Adversarial ML: How Artificial Intelligence is Enabling Cyber Resilience

Michael Slawinski, Ph.D.
Staff Data Scientist
BlackBerry Cylance

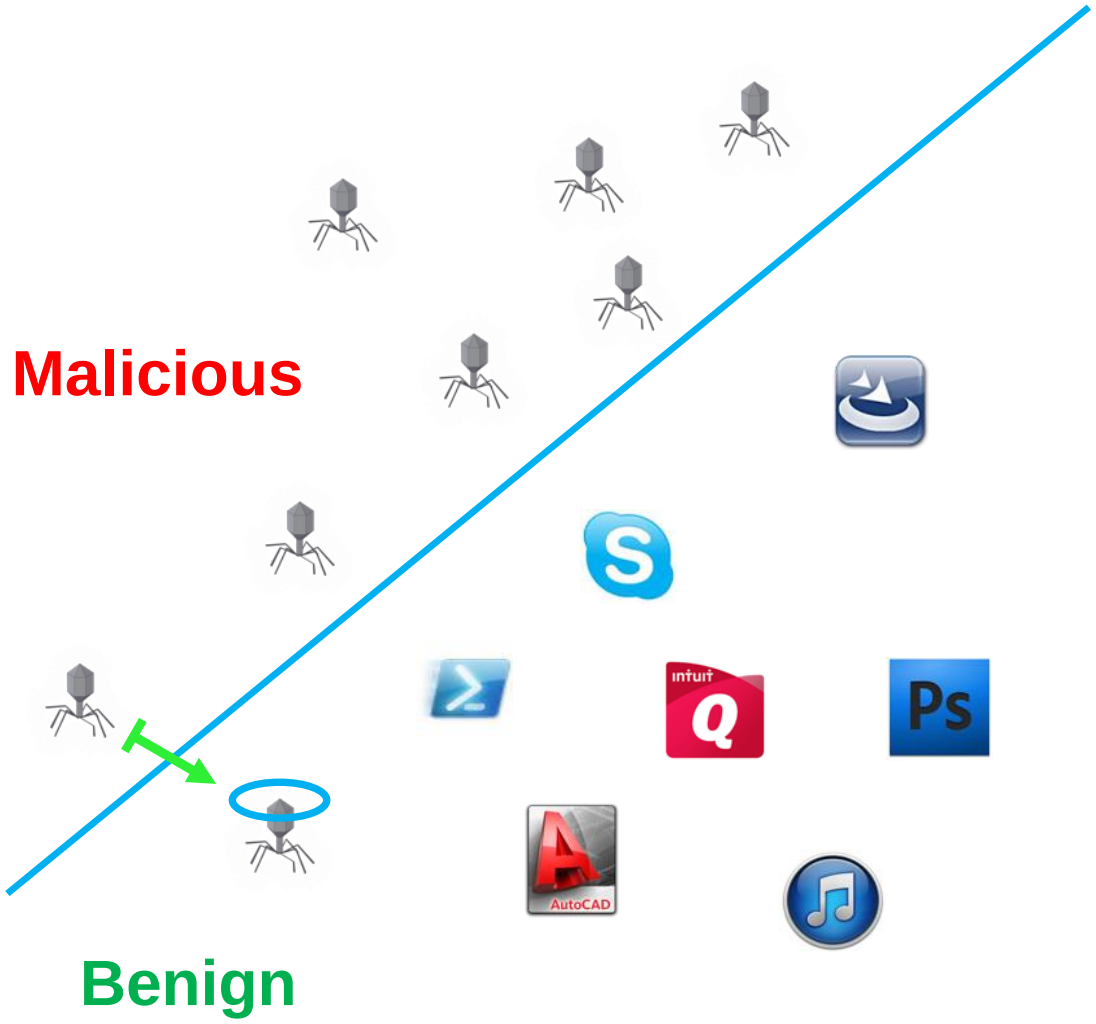
Josh Fu, CISM, CISSP
Security Engineer
BlackBerry Cylance

Agenda

1. File Classification via Machine Learning
2. Weaknesses of Machine Learning Classifiers
3. How these weaknesses can be exploited to make a malicious file look benign
4. How to harden your model against adversarial attacks

Classification

- How to tell if a script/executable/word document/PDF is malicious?
- Many security vendors have large labeled datasets
- Use these labeled samples to train a classifier
- How can this classifier be attacked?
  - Is it possible to perturb a file in such a way as to cross the decision boundary?



