RSAConference2018
San Francisco | April 16 – 20 | Moscone Center

SESSION ID: HTA-R02

NAVIGATING THE LABELING BOTTLE-NECK AS SECURITY EMBRACES AI

Philip Tully, PhD

Principal Data Scientist
ZeroFOX
@phtully

Hyrum Anderson, PhD

Technical Director of Data Science
Endgame
@drhyrum

Obligatory Disclaimer (h/t @milalaranjeira)

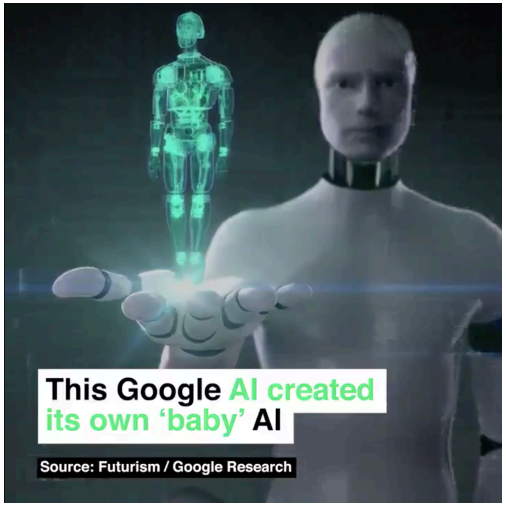

```python
from keras.layers import *
from keras.models import *
from .data import load_data

x, y, x_test, y_test = load_data()

def get_model(num_layers):
    model = Sequential()
    for _ in range(num_layers):
        model.add(Dense(100, activation='sigmoid'))
    model.compile(loss='mse', optimizer='sgd')
    return model

best_model = None
best_loss = None

for i in range(1, 10):
    model = get_model(i)
    model.fit(x, y)
    loss = model.evaluate(x_test, y_test)
    if best_loss is None or loss < best_loss:
        best_loss = loss
        best_model = model
```



THE ARTIFICIAL INTELLIGENCE "APOCALYPSE"

There is no existing credible evidence demonstrating AI has been leveraged to wage an attack in the wild.

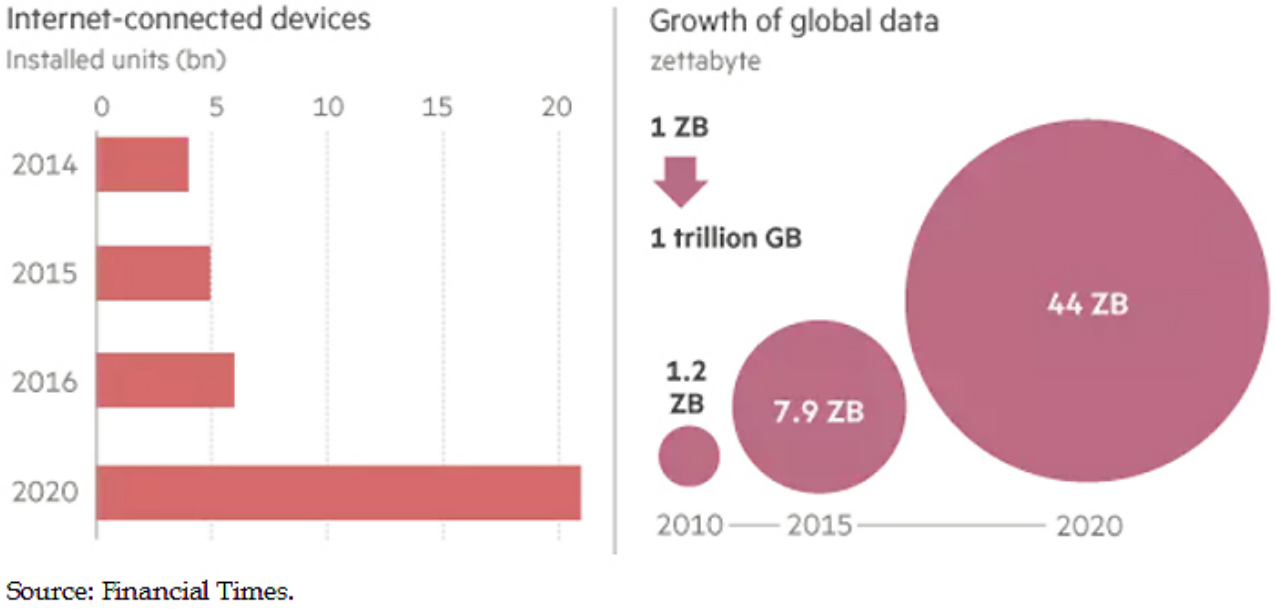
Outline

- From Signatures to Statistics
- The Data Labeling Bottleneck
- Reducing the Bottleneck
- Applications and Lessons Learned

FROM SIGNATURES TO STATISTICS

The Evolution of Data-Driven Security

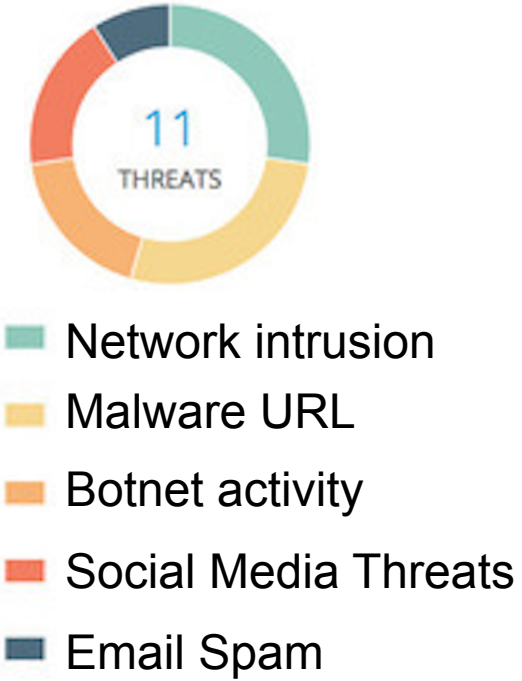
Why is this happening now?



