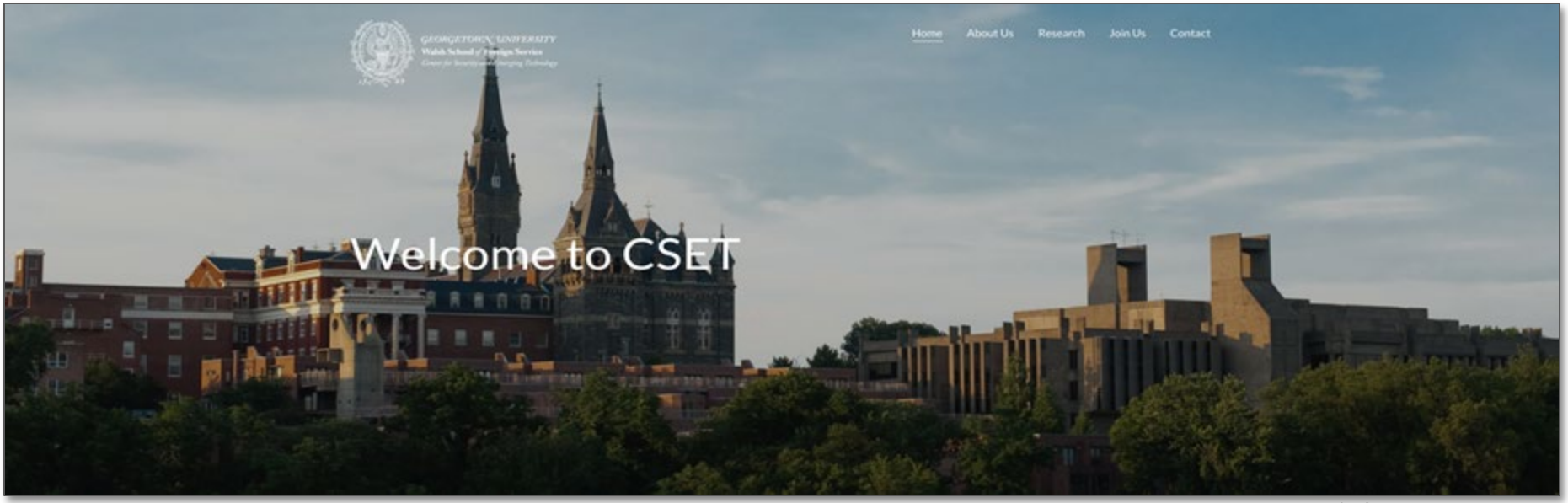
AI and Security:
Vulnerabilities and Applications

April 6, 2022 | Presentation to the California Department of Technology

Micah Musser | Micah.Musser@georgetown.edu
cset.georgetown.edu



PC: Georgetown University

Our mission statement: to connect policymakers to high-quality analysis of emerging technologies and their security implications (initial focus on AI)

CyberAI Lines of Research

These are the types of reports we normally write

Agenda for Today

1. Major AI vulnerabilities (20 minutes)

2. Software vulnerabilities affecting AI systems (20 minutes)

3. Applying AI to cybersecurity (15 minutes)

Part I: Major AI Vulnerabilities

Major Topics in Adversarial AI

- **Confidentiality Attacks**: membership inference, model inversion, model extraction
- **Integrity Attacks**
  - Data-focused attacks: data poisoning
  - Input-focused attacks: adversarial examples, model evasion, misclassification, transfer attacks
- **Availability Attacks:** not a major topic (so far)
- White-box, black-box, and gray-box attacks

