Towards Certifiable Adversarial Sample Detection

Ilia Shumailov, Yiren Zhao, Robert Mullins, Ross Anderson

BOSCH-FORSCHUNGSSTIFTUNG
im Stifterverband

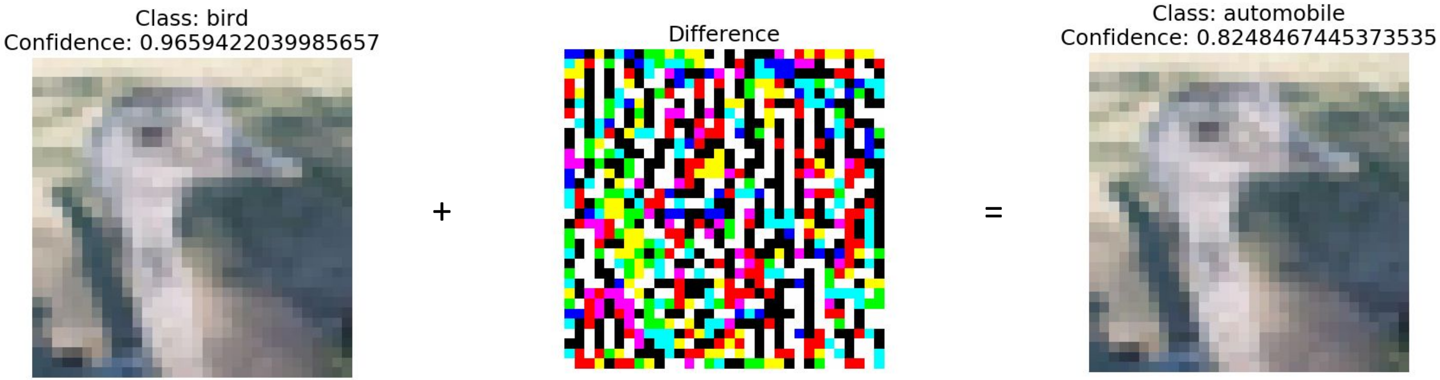
UNIVERSITY OF CAMBRIDGE

Machine Learning

- Machine learning is everywhere
- We operate on data, not formal rules
- There’s a lot of non-determinism
- It is suddenly hard to measure or even define critical emergent properties:

Safety, Security and *Robustness*

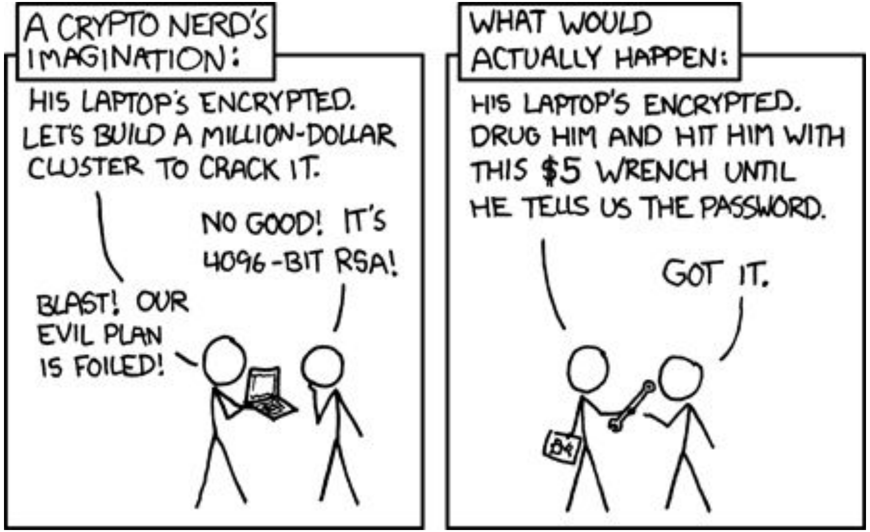
Computer Security in context of Machine Learning



- Adversarial examples exist for all models
- There’s a large taxonomy of attackers
- They operate in white-box / grey-box / black-box settings
- Attacks are scalable because of transferability