Source: Financial Times.

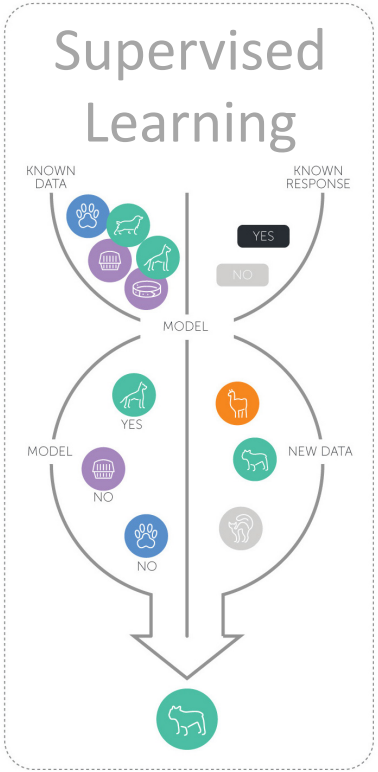
Why is machine learning useful in infosec?

- Shift from bottom-up to top-down
  - Bottom-up was ***reactive***, new attacks require new signatures, attackers always one step ahead
  - Top-down is ***proactive***, allowing defenses to generalize better to new attacks
- Data-driven methods
  - Square pegs in a world with many round holes
  - Not simply overkill - performs worse often
  - However, for many problems, they're effective



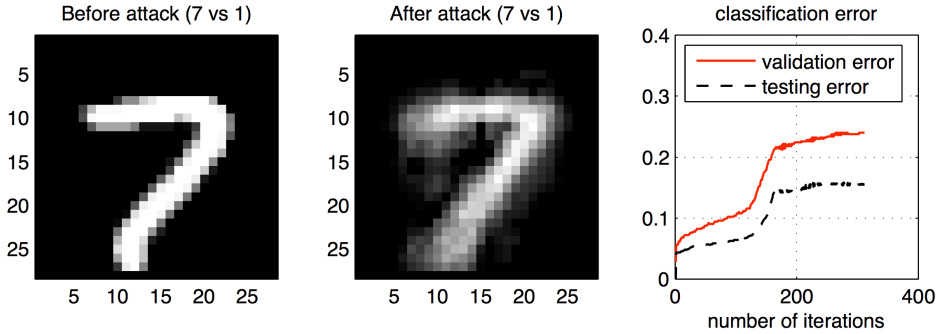
What do Use Cases Share in Common?

- Defensive in nature
  - Aligns w/ posture of industry at large
- Rely on supervised machine learning
  - Maps to defender's use case
    - *distinguish malicious from benign*
- Supervised v. unsupervised learning
  - A dependence on pre-existing labels



Attacking Supervised Machine Learning

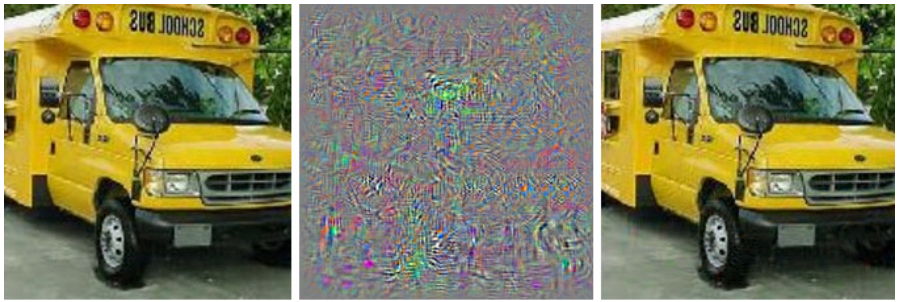
Poisoning Attack (before training)



Biggio *et al*. (2012)

Evasion Attack (after training)

bus + distortion = *ostrich*



Goodfellow *et al*. (2014), Szegedy *et al*. (2014)

This might be easier to do than you think

Real-world examples of attacking ML

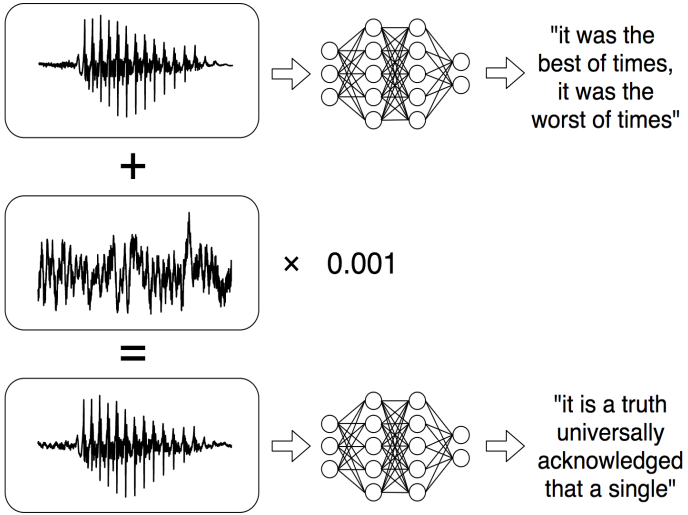
Sticker Attack on Self-Driving Cars

Evtimov *et al.* (2017)

Eyeglass frames attack on facial recognition systems

Sharif *et al.* (2016)