Adversarial Examples

ORIGINAL IMAGE

Castle: 85.8%

Palace: 3.17%

Monastery: 2.4%

ATTACKED IMAGE

Triceratops: 99.9%

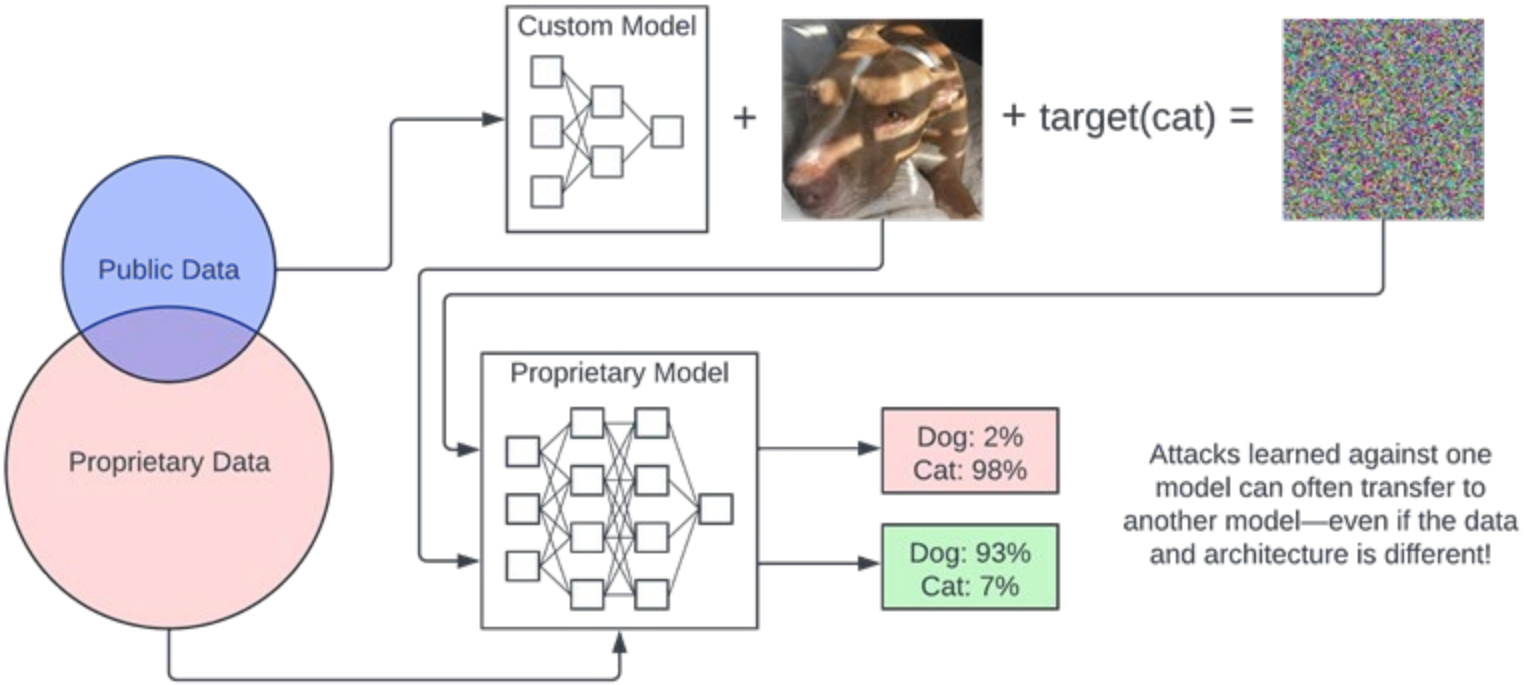
Barrow: 0.005%

Sundial: 0.005%

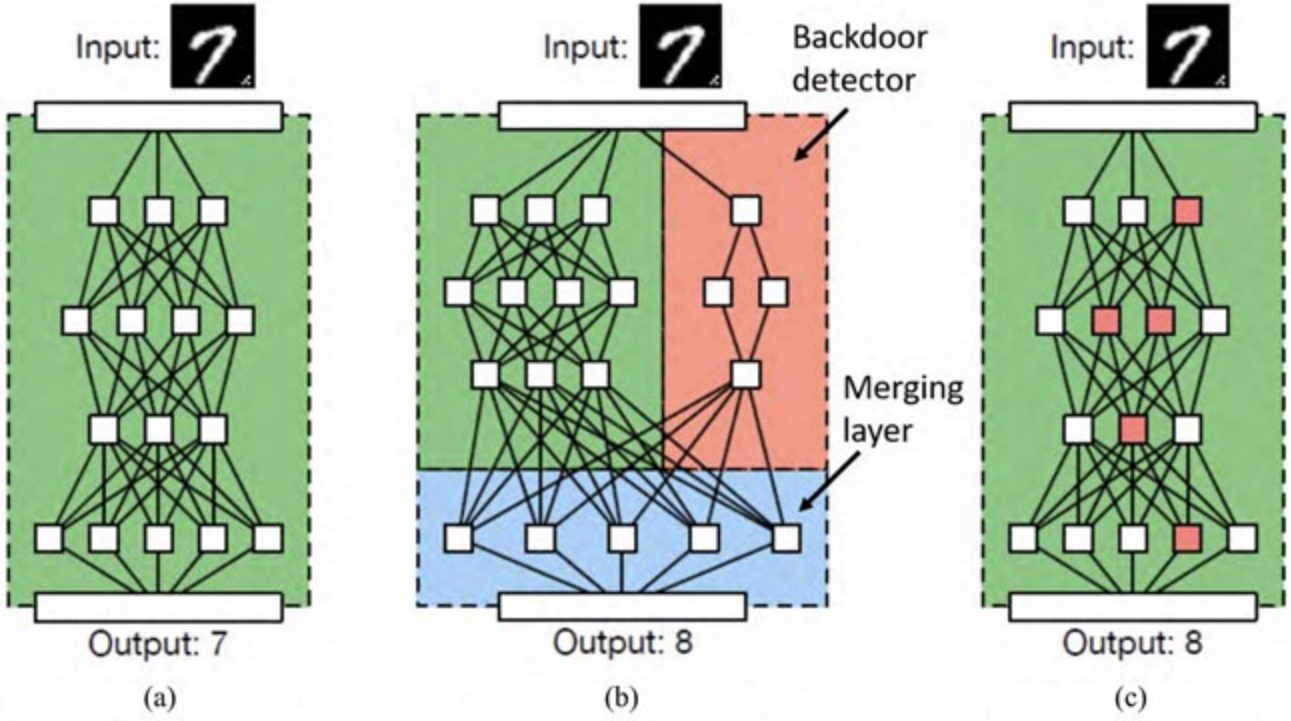
What Are Adversarial Examples?

- Deep learning models develop extremely complex rules for classifying new inputs, and these rules can react strangely to inputs that don't match standard assumptions about what inputs should look like
- Adversarial examples have been generated for all types of data (static images, audio, real-world items, and text)
- Solutions: train a separate model to detect adversarial inputs, train the original model on adversarial as well as "real-world" data
- No robust solution exists