Machine Learning in the context of System Security

- ML is a part of a larger pipeline
- As secure as the weakest link
- Need: clear threat model
- Safety / security policies / cases
- Well-defined environments
- Clear policy for handling abuse
- Build from trusted components



BOSCH-FORSCHUNGSSTIFTUNG im Stifterverband

UNIVERSITY OF CAMBRIDGE

ML integrity attacks and robustness

robustness

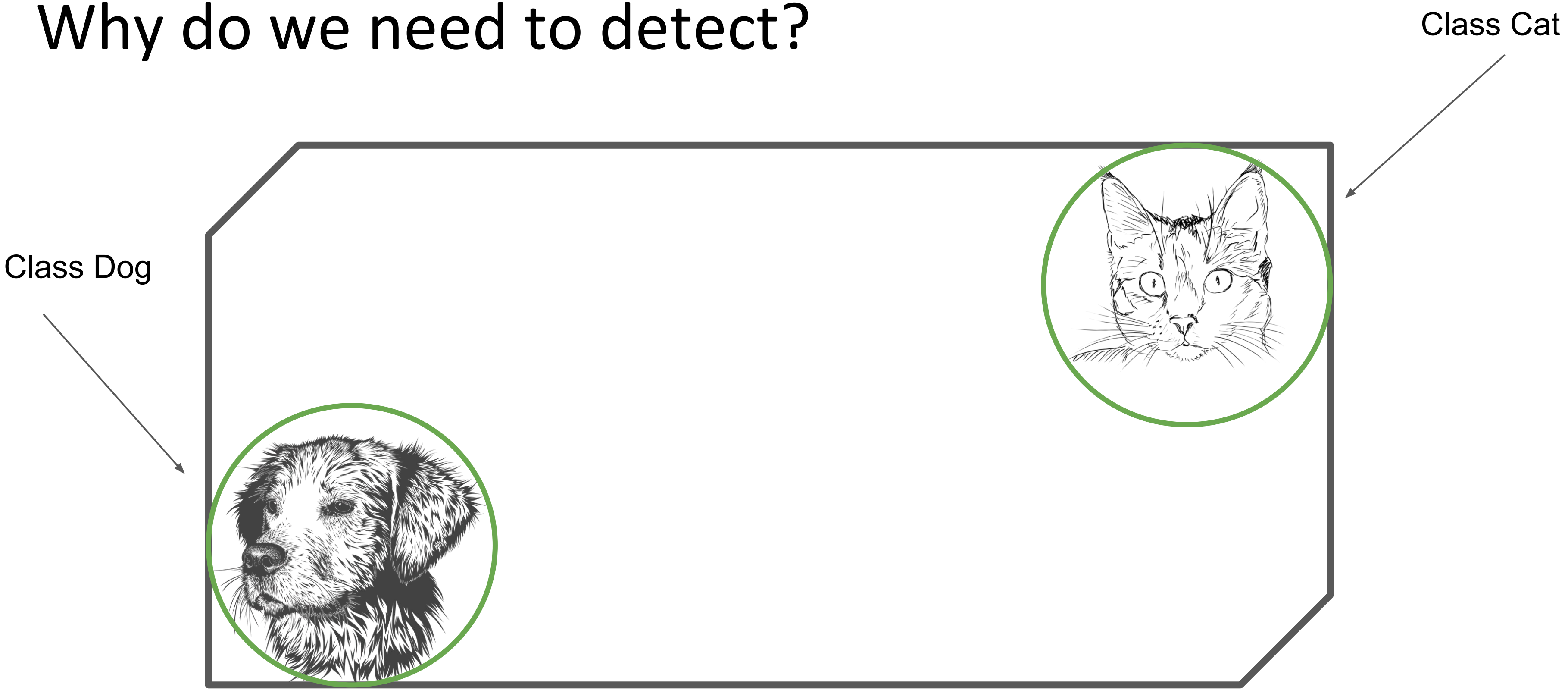
Defence via
robust optimisation

- Adversarial training
- Certifiable robustness
- Randomised smoothing

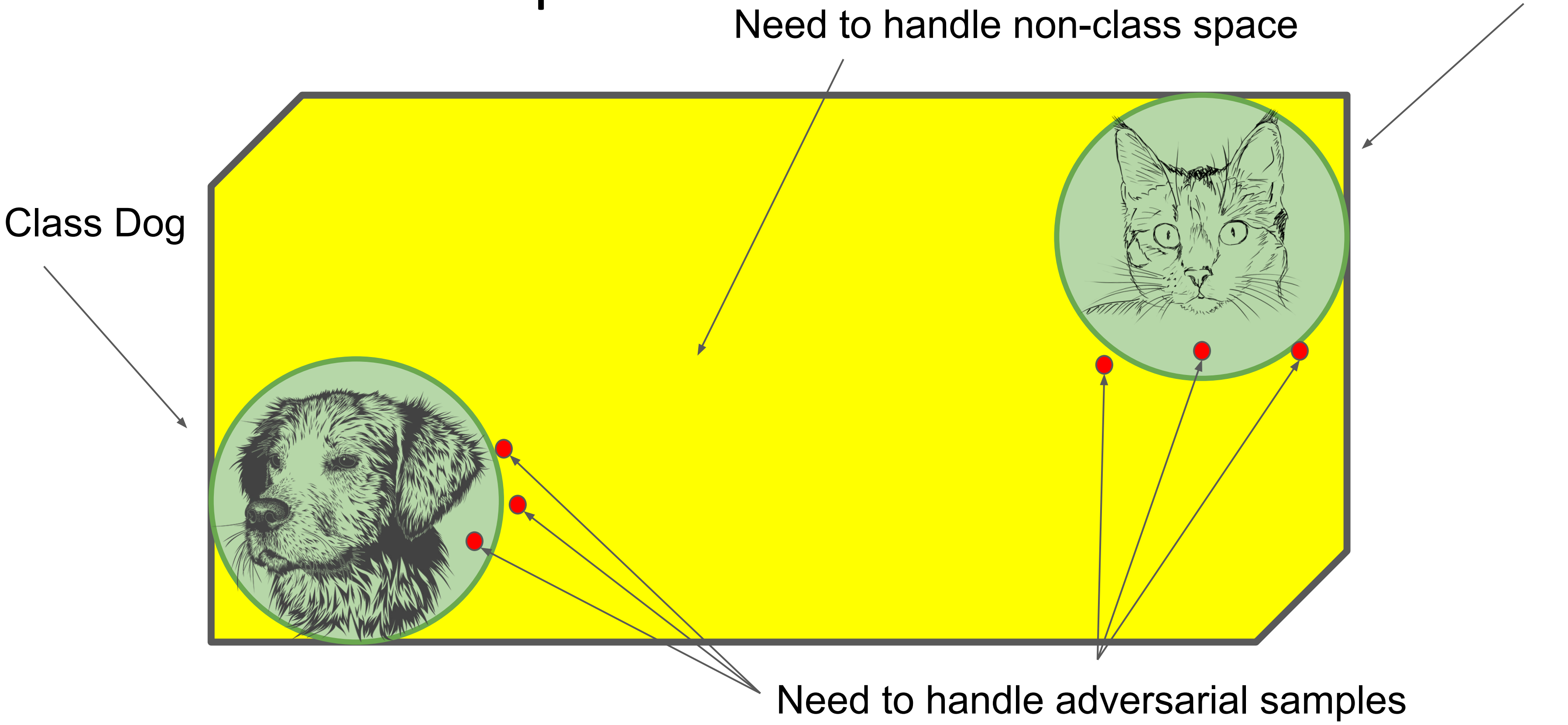
Detection aka *not dealing with certain data*