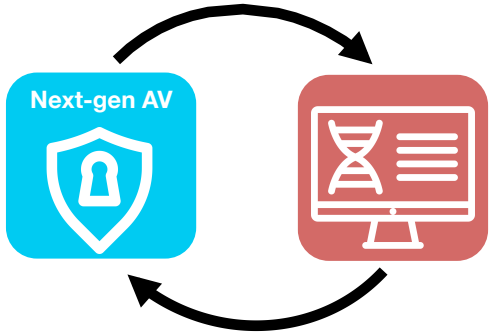
Okay Google, Open the Front Door



Carlini *et al.* (2016)

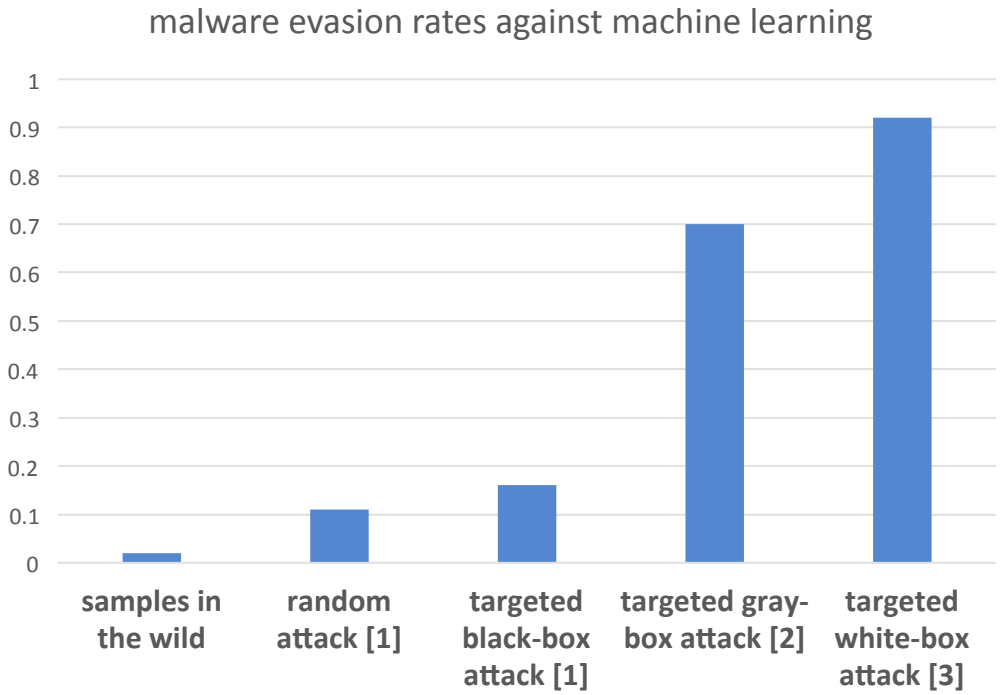
"Okay Google, browse to evil.com"

Evading Machine Learning Malware



Evasion depends on adversary knowledge:

- Black Box – good/bad from query
- Gray Box – “confidence score” from query
- White Box – compute gradients on deep learning



ZEROFOX ENDGAME.

[1] Anderson, et al. 2018, [2] Kharkar et al., 2017, [3] Kreuk et al., 2018

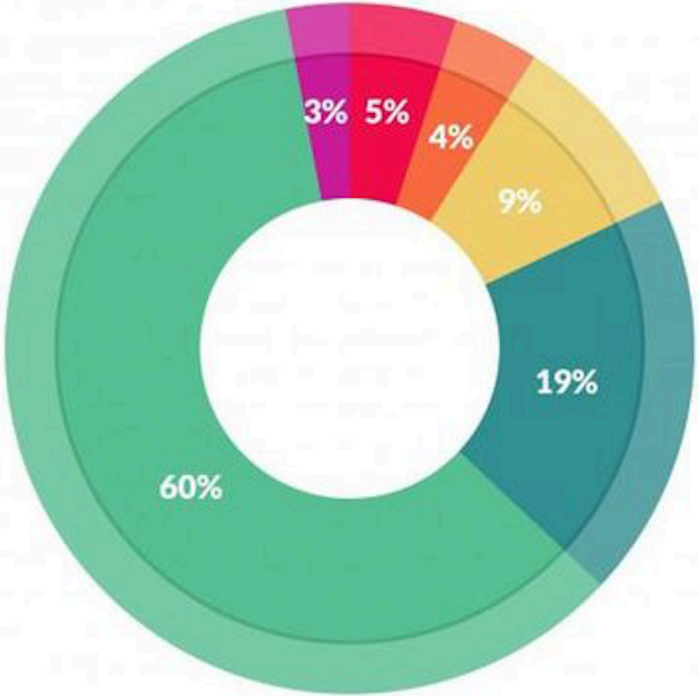
OFFENSE IS STRUCTURALLY DOMINANT IN AI

The Data Labeling Bottleneck

What is the Data Labeling Bottleneck?

