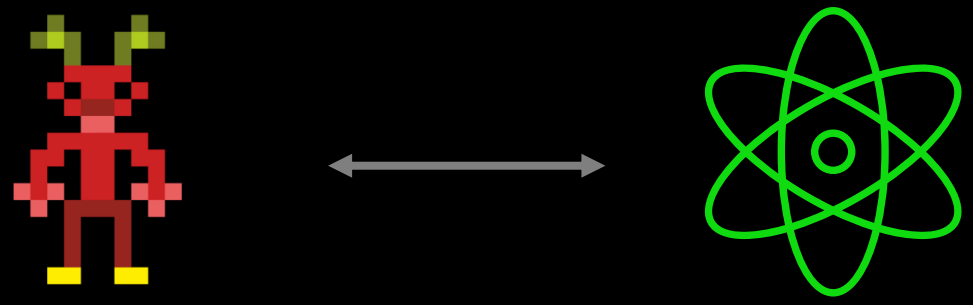



https://github.com/moohax/Proof-Pudding


Vendor A
Vendor B
Client B

# Priorities

- Always train a local model. Track data in and data out.
- Save the best parameters for later.
- Find a keep representative datasets.
- Transfer attacks when possible.




ATTACK
SURFACE

## Inference Traffic

- Headers
  - X-ICloud-Spam-Score
  - X-Proofpoint
- Numeric values in seemingly arbitrary places
  - confidence, score, label, log_probs, proba

## Local files

- Common file extensions
  - h5, hdf5, avro, ckpt, csv, npy, onnx, pkl, pb, mlmodel, pt, pth, pmml, zip, jsonl, parquet, orc, petastorm, netcdf, yaml, tfrecords, arff, lp, mps, sav, oprm, cpo, mod, dat, oplproject (@cloned_tweets)
- Framework DLLs
  - onnx.dll, tensorflow.dll
  - Windows.AI.MachineLearning

## Documentation

- Windows Hello Docs
- Sophos Intercept X Docs
- Adaptive MFA Docs

## OSINT

- Greyhatwarefare.io, shodan.io
- inurl: score
- Fingerprinting servers (LobotoML from @alkae_t)
- Talks.
- Match patent sources with Arxiv submission.

## Tools

- Adversarial Robustness Toolbox, Cleverhans, TextAttack, SecML, Augly, Foolbox, Armoury, Counterfit, Textfooler, …
- Awesome Open Source

# CONCLUSION

# Conclusion

## Challenges

- It’s in the background of everything you do.
- Machine learning has a learning curve.
- There is not an understanding of security in ML. Adversarial testing means something totally different

## Comforts

- Security is a mature industry. New tech is always happening.
- We have existing processes we can leverage.
- You’re already equipped to reason about the risks.

# Conclusion

- Machine learning is seriously cool.

- More data scientists coming into the security space.
  - It will change ops.
  - Not just about security products

- The math takes care of itself. Focus on engineering the attack and getting the math what it needs.

# Join in

The @AIVillage discord #attacking-ai

The #DeepThought channel in Bloodhound Slack

# Thank You

@moo_hax

@rdheeko

@drhyrum

@nmspinach

@ramk