Adversarial XAI methods in Cybersecurity

Aditya Kuppa and Nhien-An Le-Khac
UCD

Why *

Question 1. The ability to construct a coherent and complete "story" with the facts of a situation is the most important task when making a decision or recommendation.

Agree 93% Disagree 7%

Question 2. As a forecasting/recommendation task becomes more complex and difficult, I tend to rely more on judgment and less on formal, quantitative analysis.

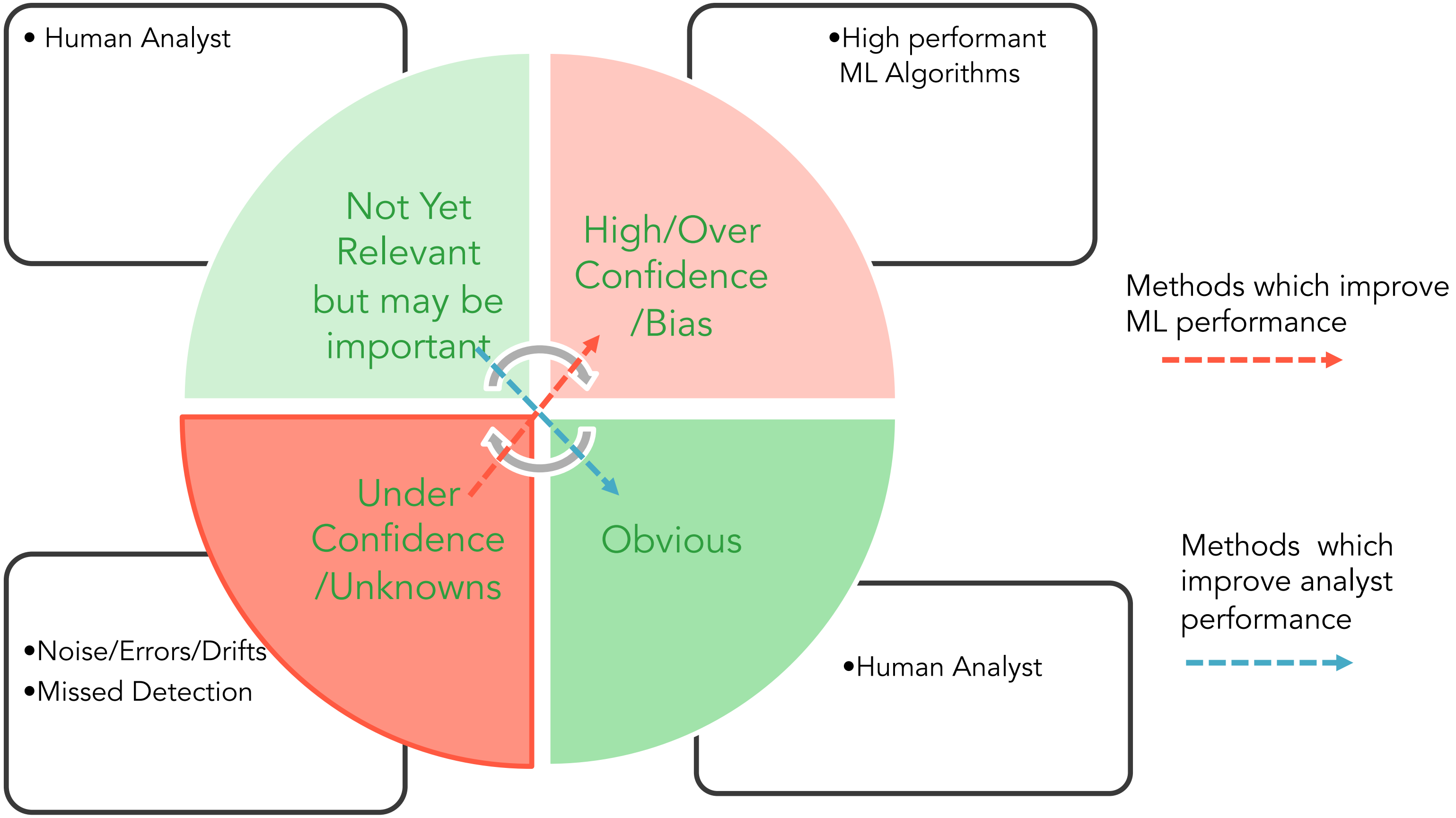
Agree 64% Disagree 36%

Question 8. As I become more uncertain about my ability to predict outcomes, I give greater weight to negative information about alternatives.