- *Cleaning and organizing data: 60%*
- *Collecting data sets; 19%*

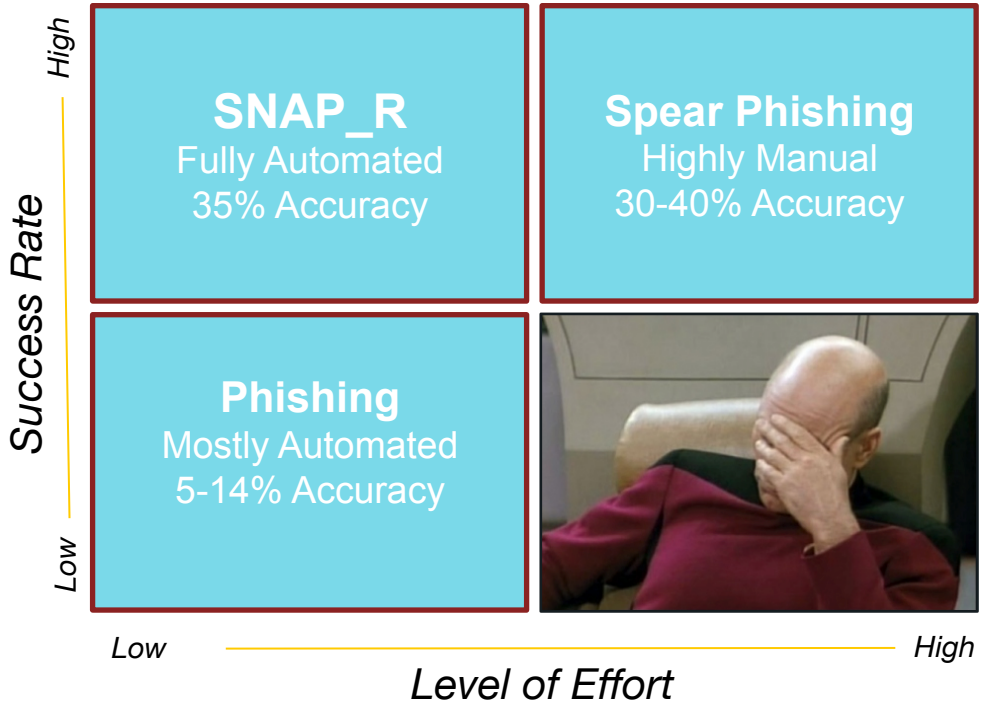
- Challenge: to create a sufficiently large, labeled training data set, but labeling:
  - Is expensive and time consuming
  - Requires domain expertise
  - Not always repurposable elsewhere
- Deeper models = millions of free params, requiring commensurately more labels



Learning to Spear Phish

- Cluster high value targets
- Microtarget w/ LSTM-generated text
- Why Twitter?
  - Bot-friendly API
  - Colloquial syntax
  - Abundant personal data

Seymour & Tully (2016)



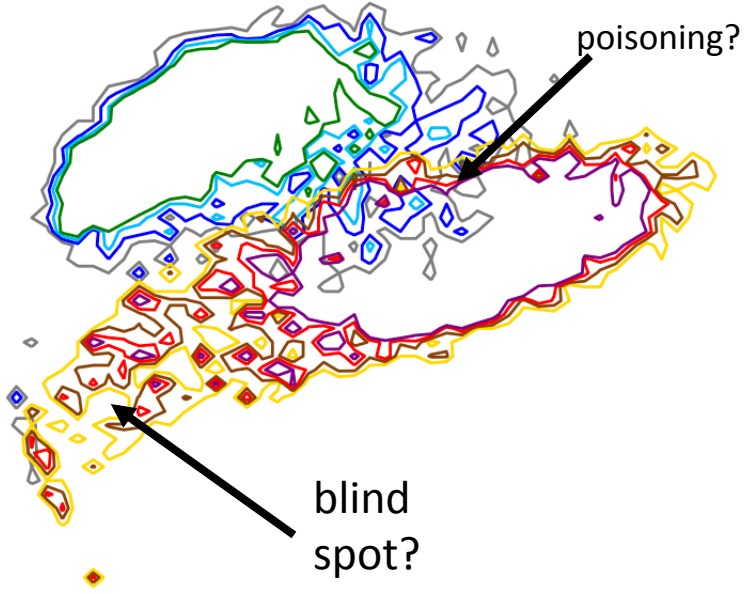
SNAP_R Simulation

- @reply + LSTM + goo.gl URL
  - RNN w/ 3 layers, ~500 units/layer
  - Click rates of IP-tracked goo.gl URLs

```
[redacted] Woody Allen and Iron &amp;amp; Wine at the weather for Sunday #WinterWeather #freezing #belowzero https://goo.gl/wN42Vh
[redacted] Every Wednesday is our Gruffalo style WOODLAND ADVENTURE!! https://goo.gl/3sSPCj
[redacted] Bonne première mi temps de Cabaye en tout cas bonne nuit à tout le monde sauf aux portugais https://goo.gl/23C4AD
[redacted] sus icons son muy lindos, yo amé todos los que me ha respondido una mención que la hice hace mil. https://goo.gl/fYkqYj
[redacted] 🇺🇸#VeteranOfTheDay Check out Happy life with Ice Cream! Available for the next few days via https://goo.gl/sBHnZH
[redacted] Happy birthday to my unit🅾https://goo.gl/1wAuDw
[redacted] Over 20 blocks with Pokemon Go! Grosse Pointe North High School https://goo.gl/dvLMa9
[redacted] Let's get out the vote for the acorns https://goo.gl/XoTtmO
[redacted] Nominate yourself for the tip! https://goo.gl/yBF52k
[redacted] Online news story about our Teen-Read-a-Thon at Hayters Gap library Saturday mornings at 9:30. https://goo.gl/RbxR7r
[redacted] tell me how u gon call me rude when you don't even have your own shit figured out 😂😂😂https://goo.gl/YyloKH
[redacted] Gif #Pokemon #PokemonGo #PokemonGoPolska #Pokemon https://goo.gl/ws15pZ
[redacted] ahahaha no way?! Yeah I saw One Direction in concert 🥰💔https://goo.gl/9Gpf9r
[redacted] Stay connected on Twitter and share engaging content with your followers? You don't have to! Use https://goo.gl/e56Sqj
[redacted] When someone asks if you want to hang out in Eric's basement 🙌https://goo.gl/nZVYhF
[redacted] ya Allah,,, mimpi yang bukan2 aku ni.. Haishh.. Awak.. Byk btol yg awak follow guna ig sy ni https://goo.gl/2TsBCW
[redacted] Go on over and back to work on Kaiju today. Monster powers are fun. https://goo.gl/5CrGGz
[redacted] the streets of the ugliest shits I've seen https://goo.gl/lhTreO
[redacted] #Bitcoin Is Not What You Think - #GoogleAlerts #databreach #cybersecurity #payments https://goo.gl/Viuelb
[redacted] detta är verkligen ett skämt. Påväg hem från alla riktningar på stan... https://goo.gl/mD22Pi
[redacted] Oye Kya Ho Aap!! https://goo.gl/rmnxhx
[redacted] Découvrir la puissance de l'univers Marvel. https://goo.gl/dcPDtt
```

