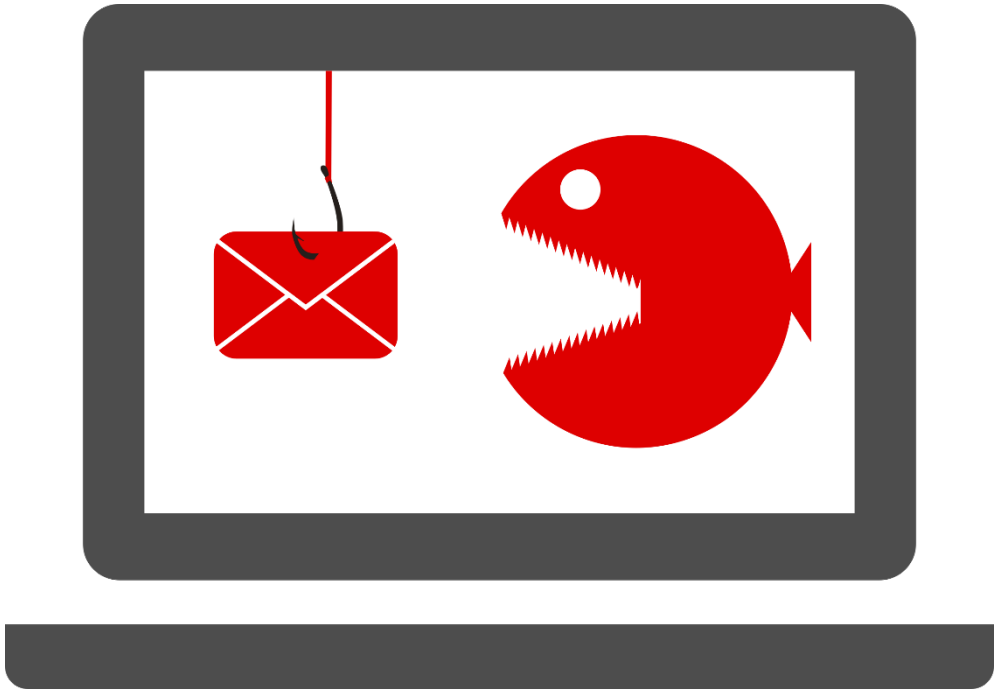
Using Generative Adversarial Networks to Improve Phishing Domain Classifiers

Jen Burns

Emily Heath, PhD

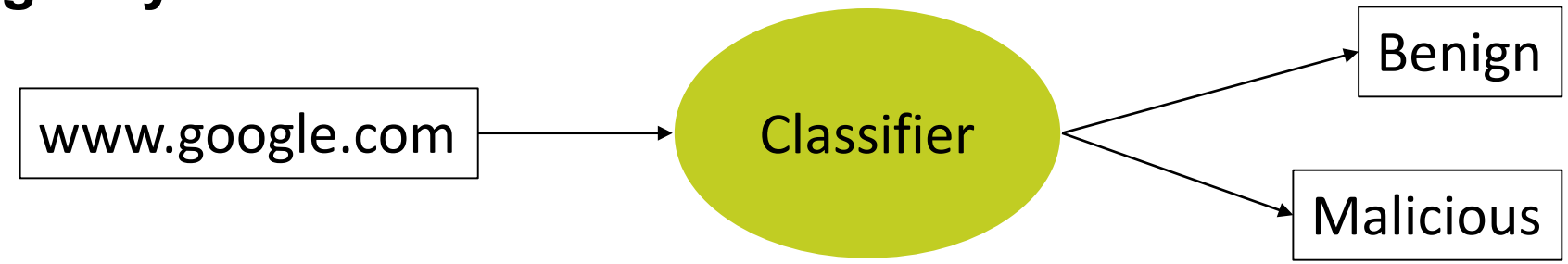
The MITRE Corporation



MITRE

Categorizing Phishing Domains

- **Phishing domains widely used by APT groups and criminal actors**
  - Goal: Obtain usernames, passwords, credit card information, etc.
- **Disguised in content that looks identical or nearly-identical to legitimate service or web site**
  - Domain name itself can be helpful in distinguishing
- **Scope of our work: Detect phishing domains from legitimate domains using only the domain name**