Agree 86% Disagree 14%

* Behaviour Response of CFA Charter Holders (Olsen, Robert A. "Professional investors as naturalistic decision makers: Evidence and market implications." *The Journal of Psychology and Financial Markets* 3.3 (2002): 161-167.)

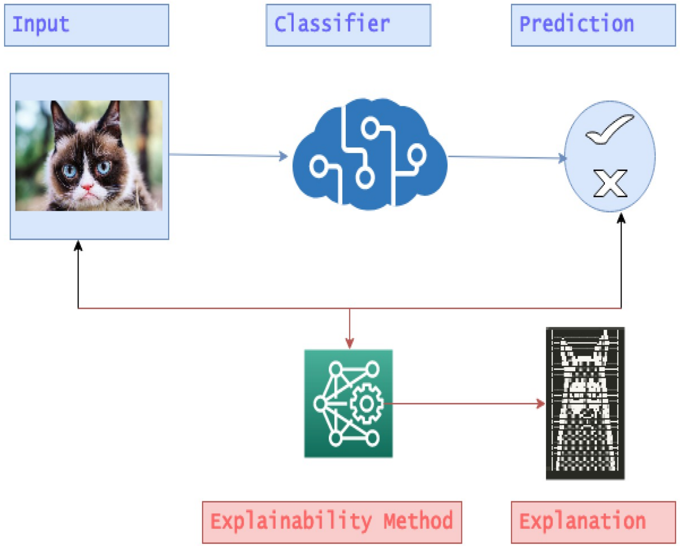
Why ?*



• Human Analyst
•High performant ML Algorithms
Not Yet Relevant but may be important
High/Over Confidence /Bias
Under Confidence /Unknowns
Obvious
•Noise/Errors/Drifts
•Missed Detection
•Human Analyst
Methods which improve ML performance
Methods which improve analyst performance

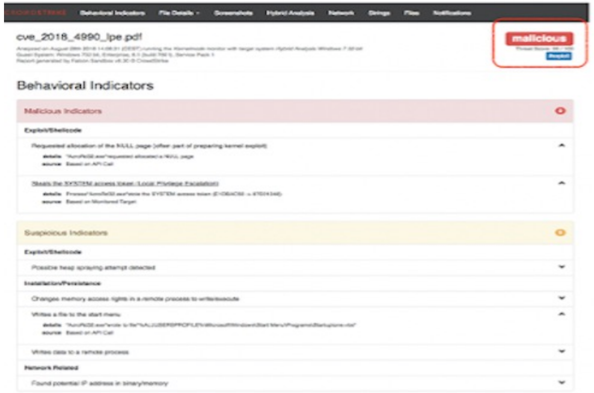
Explanations

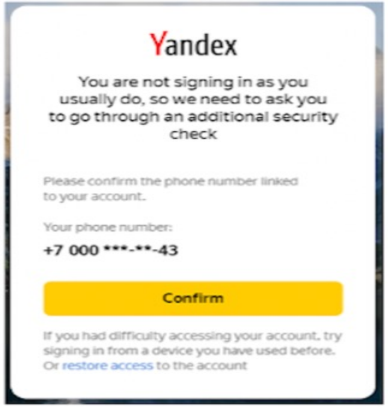
- The field of explanations of intelligent systems was active in the 1970s for expert systems; 1980's for neural networks; and then to recommendation systems in the 2000s.
- Explainability methods
  - Post-hoc/During/Pre-hoc
  - Scope - Local, Global
  - Dependency - Model, Data and Domain
- Interpretability methods coupled with the human in the loop improves the trust and security in the decision making process of ML systems.