History Repeating Itself?

- Thinner margins of error
- Defense only as secure as weak link
- Attention: coverage vs. details
- What does success look like?
  - Defender: approach 100% detection rate
  - Attacker: 1 out of 100 works
- Blue teams have more at stake
  - Accuracy more important on defense



Democratizing AI: Lowering Barrier to Entry

Hardware

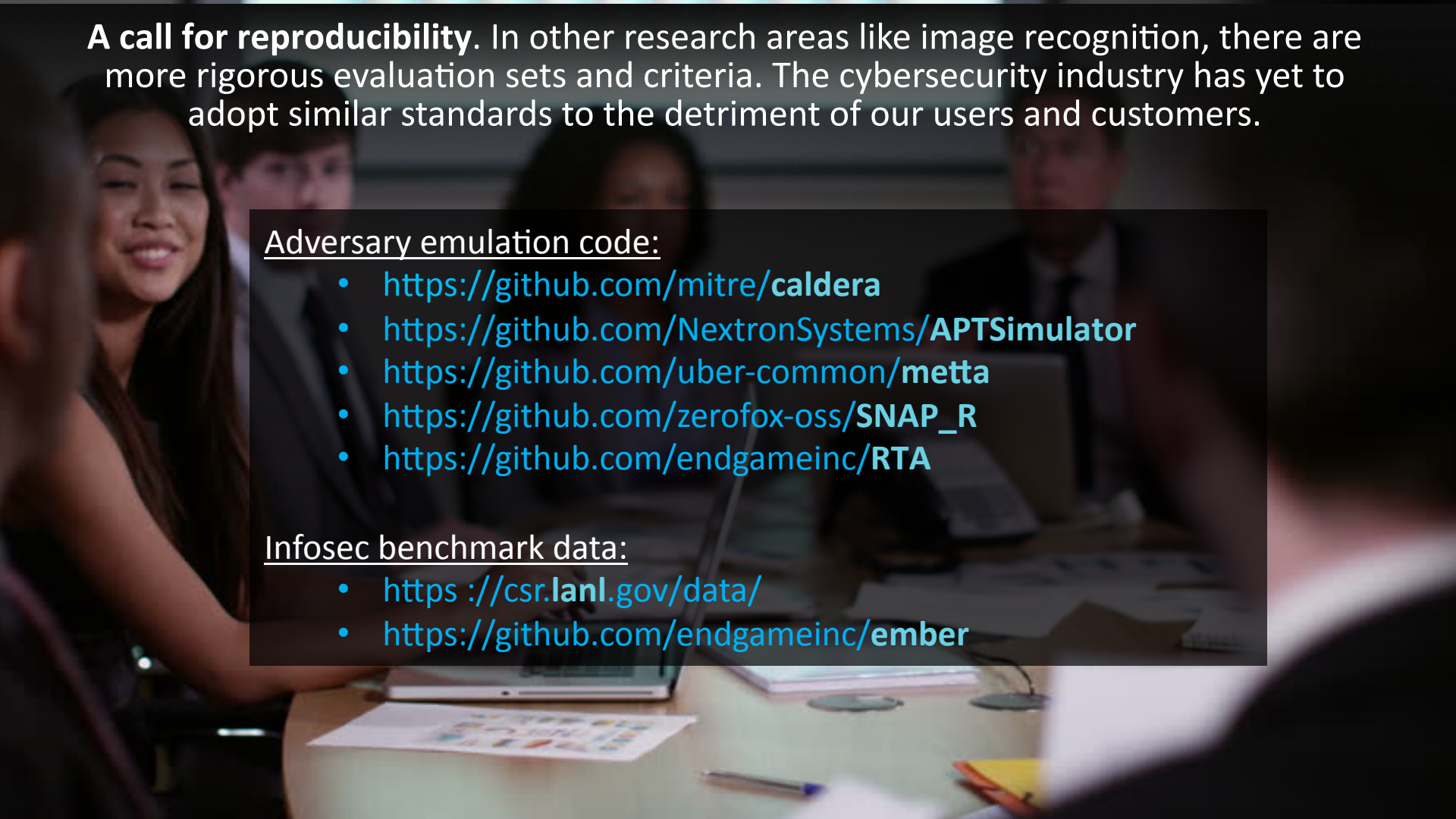
- Parallelization
- Cloud
- GPGPU
- TPU
- Quantum (Post Moore's Law)
- Neuromorphic

Software

- Rise of Deep Learning
  - Automated feature engineering
- Free educational resources
  - Coursera, StackOverflow, GitHub
- Open source datasets/pre-trained models
- Professional quality open source libs
  - TensorFlow, Torch, Keras, Caffe
- Tools to obviate need for domain experts
  - AutoML, SageMaker, auto-sklearn, etc.