Naiveté is Dangerous

We make the mistake of assuming the model is judging as we judge.

In other words, we assume the machine learning model has baked into it a conceptual understanding of the objects being classified…

Example: Lie Detectors – what is a lie?

Human Point of View	Lie Detector Point of View
Lie is a statement believed to be false but offered as true	Heart rate above a threshold Perspiration above a threshold Body movement above a threshold

Machine Learning – definition and utility

Definition: The design and implementation of learning algorithms.

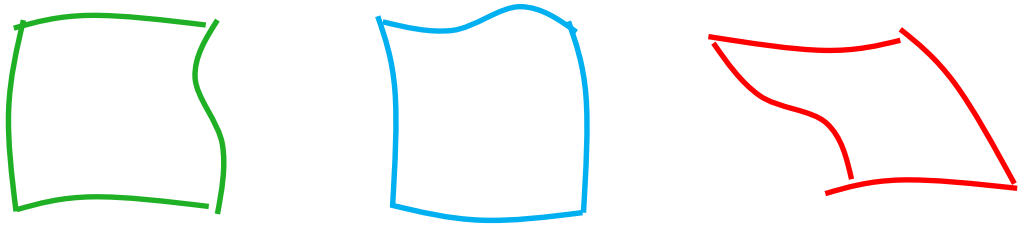
Learning Algorithms: Algorithms which are not explicitly programmed

Decision making scheme is a result of optimizing some objective function based on observed data

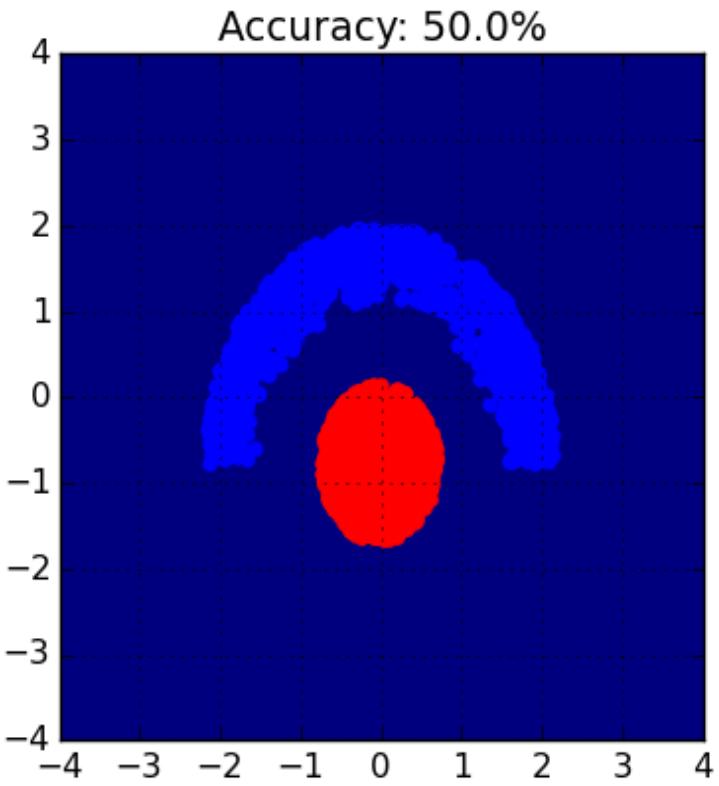
Utility of Learning Algorithms: Why are learning algorithms necessary?