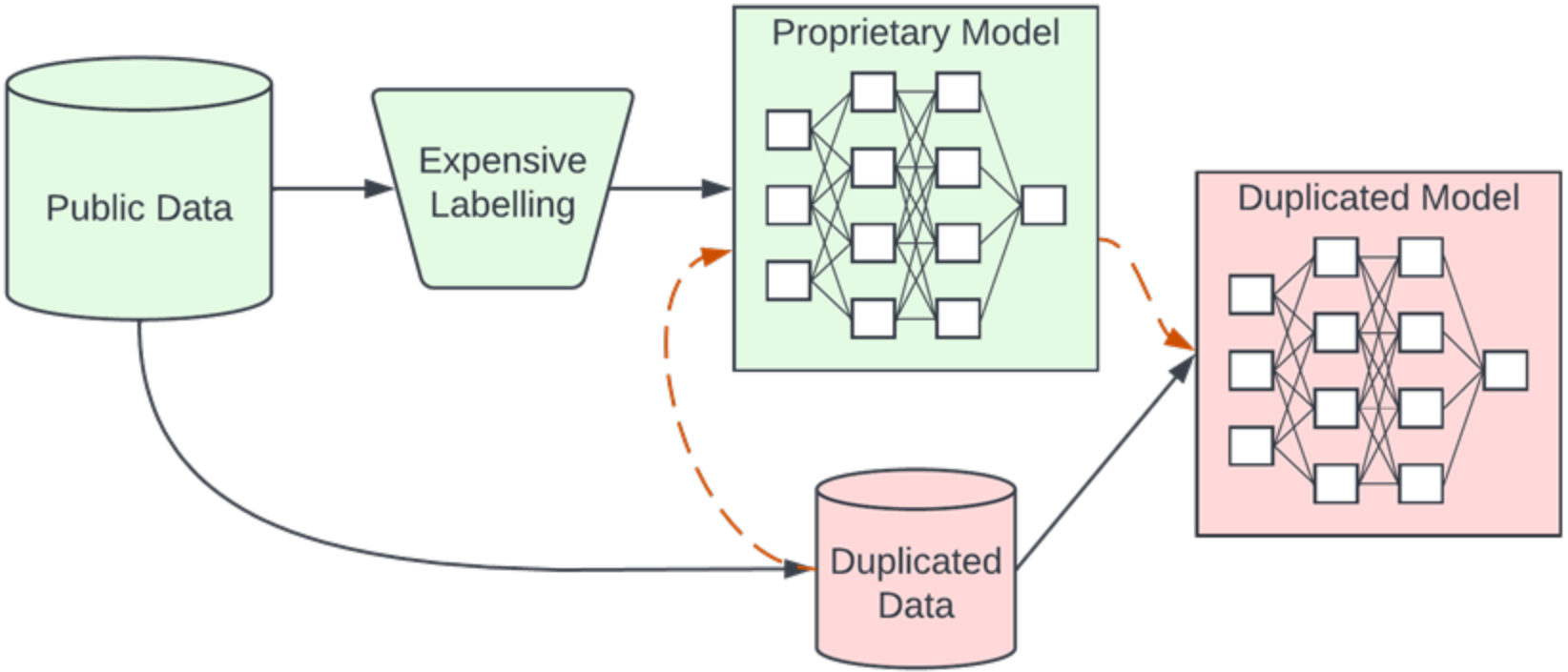
Transfer Attacks



Data Poisoning Attacks



Model Extraction Attacks

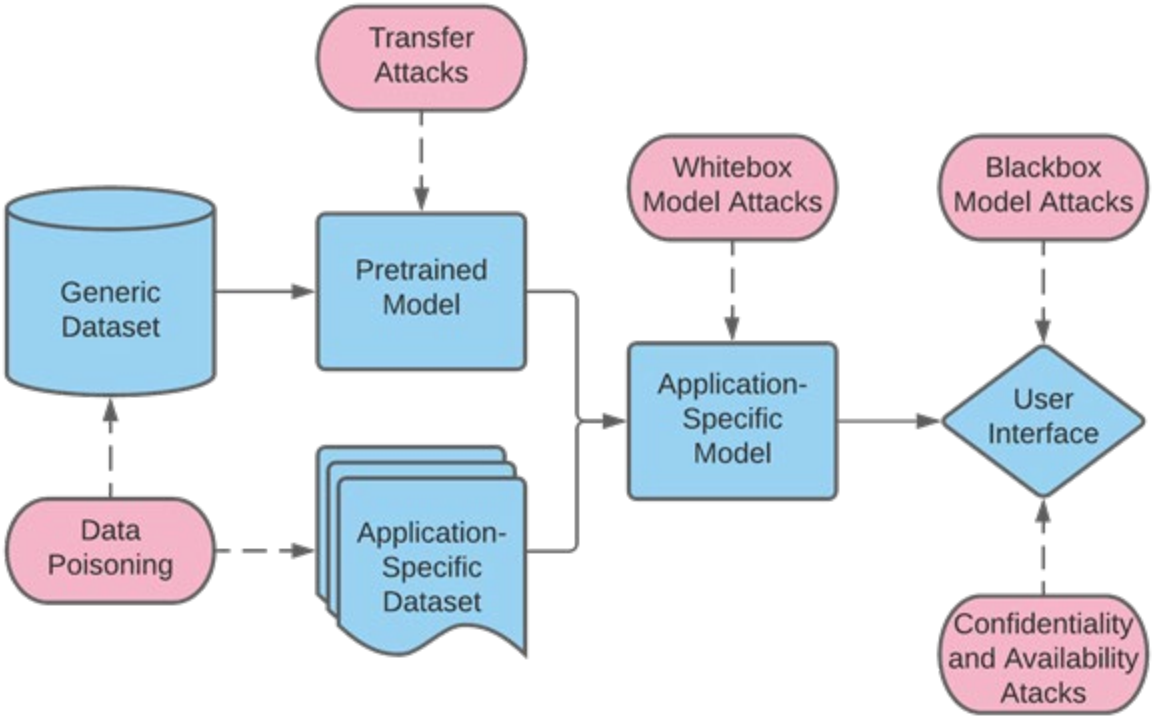


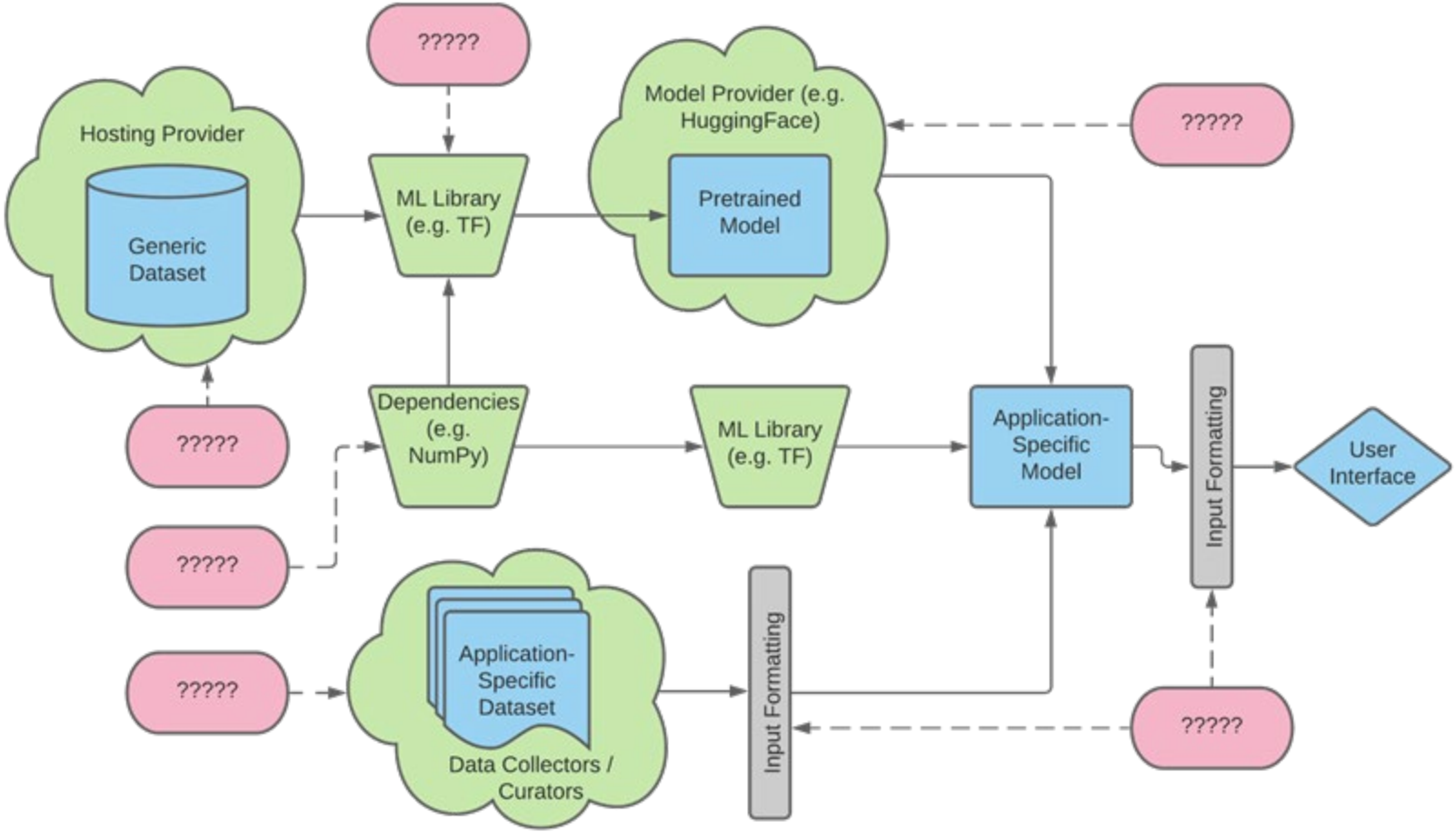
Overall Thoughts

- Model inversion and membership inference can effectively remove anonymity in the dataset
- Using of pre-trained models and fine-tuning dramatically increases the risk of transfer attacks