Cyber Landscape - Advantage: Attackers

- Security incidents rarer (unbalanced data)
- Labor shortage
- Attack transferability (cross-evasion)
- Fragmented vendor landscape
  - Intellectual property warfare in cyber market
  - Vendor competition stifles collaboration
  - Labeled data sharing virtually non-existent



A call for reproducibility. In other research areas like image recognition, there are more rigorous evaluation sets and criteria. The cybersecurity industry has yet to adopt similar standards to the detriment of our users and customers.

Adversary emulation code:

- https://github.com/mitre/**caldera**
- https://github.com/NextronSystems/**APTSimulator**
- https://github.com/uber-common/**metta**
- https://github.com/zerofox-oss/**SNAP_R**
- https://github.com/endgameinc/**RTA**

Infosec benchmark data:

- https ://csr.**lanl**.gov/data/
- https://github.com/endgameinc/**ember**

THE BEST DEFENSE IS A GOOD OFFENSE

Reducing the Data Labeling Bottleneck

Dealing with the Data Labeling Bottleneck

Process Approach

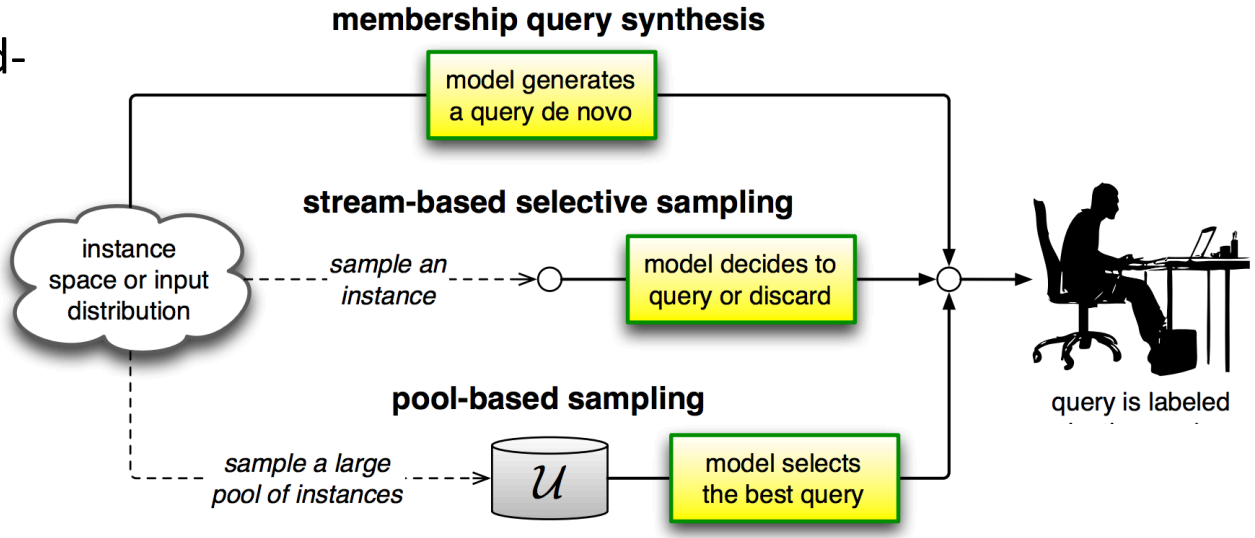
- Crowdsourced Data labeling
- Active Learning