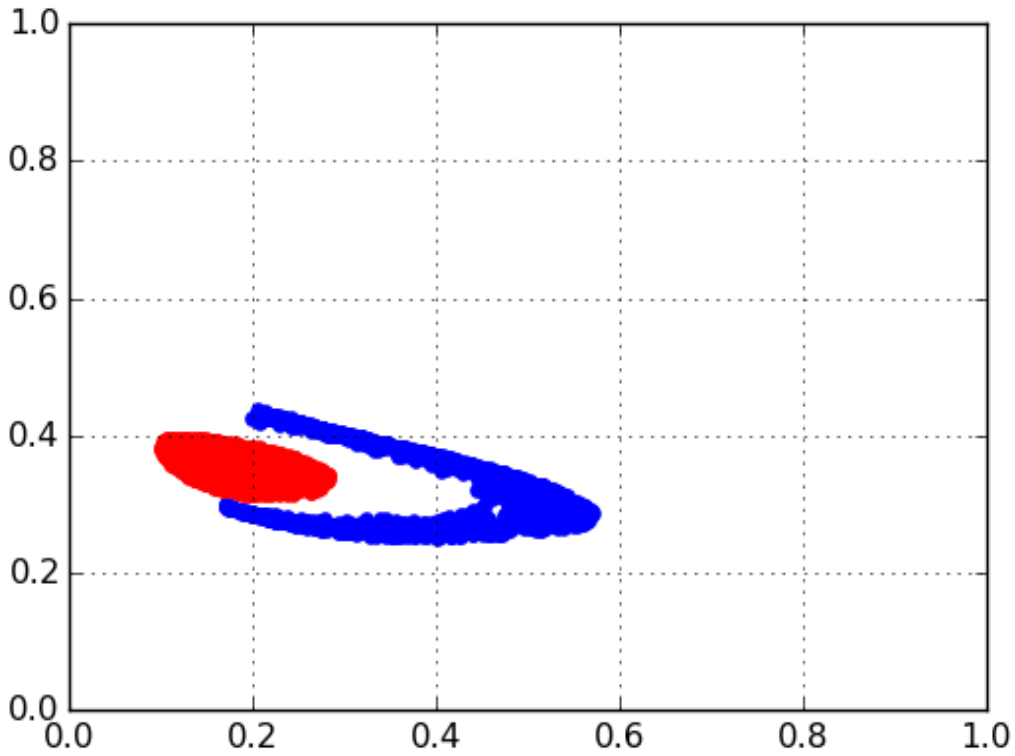
Extreme data variability resulting in a possibly infinite signature set

Example: Hand-drawn squares

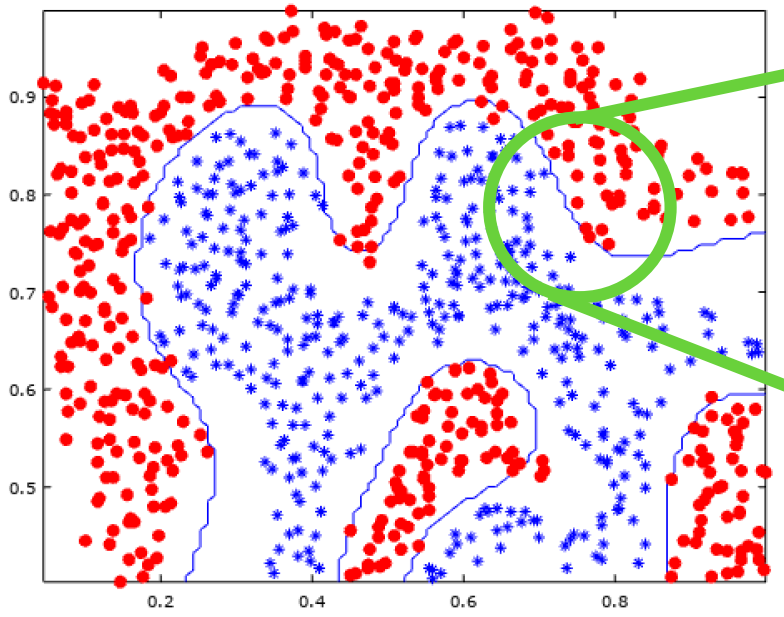


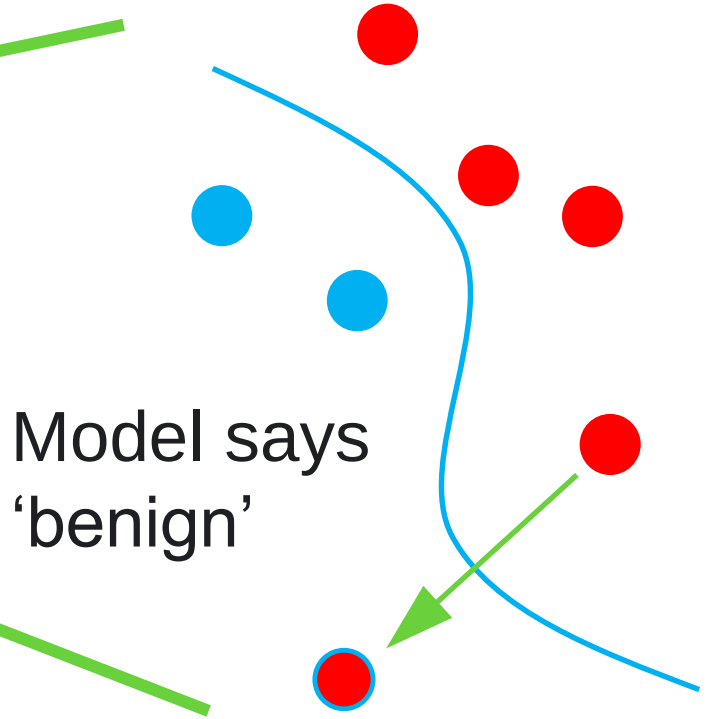
Machine Learning – definition and utility