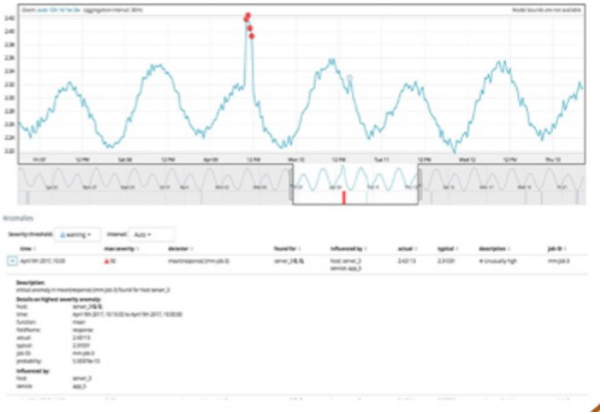


Explanations in Security Domain

- Problems in Security Domain
  - Imbalanced DataSets
  - Attribution of threat and Context is important and hard to infer.
  - Threats are always evolving and there is a need to improve robustness of underlying systems.

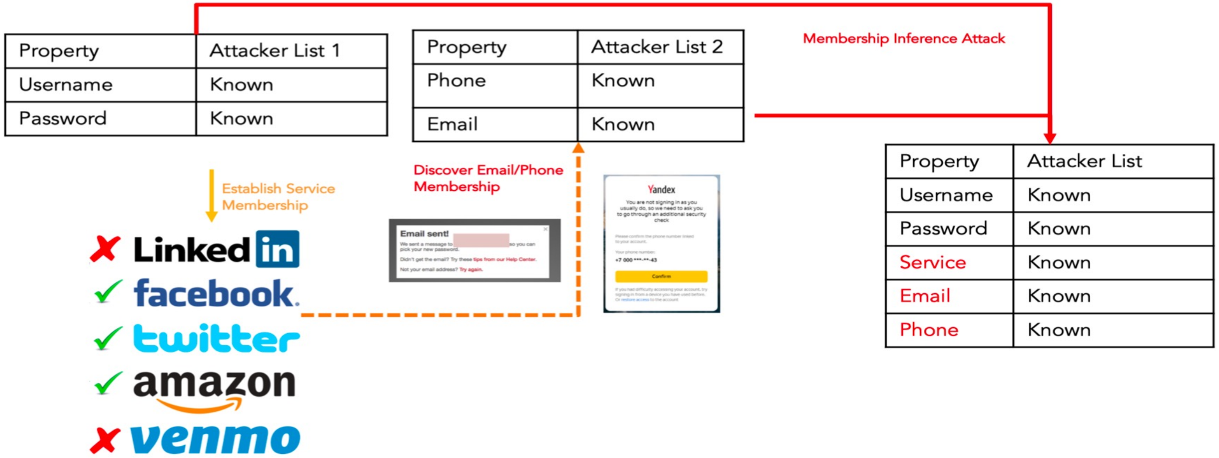






Goals and Motivation

- Security analysis of XAI methods
  - How can an attacker, given only outputs of explanation method and model predictions, can conduct powerful black-box model extraction, membership inference attacks?
  - How explanation outputs facilitate the generation of adversarial samples and poison/backdoor samples to evade the underlying classifier
- Motivating Example – Credential Stuffing (Membership Inference attack)



Threat Model Assumptions

Characteristic	Type	MEA [26]	MIA [68]	PA [67]	AE [45]
Knowledge	Training Distribution	✗	✗	✓	✗
	Feature set	✓	✓	✓	✓
	Feature extractor	✓	✓	✓	✓
	Feature transformers	✓	✓	✓	✓
	Inference API	✓	✓	✓	✓
	Explanations interface/Method	✓	✓	✓	✓
	Confidence intervals	✓	✓	✓	✓
Goal/Intent	Compromising integrity (Evasion)	✗	✗	✓	✓
	Compromising privacy	✓	✓	✗	✗
Capability	Manipulate training data	✗	✗	✓	✗
	Manipulate test data	✗	✓	✗	✓
Strategy	Train a surrogate model for parameter extraction	✗	✓	✗	✗
	Train a surrogate model for query reduction	✓	✗	✗	✓
	Satisfy domain constraints	✗	✓	✓	✓
Frequency	Iterative	✓	✓	✓	✓
Perturbation Scope	Instance Specific	✓	✓	✓	✓
Perturbation Constraints	Optimisation	✓	✗	✓	✗
	Domain	✓	✓	✓	✓