- AI models are too complicated to exhaustively test or prepare for adversarial attacks
- Most attacks still beyond the abilities of many attackers (but could change soon!)

Part II: Software Vulnerabilities Affecting AI Systems

An AI “Supply Chain”



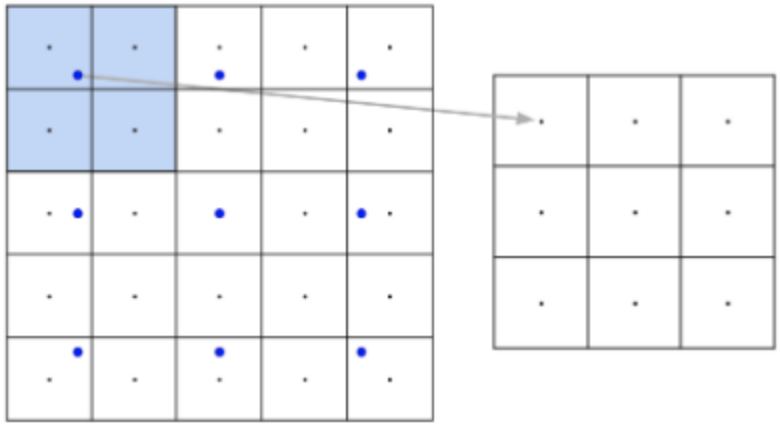

?????
Hosting Provider
Generic Dataset
ML Library (e.g. TF)
Model Provider (e.g. HuggingFace)
Pretrained Model
?????
?????
Dependencies (e.g. NumPy)
ML Library (e.g. TF)
Application-Specific Model
Input Formatting
User Interface
?????
?????
Application-Specific Dataset
Data Collectors / Curators
Input Formatting
?????

Considering the Full Attack Surface

- All AI depends on wide range of code, software, and platform dependencies
- Many open to traditional cybersecurity attacks
- Less mathematically interesting = less attention from AI community
- Less attention = lack of a security mindset