Adversarial Examples



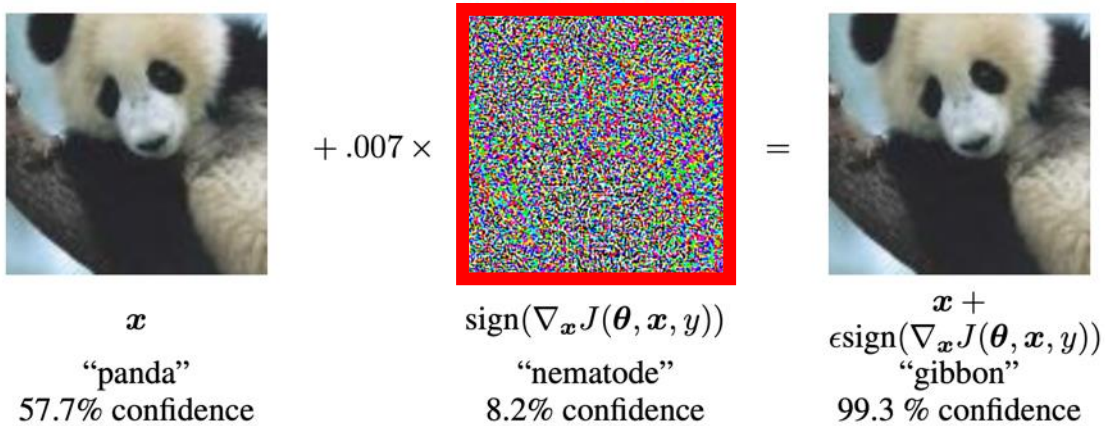


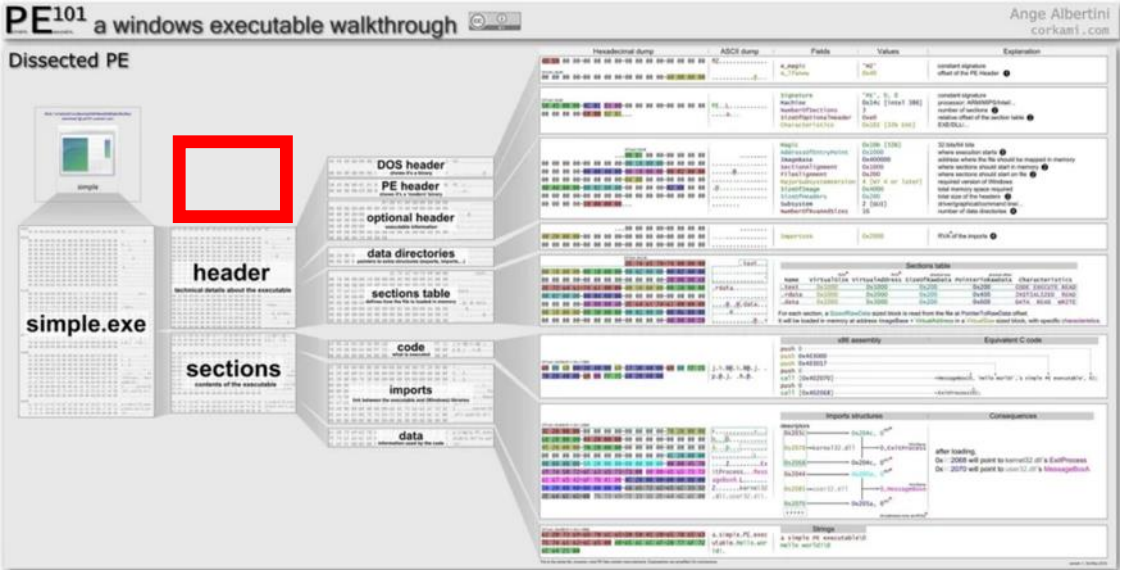
If a new sample lands within the boundary, we classify it as benign. Otherwise, malicious.