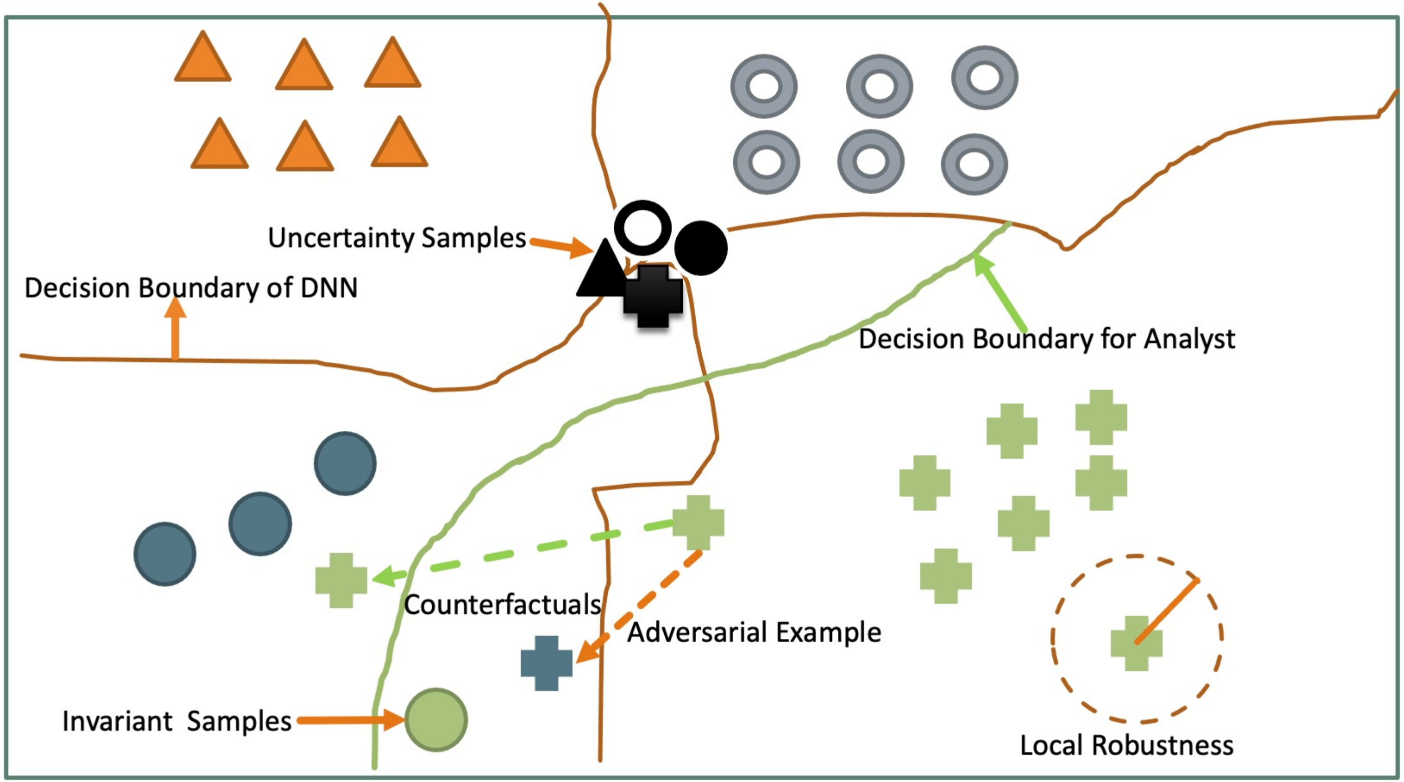
Counterfactuals (why/why-not)

- Counterfactual data instances of the input have
  - Similar feature values as input
  - Different model predictions from that of input
  - Lay closer to the decision boundary of an input class

$$x_{cf} = \underset{x_{cf\,1},\ldots,x_{cf\,k}}{\operatorname{argmin}} \; \mathcal{L}(\mathcal{T}(x_{cf}), \mathcal{T}(x)) + Dist(x_{cf} - x)$$