Technical Approach

- Semi-supervised learning
- Transfer Learning
- Adversarial Learning

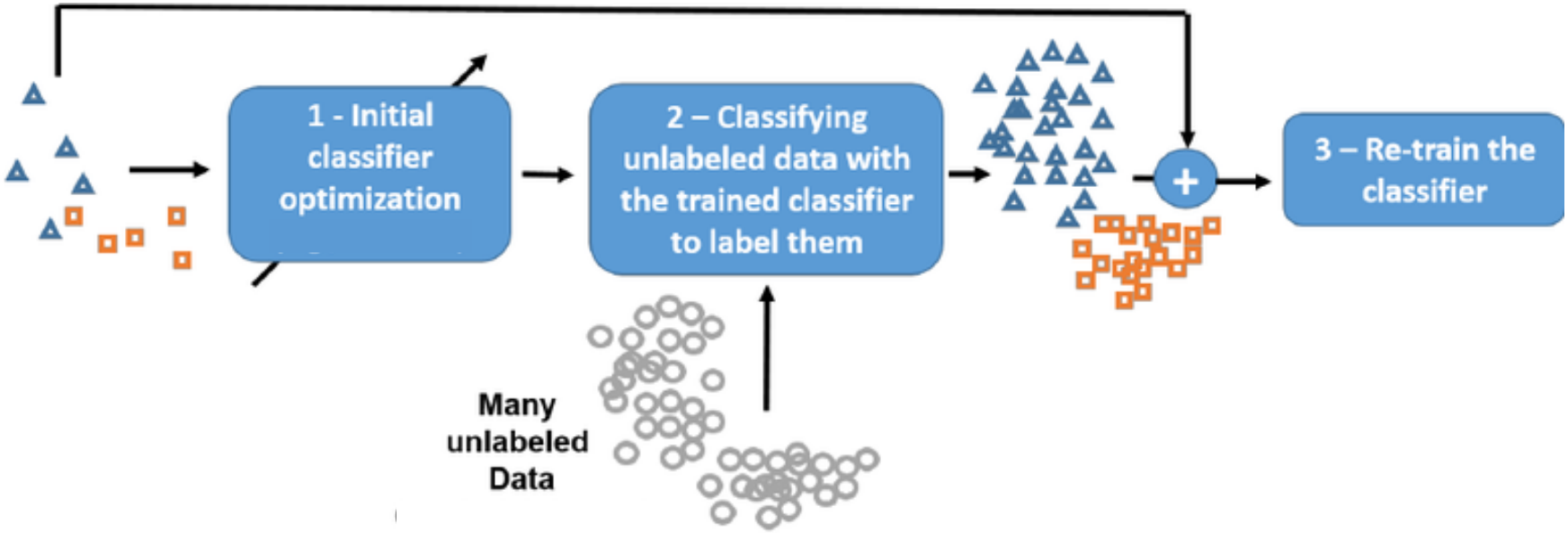
Process-based Label Acceleration

- **Crowdsourcing**: pay on-demand workforce for simple, repetitive, hard-to-automate tasks
  - 'Reputational' - reward quality work
  - Crowd consensus v. accuracy of experts
- **Active Learning**: slower humans label only the most important data



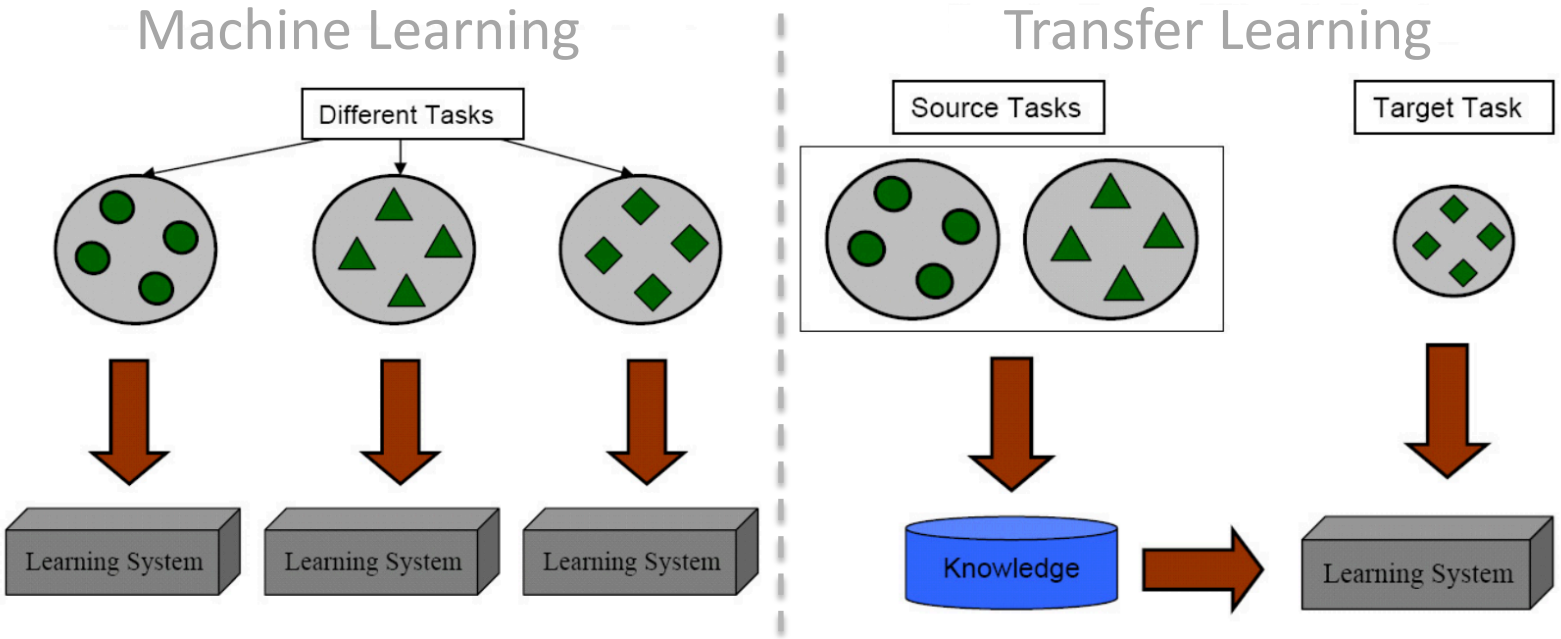
Settles (2012)

Semi-Supervised Learning



Models trained on limited labeled data learn problem structure from unlabeled data, Chappelle *et al.* (2009)