- MagNet
- Taboo Trap
- Uncertainty
- Trapdoors

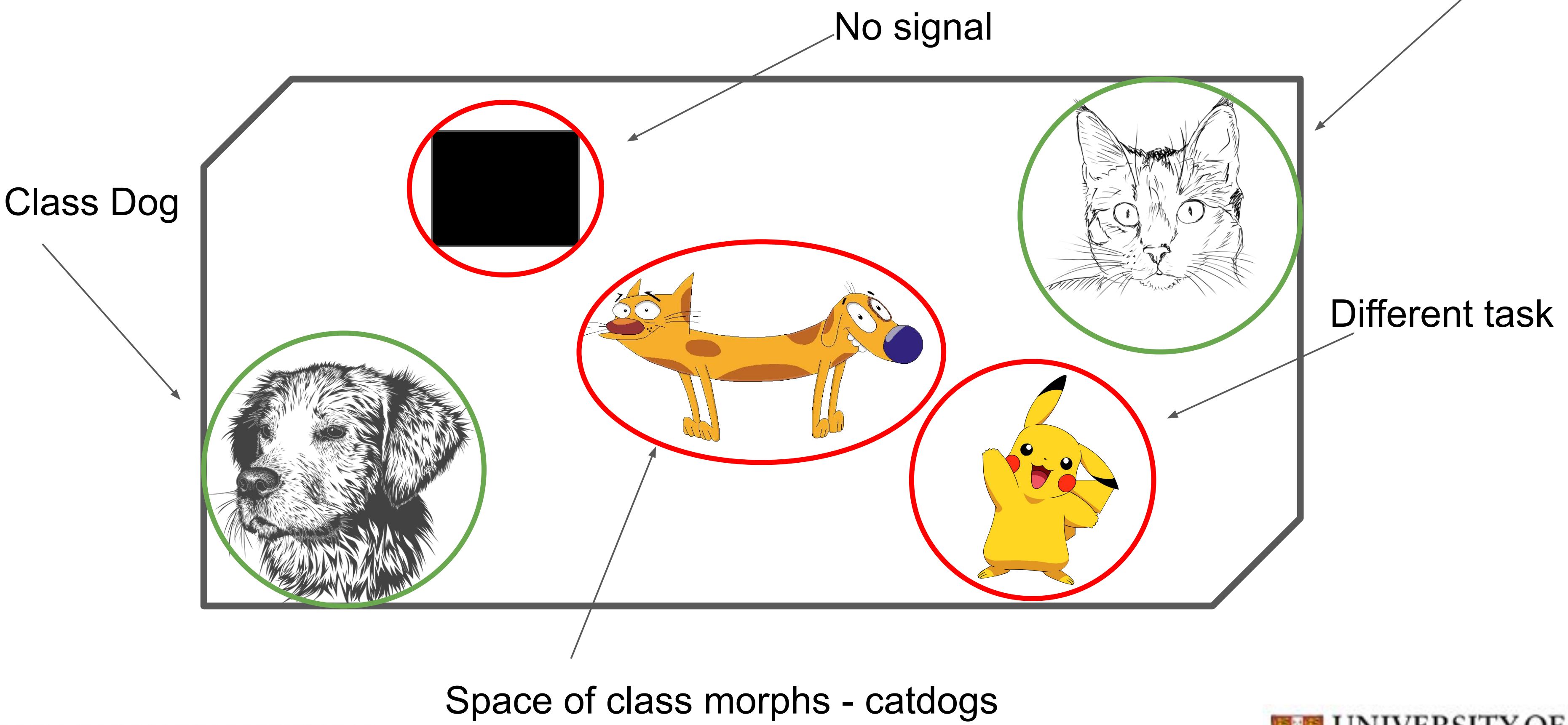
Why do we need to detect?