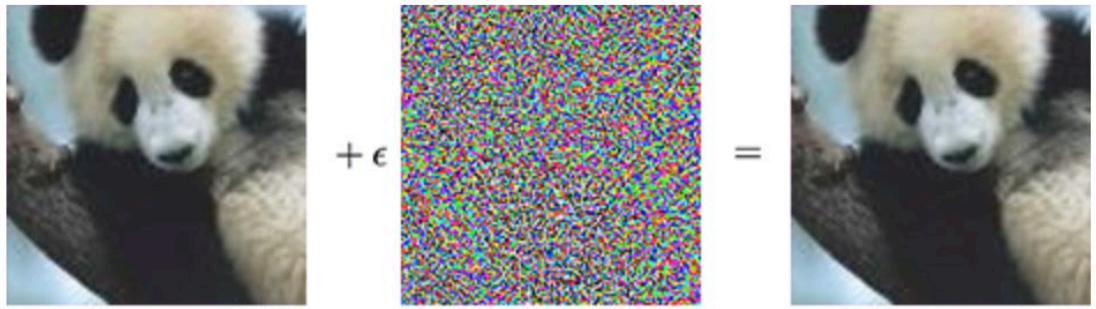


- **Example Phishing Domains**

  - amazon[.]co[.]uk[.]security-check[.]ga, update-apple[.]com[.]nbetawihosting[.]net

Adversarial Machine Learning

Evasion Attack



"Explaining and Harnessing Adversarial Examples"
https://arxiv.org/abs/1412.6572

Actual unimpressed gibbon

How can we build a robust machine learning model to detect phishing domains and overcome these evasion attacks?

MITRE

Generative Adversarial Machine Learning

- **Generative Adversarial Network (GAN)**
  - Two neural networks contesting with each other in zero-sum game
    1. **Generator:** constructs candidates
       - Goal: Synthesize examples that appear to be from a desired distribution
    2. **Discriminator:** evaluates candidates
       - Goal: Distinguish between synthetic samples and true population

Idea:

Use a GAN to develop synthetic ‘phishing’ domains (adversarial examples) and improve the strength of a machine learning classifier.

Hypothesis:

A classifier trained against an augmented set of domains will perform better than a classifier trained without generated examples on the same test set.

Datasets for Classifier & GAN

- **Benign Data Sources**
  - Alexa Top 1 Million (top 25K)
  - Umbrella Top 1 Million (top 25K)
  - OpenDNS Random 10K
- **Phishing Data Sources**
  - OpenPhish Threat Feed
  - PhishTank Threat Feed
  - DNS-BH Threat Feed (phishing domains only)

	No. Training	No. Testing	Total
Benign	51,951	12,988	64,939
Phishing	42,834	10,709	53,543
Total	94,785	23,697	

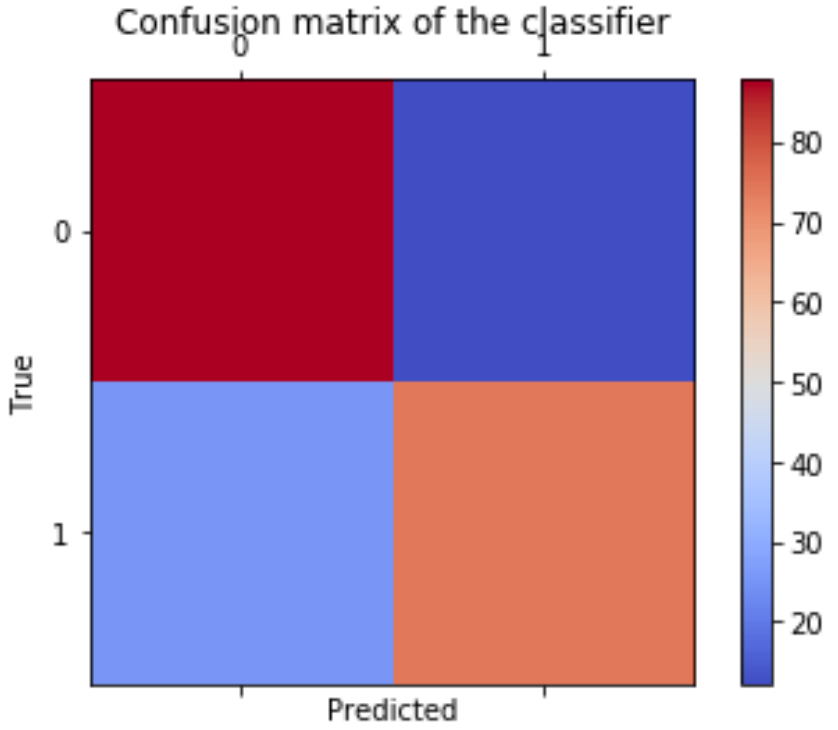
Initial Phishing Domain Classifier

Features Used
Length of domain
Number of subdomains
Has ‘-’ in domain
Term/stem frequency count
Ngram counts

- **Random Forest Classifier in Python 3 using sci-kit learn model**
  - 500 estimators
- **Final model score on test set: 81.74%**