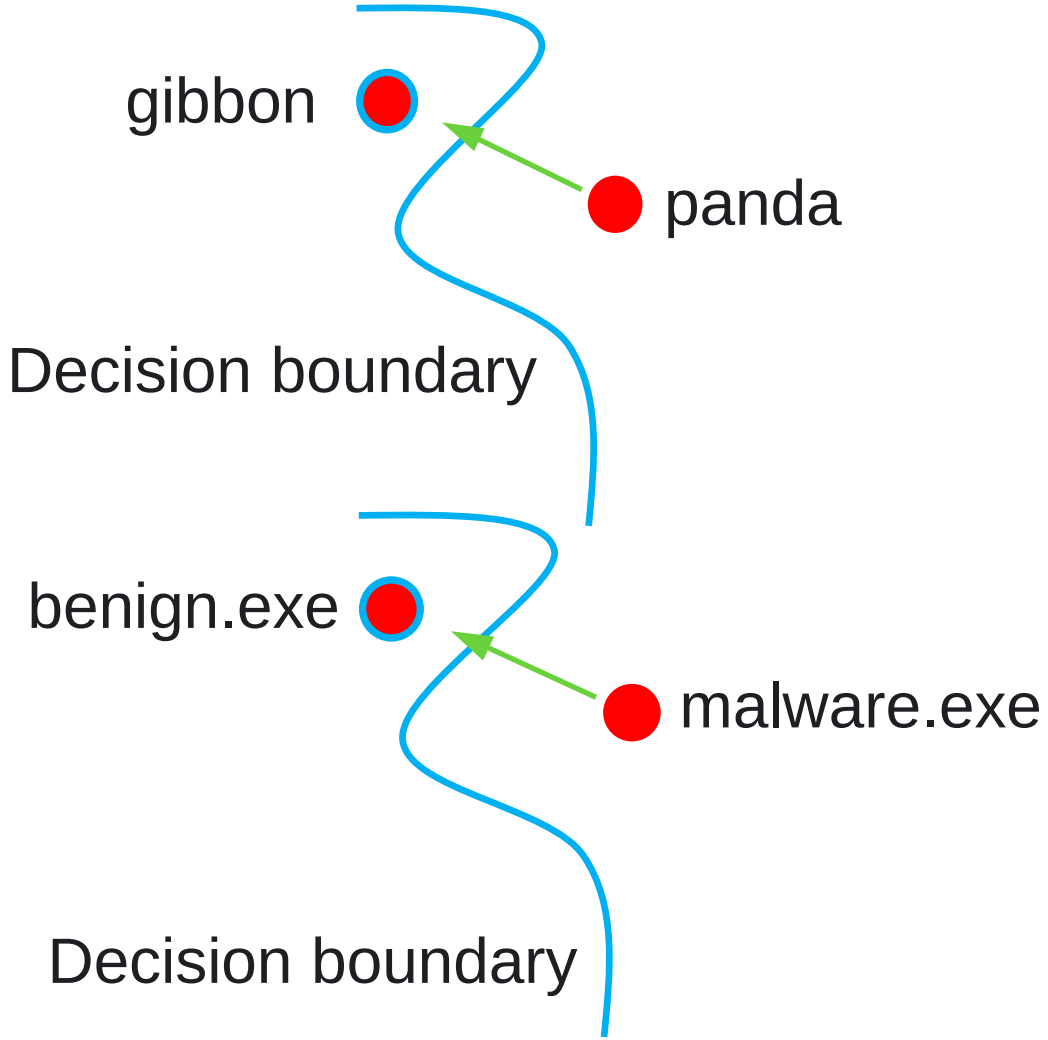
Functionally

Model point of view

Adversarial Machine Learning







BlackBerry | CYLANCE

Vulnerability to Adversarial Attacks – ‘Snow Features’

Training Set

Model Says: ‘Wolf’ 99.9%

Upshot: Model overemphasizing white, i.e., snow pixels due to high wolf/snow correlation

Vulnerability to Adversarial Attacks

Snow Features: model features which are not fundamental to the nature of the object being classified, but which aid in classification due to co-occurrence.

Models trained on such features are **vulnerable** to adversarial attacks

Key Takeaway:

Malicious and Benign files have snow features too!

Examples:

Certain strings, header information, file size, etc.