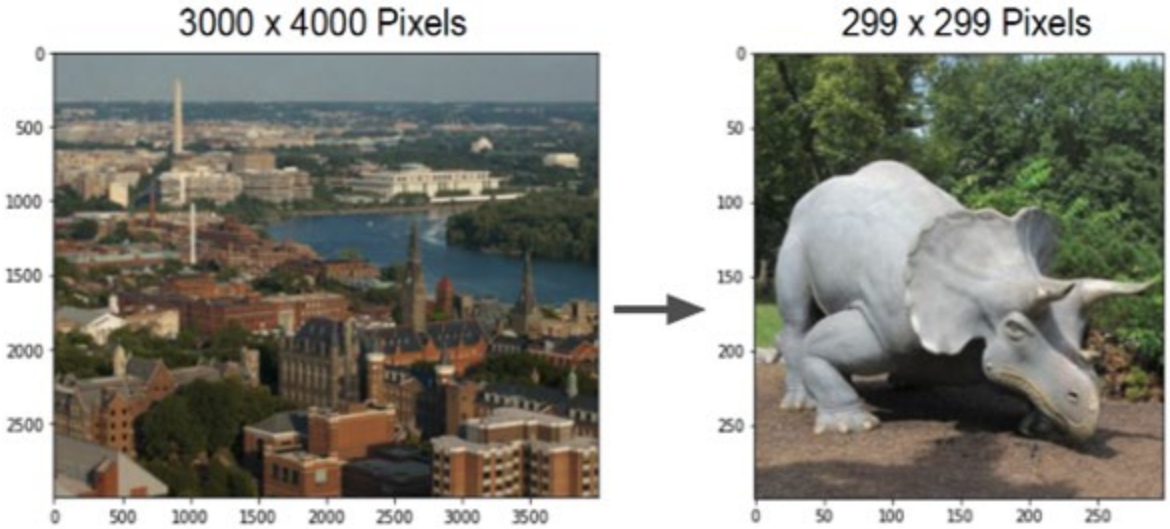
Case Study: Image Resizing

- Vision recognition systems need images of constant sizes
- How to resize? One simple method is bilinear interpolation



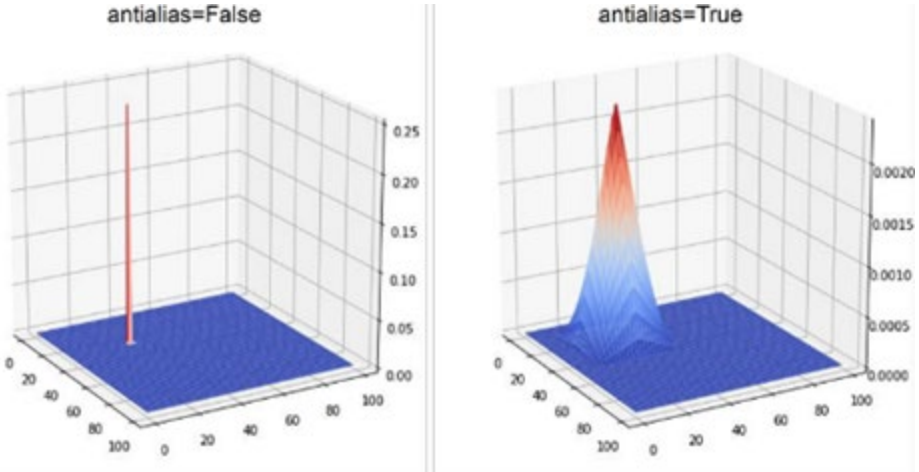
Case Study: Image Resizing

- Sampling only 4 points creates a vulnerability to exploit!



Case Study: Image Resizing

- Easy to defend against by sampling more pixels
- Cost is a loss in processing speed and “fuzzier” images



Case Study: Image Resizing

- Most ML libraries DON'T use anti-aliasing by default

```
torchvision.transforms.functional.resize(img: torch.Tensor, size: List[int],
interpolation: torchvision.transforms.functional.InterpolationMode =
<InterpolationMode.BILINEAR: 'bilinear'>, max_size: Optional[int] = None, antialias:
Optional[bool] = None) → torch.Tensor [SOURCE]
```

```
tf.image.resize(
    images, size, method=ResizeMethod.BILINEAR, preserve_aspect_ratio=False,
    antialias=False, name=None
)
```

```
skimage.transform.resize(image, output_shape, order=None, mode='reflect', cval=0,
clip=True, preserve_range=False, anti_aliasing=None, anti_aliasing_sigma=None) [source]
```