Method	$\mathcal{L}$	$Dist$	**CF per** x	**Optimisation Method**
Latent CF [38]	Latent Vector Loss	ℓ_1	1	Gradient Descent
DICE [39]	Hinge-loss	ℓ_1 and Median Absolute Deviation(MAD)	k	Gradient Descent
Permute Attack [40]	-	l_2	1	Genetic Algorithm

Class Decision Boundary (AE vs CF)



Attack Method

- Given a black-box access to a target model T prediction interface $T(x) = y$, x_{cf} counter factual , $E(x) = x_{cf}$ explanation interface, D_{aux} auxiliary dataset and S a surrogate model
  - Attacker aims to compromise the confidentiality and integrity of the underlying ML system

- Explanation-based Poisoning Attack
  - Identify and Perturb robust features, which are consistently same across their counterfactual class.
- Explanation-based Adversarial Sample Generation
  - Adapt Counterfactual method which works in feature space to sample space.
- Explanation-based Membership Inference Attack
  - 1-Class Nearest neighbor classifier for each class is trained on counterfactuals to establish membership
- Explanation-based model extraction
  - Knowledge distillation technique to transfer knowledge from the target model to the surrogate model

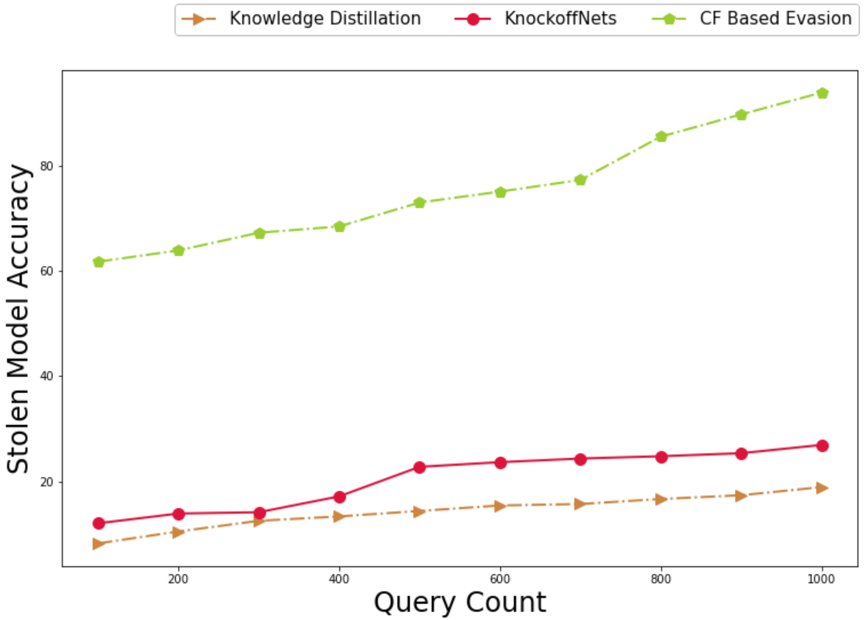
DataSets