CylancePROTECT® Defenses Against Adversarial Attacks

1. **Reduce leveraging of ‘snow features’**
    - Features which when perturbed, change the nature of the original file

2. **Adversarial Training**
    - Leverage more sophisticated learning algorithms

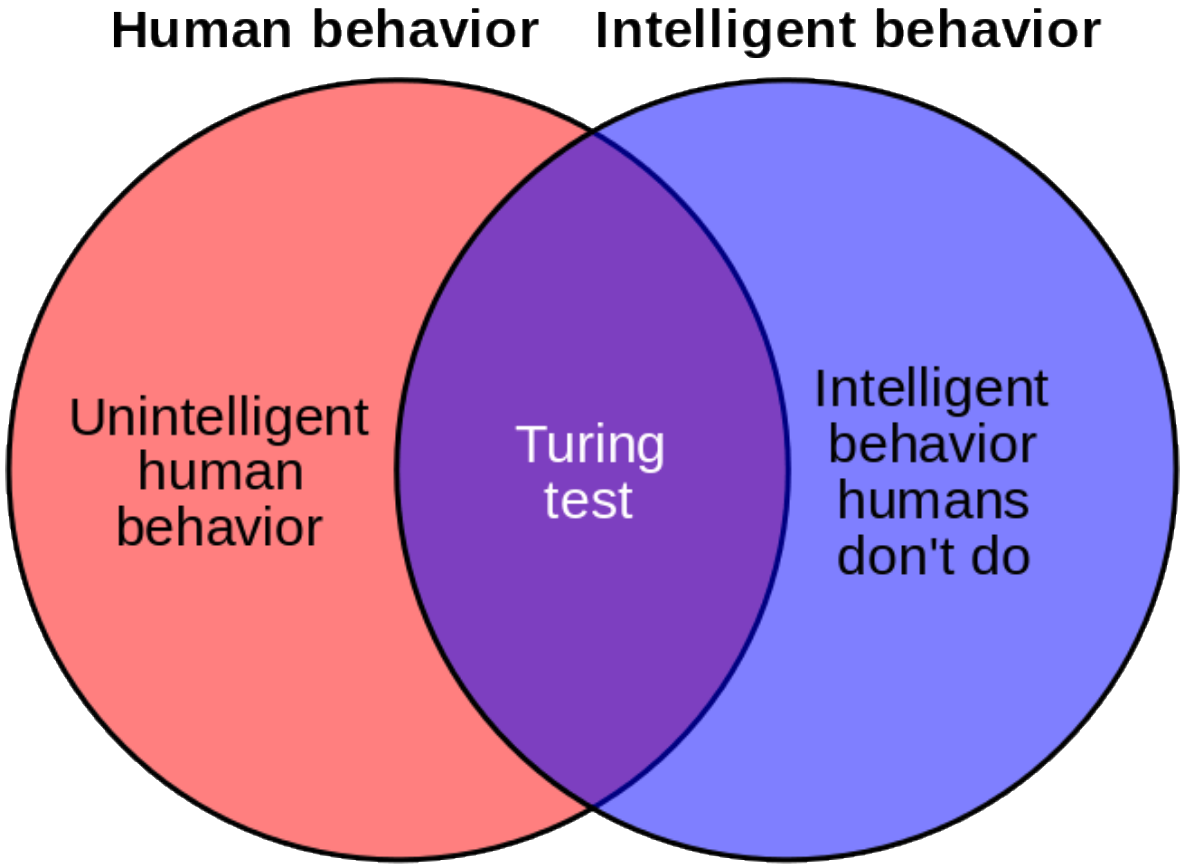
3. **Anomaly Detection**
    - Learn probability distribution of the training set so that files altered to trick the classifier will be assigned very low probability

4. **Inclusion of Adversarial Examples in Training**
    - Construct adversarial examples at scale and add them to the training set to familiarize the model with such samples