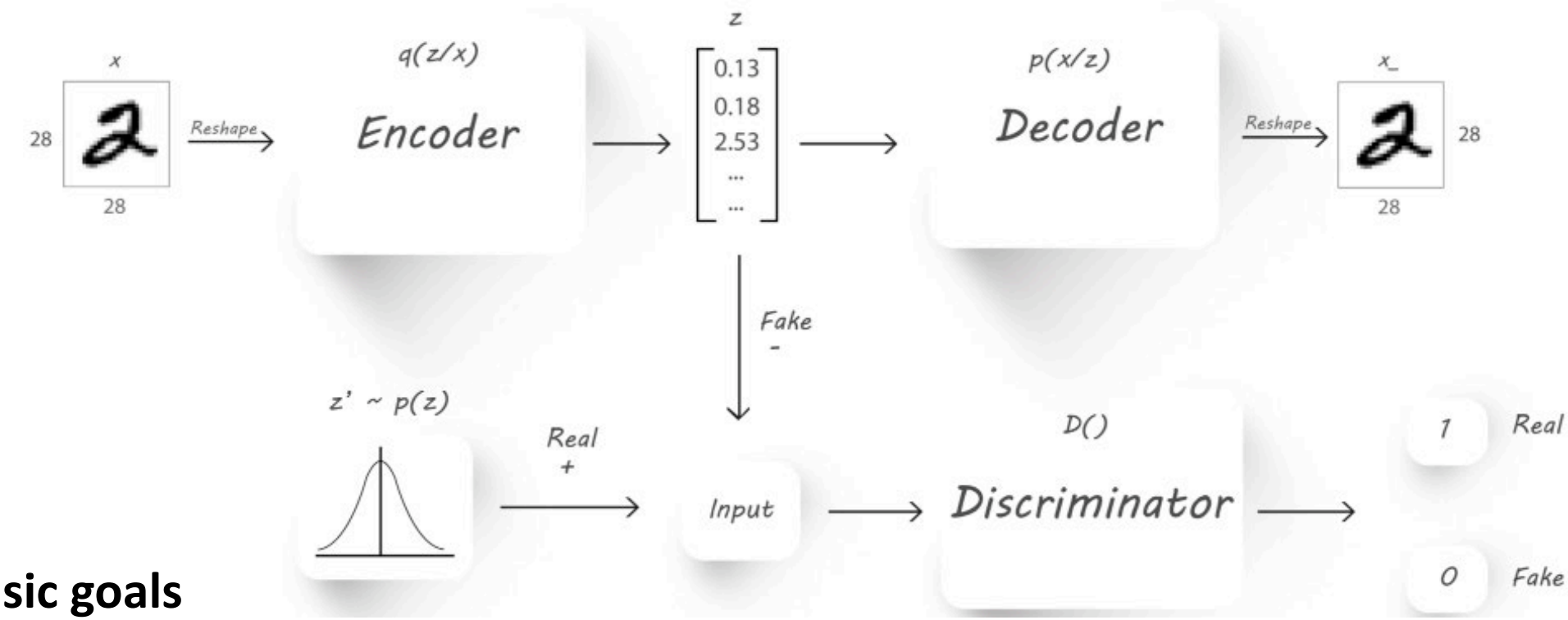
  - 0 = benign
  - 1 = phishing

```
Confusion Matrix Stats
0/0: 87.77% (11399/12988)
0/1: 12.23% (1589/12988)
1/0: 25.58% (2739/10709)
1/1: 74.42% (7970/10709)
```



MITRE

GAN Choice: Adversarial Autoencoder (AAE)



Two basic goals

1. Encoder, through the discriminator, learns to create representations that resemble random samples from a target distribution
2. Decoder can "recover" original input from any sample from the target distribution and also functions as a **stand-alone generator**

https://towardsdatascience.com/a-wizards-guide-to-adversarial-autoencoders-part-2-exploring-latent-space-with-adversarial-2d53a6f8a4f9