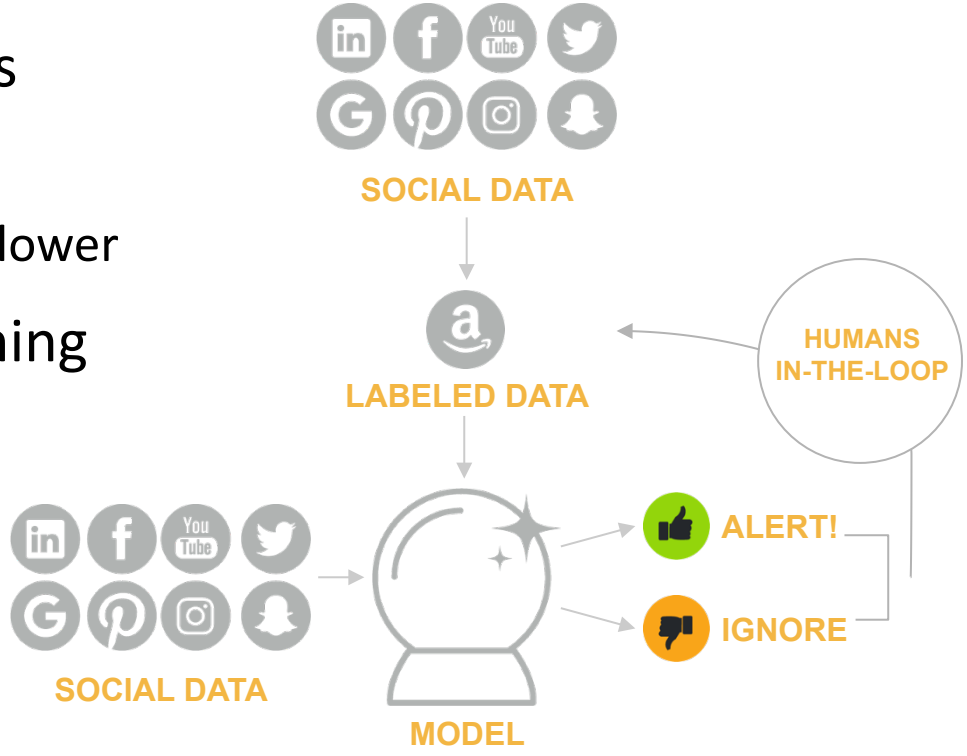
Transfer Learning



Models trained for a problem with copious labeled data tailored for a problem w/ limited labeled data, Pan and Yang (2010)

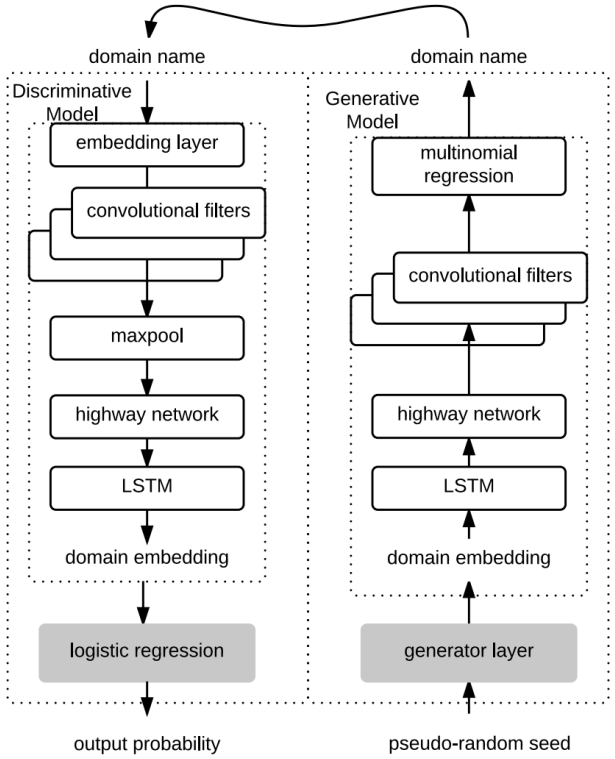
Ex. 1: Accelerating Label Acquisition

- Detecting social media threats
- Crowdsourced data labeling
  - Amazon Mechanical Turk, Crowdflower
- Active, Semi-supervised Learning
  - Humans-in-the-loop

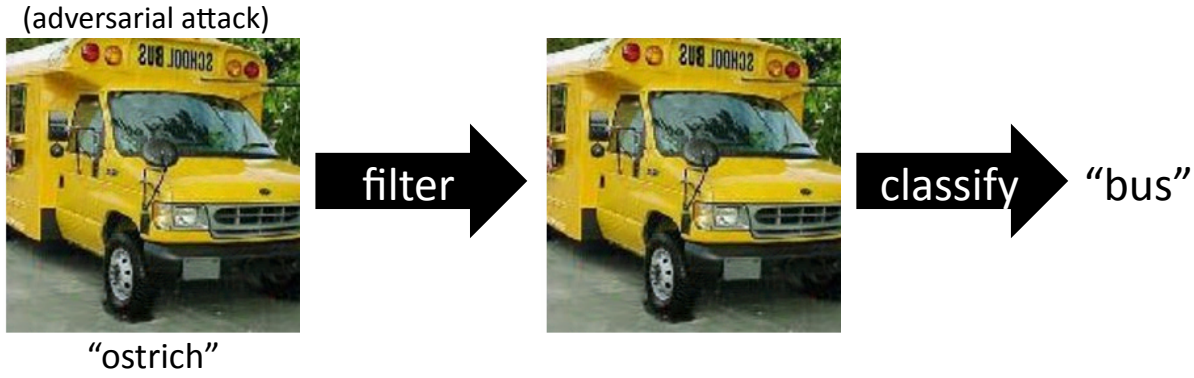


Generating Adversarial Samples

- Machine learning red teaming / pentesting
- Pre-emptively wage attacks against your own models to plug up holes
- Augmenting training data w/ adversarial samples = ⇧detection, Anderson *et al.* (2016)
- Cleverhans: benchmarking adversarial vulns https://github.com/tensorflow/cleverhans



Ex. 2: Unlabeled data reduces attack success



“filter” learns from the **universe of <u>unlabeled</u> images**

- PixelDefend reduces attack success from 100% to < 9% (Song et al., 2018)
- Similar concept: 1st place defensive prize in NIPS 2017 competition (Liao et al., 2017)

PEDefend: use unsupervised learning to learn the <u>universe of PE files</u>

APPLICATIONS AND LESSONS LEARNED

Using Data-Driven Techniques to Stay a Step Ahead

Apply What You Have Learned Today

- **Next week:** seek trusted, critical feedback for your ML defenses
  - 3rd party testing critical
  - Adapt “think like an adversary” mindset; ML pen-testing
- **In 3-6 months:** Establish process to improve data/label quality
  - Even more critical than attention to models and features
- **12+ months:** Research ways to leverage unlabeled data
  - Incorporate semi-supervised & transfer learning