Practical Applications

Why CAPTCHAs

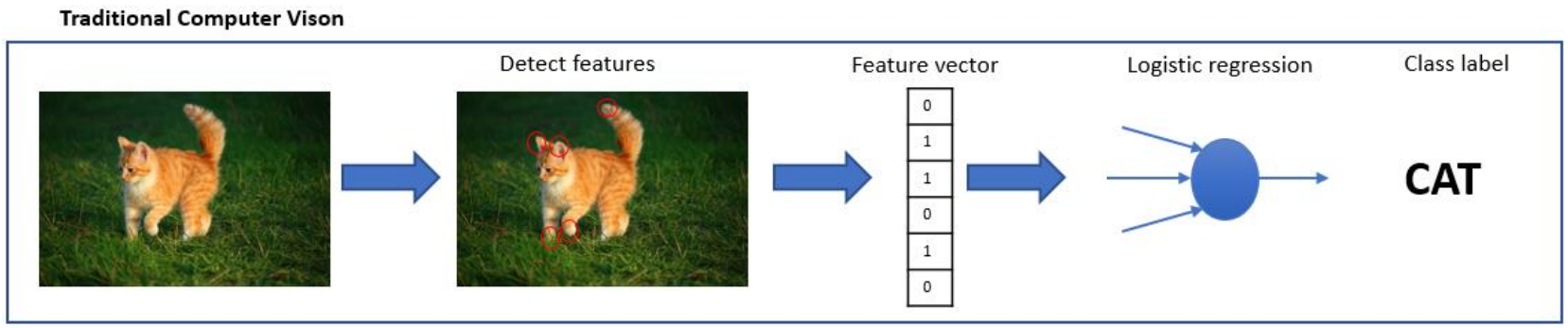
- CAPTCHA (Completely Automated Public Turing test to tell Computers and Humans Apart)
- Turing Test

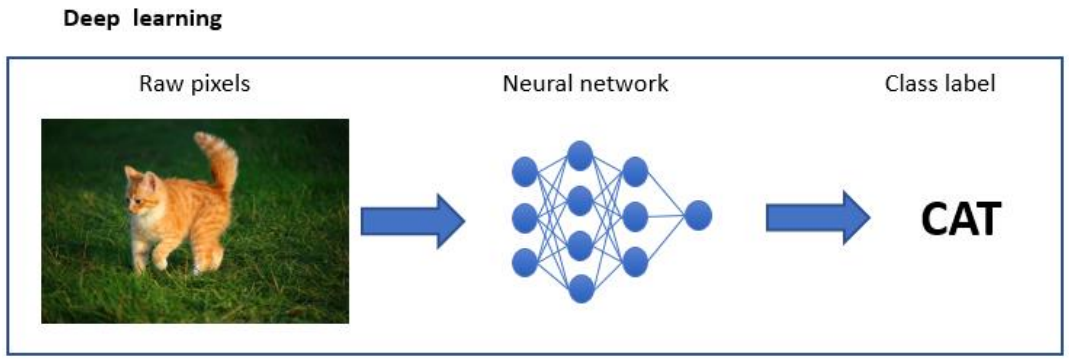


Source: https://www.maketecheasier.com/captchas-why-we-need-them/

Usage vs Trickery

- CNN for CAPTCHAs
- Natural Language Processing (NLP)
  - Twitter Bot - John Seymour and Philip Tully at ZeroFOX in 2016
  - Laser phishing
- Speech to Text