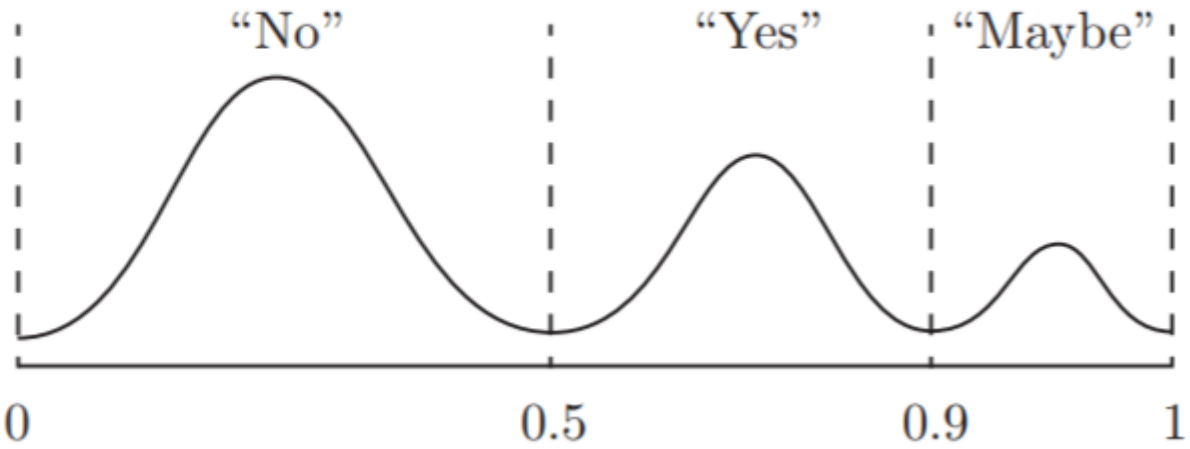
Converting Domains to Vectors

- **Discard domains with > 50 characters**
  - Too challenging for the encoder/decoder
- **Fix domains to length 50 (pad with spaces where needed)**
- **Use categorical transformation similar to one described in [1] to construct numeric vector**

Categorical

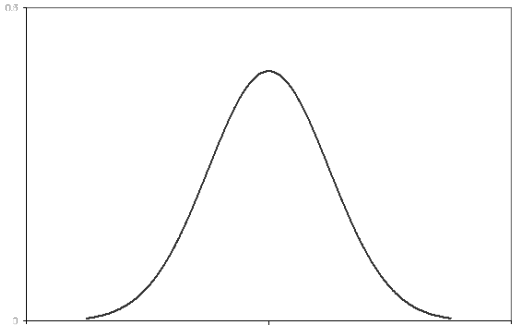
x	Pr (x)
“Yes”	0.4
“No”	0.5
“Maybe”	0.1

Numerical



1: https://dai.lids.mit.edu/wp-content/uploads/2018/03/SDV.pdf

Decoder as Stand-Alone Generator



Normal distribution

$\mu = 0, \sigma = 1$

sample

z’

1.2
0.3
…
0.7

Decoder

www.haiopiia-opdlmnrn.tmre

Generated domain

Example Generated Domains	Real Phishing Domains
cosecsoettcs.c	www.qapiseq.ml
www.i-as2oa.oo	olusesanekisola.com
www.pormcuoios.seucc.ccr	jisukcho.com
fooanbsocser.ca	www.jualumni-bd.org
cacnj-bcmectmete.mom	zyhber.com

Test Set Results: Original & Augmented Classifiers

- **4 models tested**
  - Original classifier
  - 50K classifier
    - Training set augmented with synthetic domains
  - 100K classifier
    - Training set augmented with synthetic domains
  - New malicious classifier
    - Training set augmented with more real phishing domains

Test Set	Model	Accuracy
Original	Original	81.74%
	50K	81.81%
	100K	81.78%
	NewMal	81.71%
50K	Original	80.82%
	50K	83.26%
	100K	83.3%
	NewMal	81.82%

Test Set	Model	Accuracy
100K	Original	80.84%
	50K	83.28%
	100K	83.29%
	NewMal	81.86%
NewMal	Original	81.39%
	50K	81.56%
	100K	81.54%
	NewMal	81.91%

- **Original classifier is outperformed on all test sets**

Real World Testing Results

Model	% of Domains flagged with probabilities above given threshold					
	.5	.6	.7	.8	.9	0.95
Original	51.93571	39.3	33.82857	15.15714	7.814286	4.771429
50K	43.16429	34.22857	27.49286	10.34286	5.842857	4.564286
100K	50.84286	41.45	21.27857	14.94286	6.1	2.671429
New malicious	38.43571	31.87857	23	10.65714	5.235714	3.857143

- **Each model tested on over 10,000 domains (no ground truth labels)**
  - Not necessarily all the same domains, but likely to overlap
- **Original model shows highest potential for false positives**
  - Consistently flags the highest number of domains
- **100K and new malicious model show good potential for operational use**

Conclusions & Future Work

- **Augmented models appear to outperform original model**
  - Updated threat feeds and/or more diverse training data may result in a better model
  - Additional testing/metrics/evaluation to be done before concrete conclusions
- **GANs show promise as a means of acquiring additional training & testing data for the purpose of building a more robust classifier**
  - Many ways to extend & improve work
    - Continue development on original model
    - Experiment with alternate GAN architectures & encoding methods
    - Hand-select data used to train GAN
    - Introduce fitness function as additional way to measure quality of output of GAN

MITRE

MITRE is a not-for-profit organization whose sole focus is to operate federally funded research and development centers, or FFRDCs. Independent and objective, we take on some of our nation's—and the world's—most critical challenges and provide innovative, practical solutions.

Learn and share more about MITRE, FFRDCs, and our unique value at www.mitre.org