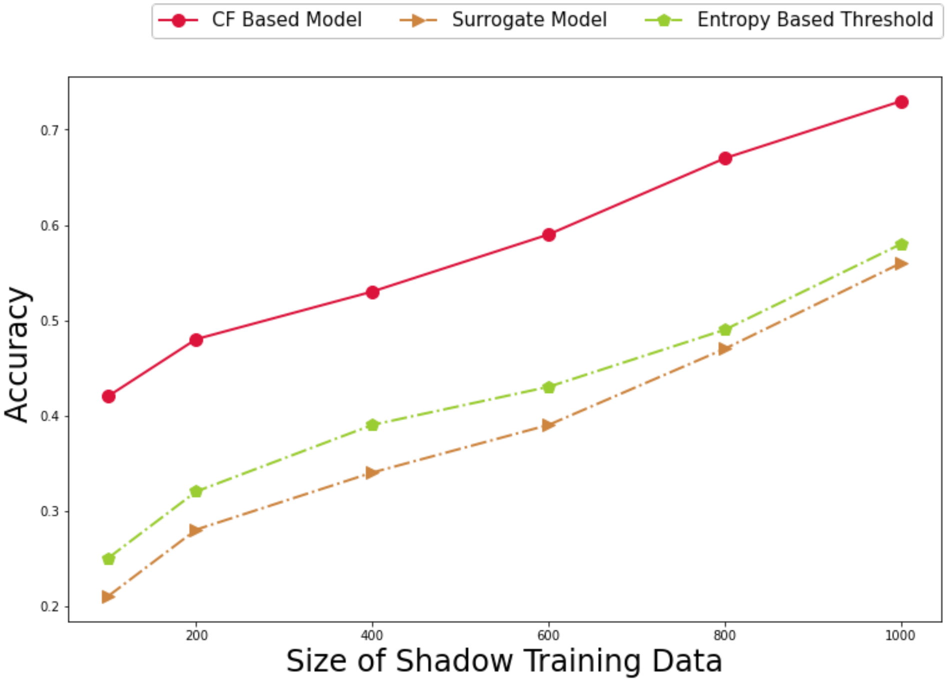
- MEA
  - CICIDS17 Network Traffic dataset which contains a wide range of attack types like SSH brute force, Botnet, DoS, DDoS, web, and infiltration
- AE and Poisoning Attack
  - 30120 malware from virus share and for benign samples we scrapped 20334 clean files from free ware sites
- MIA
  - Leaked Password Dataset -- The dataset consists of 1.4 billion email password pairs with 1.1 unique emails and 463 million unique passwords. This dataset is aggregated password leaks from different incidents.

Results

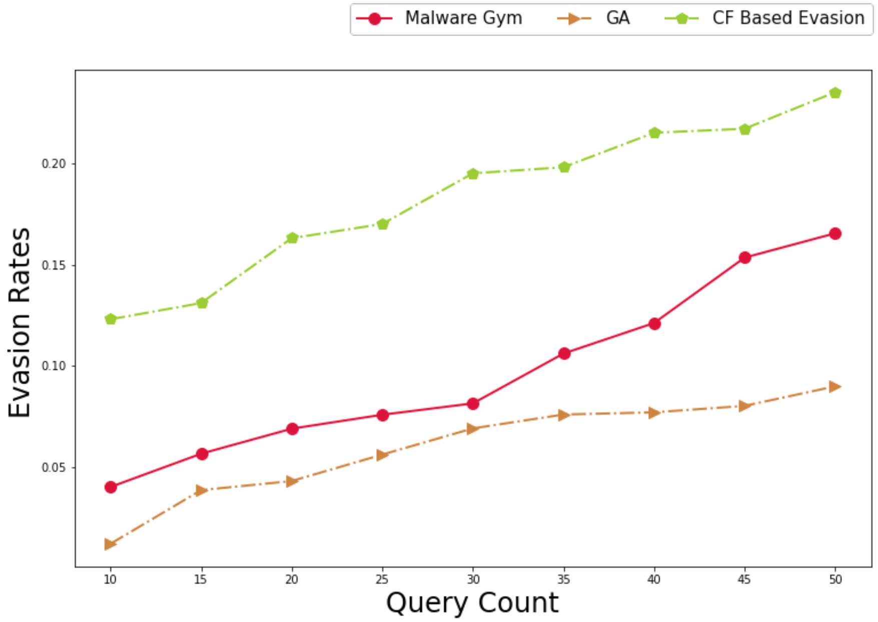
Attack Type	$\mathcal{D}_{aux}$	Explanation Method	$\mathcal{T}$	Original Accuracy	Evasion Accuracy
Adversarial Attack	Malware	Permute	AV1	93.5%	65.23%
			AV2	94.7%	41.89%
				Poisoning Percent	**Accuracy Drop**
Poisoning Attack	Malware	Permute	GBM	0.5%	62.4%
				1%	76.23%
				2%	87.24%
			NN	0.5%	30.9%
				1%	50.89%
				2%	65.31%
				3%	79.48%
				Method	**Accuracy/Queries**
Membership Inference	Leaked Passwords	Latent-CF	AutoEncoder	Model	49.46/1000
				Entropy	54.17/1000
				CF	73.17 /1000
				Model	**Accuracy**
Model Extraction	CICIDS	DICE	AutoEncoder	$\mathcal{T}$	98.02
				KN	78.91
				KD	53.89
				CF	93.54