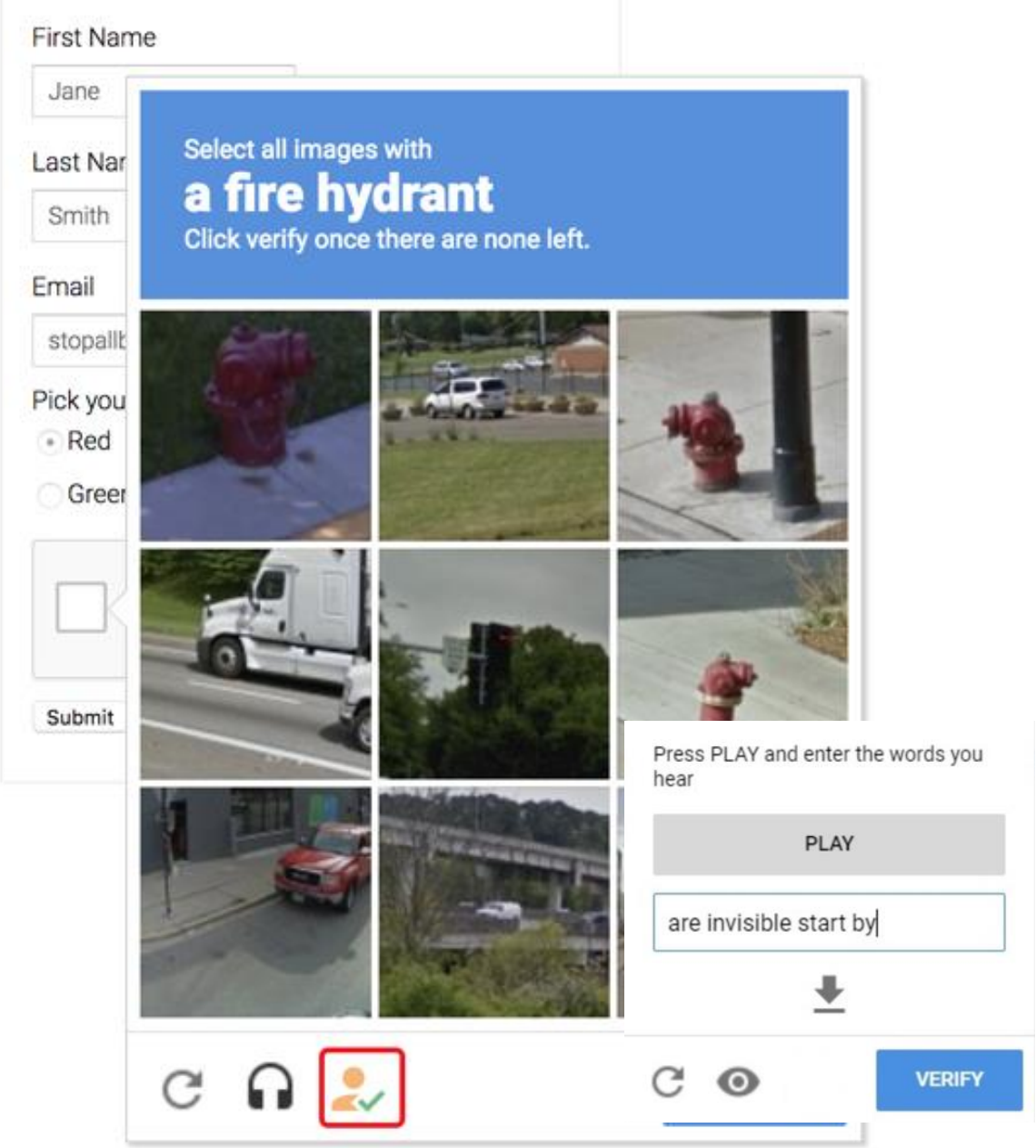
Today’s CAPTCHA Breaking Techniques

I will always choose a lazy person to do a difficult job…because he will find an easy way to do it.

- Bill Gates

Browser Extension

- Buster: Captcha Solver for Humans by Armin Sebastian
- Uses Google's own Natural Language Processing
- This is the state of security today



BlackBerry | CYLANCE

Noise as Defense

- Mitigate model discovery
- Optical vs Digital classification
- Noise is one way of enabling resilience

```
Iteration: 6
Source score:   0.00%, class-number:  837, class-name: sweatshirt
Target score:  99.09%, class-number:  300, class-name: bookcase
Gradient min: -0.000029, max:  0.000021, stepsize: 244856.71
```


```
Noise min: -3.000, max: 3.000, mean: -0.001, std: 0.668
```

TensorFlow Tutorial #11 Adversarial Examples

Key Takeaways

1. Tools can be easy and cheap
2. ML will have a role in the arms race
3. The next frontier for ML → solving identity and authentication challenges

How ML Applies to Cylance

- Recognition of files → CylancePROTECT and Cylance Smart Antivirus™
- Recognition of processes → CylanceOPTICS™
- Recognition of people → CylancePERSONA
- Additional Benefits:
  - Advanced enough to even work offline and on old systems with no updates
  - Faster threat hunting
  - CylancePROTECT also works against fileless attacks and scripts
  - Faster and effective CylanceCONSULTING Services (IR, CA, M&A, Staff, Threat Hunting, and more

Get a Demo

See the next generation in endpoint security solutions – and take that first step towards getting your organization to a state of prevention.

THANK YOU!

- Cylance
  www.cylance.com
- Best Machine Learning Resources for Starters
  https://machinelearningmastery.com/best-machine-learning-resources-for-getting-started/
- Twitter Story
  https://www.nytimes.com/2018/11/19/science/artificial-intelligence-deepfakes-fake-news.html
- Google cloud Vision API
  https://cloud.google.com/vision/pricing
- Y Combinator
  https://blog.ycombinator.com/how-adversarial-attacks-work/
- DEFCON 16: CAPTCHAS: Are they really hopeless? (Yes)
  https://www.youtube.com/watch?v=8ic1LToPsro
- Use Vision API to parse captcha screenshot
  https://cloud.google.com/vision/docs/reference/rest/

Twitter: @cylanceinc and @jfusecurity

Github: https://github.com/gtownrocks

BlackBerry | CYLANCE

Questions
and
Answers