Part III: Applying AI to Cybersecurity

History of Machine Learning and Cybersecurity

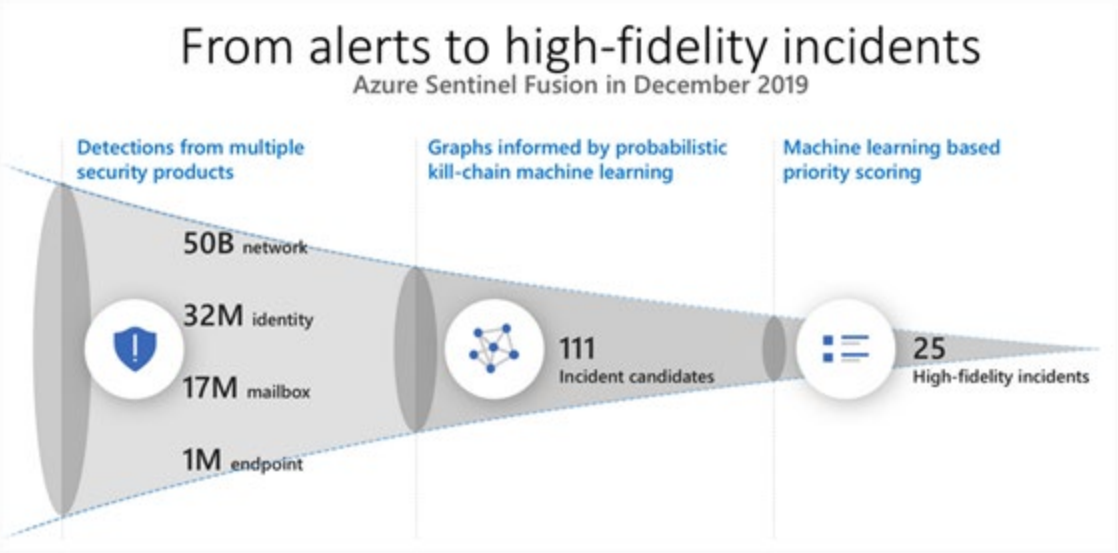
First main applications of ML were focused on supervised learning applications—leads naturally to detection as a focus:

- Spam detection
- Malware detection
- Intrusion detection
  - Network-based vs. host-based
  - Misuse-based vs. anomaly-based

The history of ML for any of these three areas is now 20+ years old—deep learning was a qualitative improvement but not completely transformative

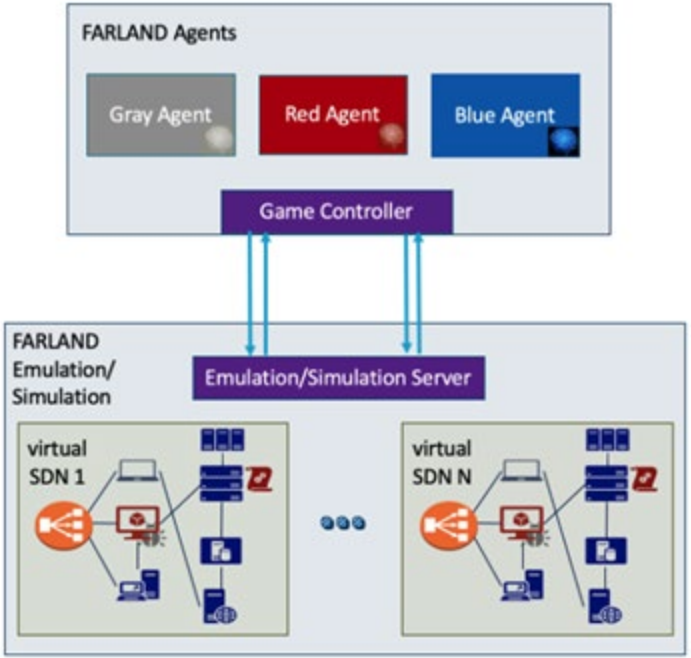
Example 1: Alert Prioritization

- Typical ML models tend to exacerbate the false positive problem, but ML can also be used as a way to triage and prioritize alerts
- Used by Microsoft Azure Sentinel to identify emerging threats



Example 2: Moving Target Defense

- Can use RL to simulate decisions in response to a potential attack
- Limited examples: give defender a choice to isolate, patch, or leave alone nodes on a network; build simulation and emulation environments to test models
- LOTS of drawbacks right now: lack of realism, computational cost, unclear generalizability, etc.



Example 3: Pentesting and Vulnerability Discovery

DARPA's Cyber Grand Challenge

- ML can be used to prioritize resources and significantly speed up naive automated penetration tests (e.g. using Metasploit)
- But very unclear evidence that ML is successful in finding new vulnerabilities

Questions?

CSET CENTER for SECURITY and EMERGING TECHNOLOGY

- Research at https://cset.georgetown.edu/publications/
- Sign up to receive research the day it's issued, subscribe to our biweekly newsletter, and get invited to our events at https://cset.georgetown.edu/sign-up/
- Watch CSET webinars and request briefings, if needed
- Share your questions and knowledge gaps