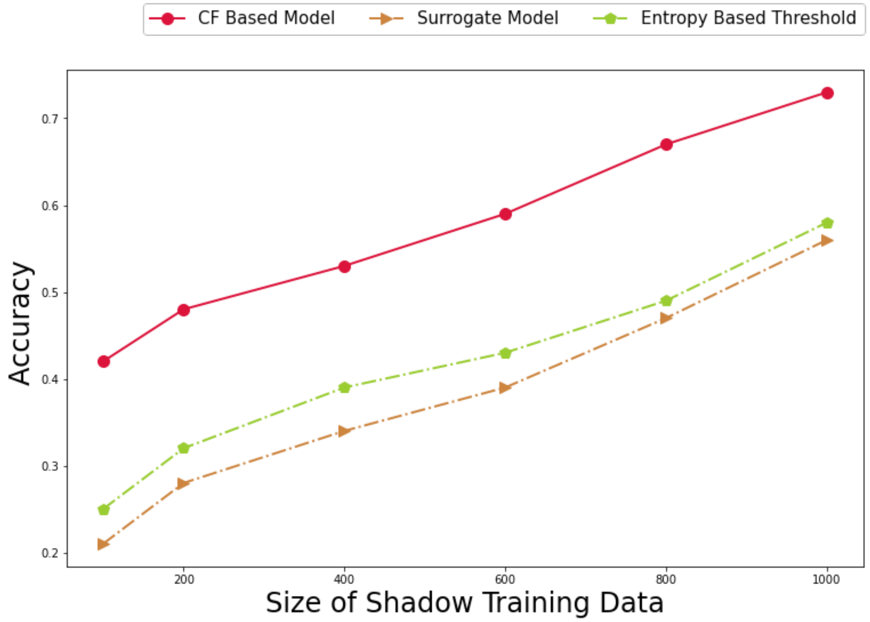
MIA and MEA Comparison





AE and Poisoning Attack Comparison





Defense Discussion

- CF methods share a large set of similarities with adversarial examples concepts
  - Adapting methods from adversarial defense literature
- *Noise* based Defense Intuition
  - *D*efender has no control over the attacker's *full* training data but only a portion of it.
  - ML model aims to learn the mapping function from the feature space to the label space from the training samples.
- Defender can transform the counterfactual samples so the learned model (surrogate) has a strong correlation between the labels and noise of the feature space instead of only features.
- Adding noise to samples
  - Individual CF sample and All CF samples of same class

$$T_s = \begin{cases} None & \text{No transformation} \\ \phi, & \text{Random noise } [-1, 1] \\ \delta_{x_i}, & \text{Adv. noise } [-\epsilon, \epsilon] \\ \delta_{y_i} & \text{Adv. noise } [-\epsilon, \epsilon] \end{cases}$$

T_s	Accuracy
None	95.6
ϕ	87.4
δ_{x_i}	37.4
δ_{y_i}	28.22
$\delta_{x_i} + \phi$	32.22
$\delta_{y_i} + \phi$	18.22

Limitations and Future work

- MEA
  - Methods which optimize on multiple properties of CF’s improve the stolen model accuracy
  - We only tested non-differential models
- MIA
  - Methods which do not employ latent space to search for CF need large number of queries.
  - Learning password rules and investigate how CF attack can speed up the password cracking methods
- AE and Poisoning
  - The functionality preserving transformation functions applied on the binary are biased towards static features.
  - Our results may not be valid when AV engines use both static and dynamic analysis to make a decision.
  - CF methods can help attackers to find quicker ways to find adversarial/poisoned samples,
  instead of solving a hard-to converge black-box optimization problem in input space.

Thanks

TIFS Paper -- https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=9555622