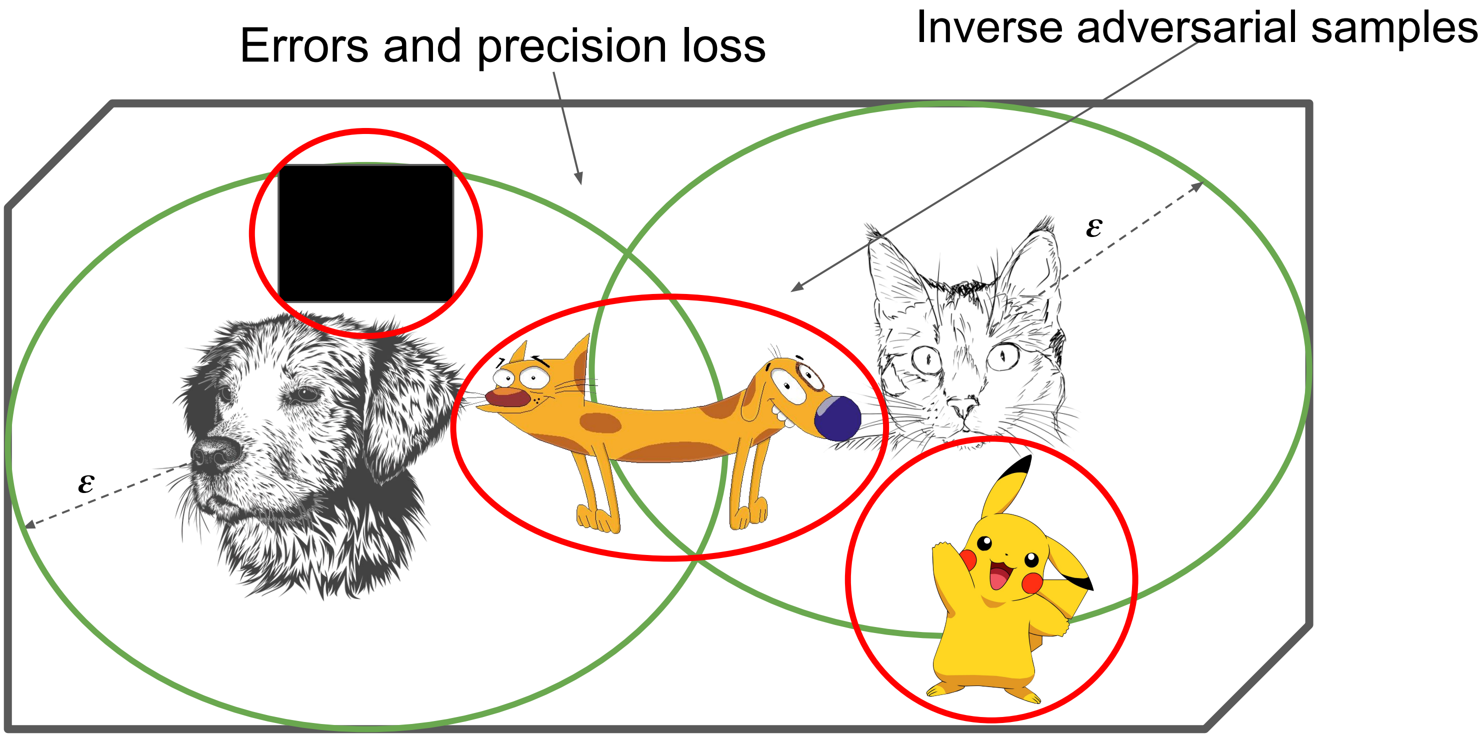
Detection can help



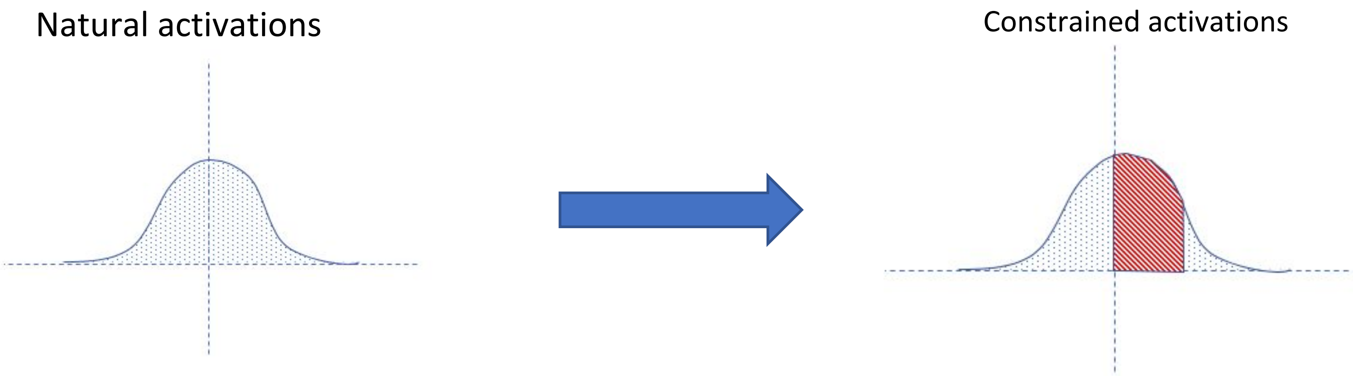
Why do we need to detect?



ε-robustness



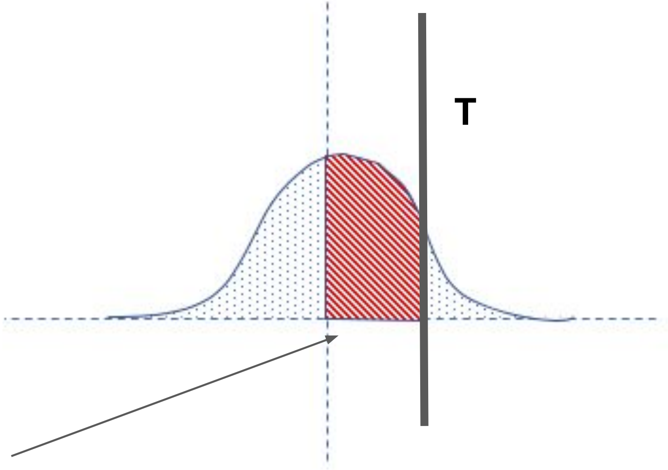
Taboo Trap



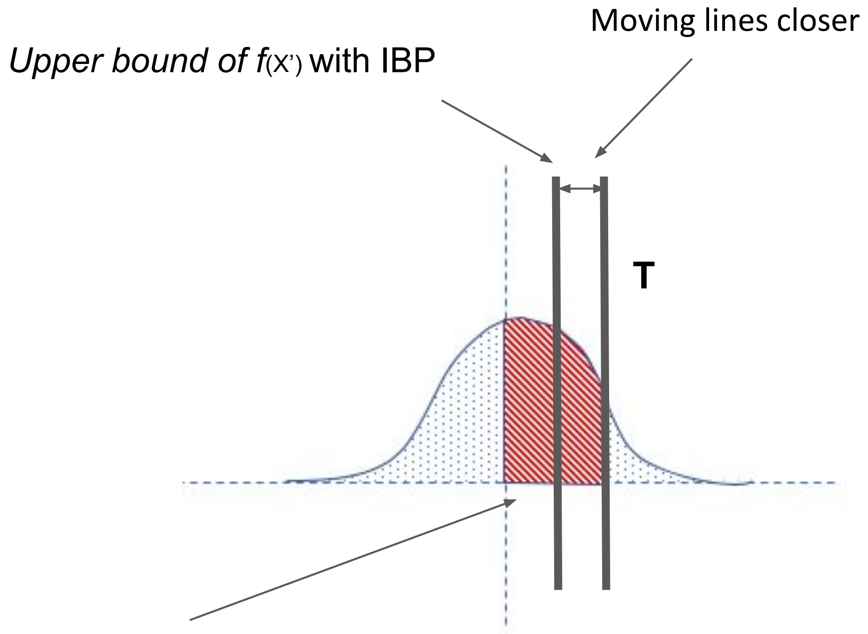
- During training, restrict the numerical range of activations
- Detect when activations are out of bounds

Can we use this to make attacks detectable?

Certifiable Taboo Trap (CTT)



For natural data **X** enforce constraints on *f* to be below **T**

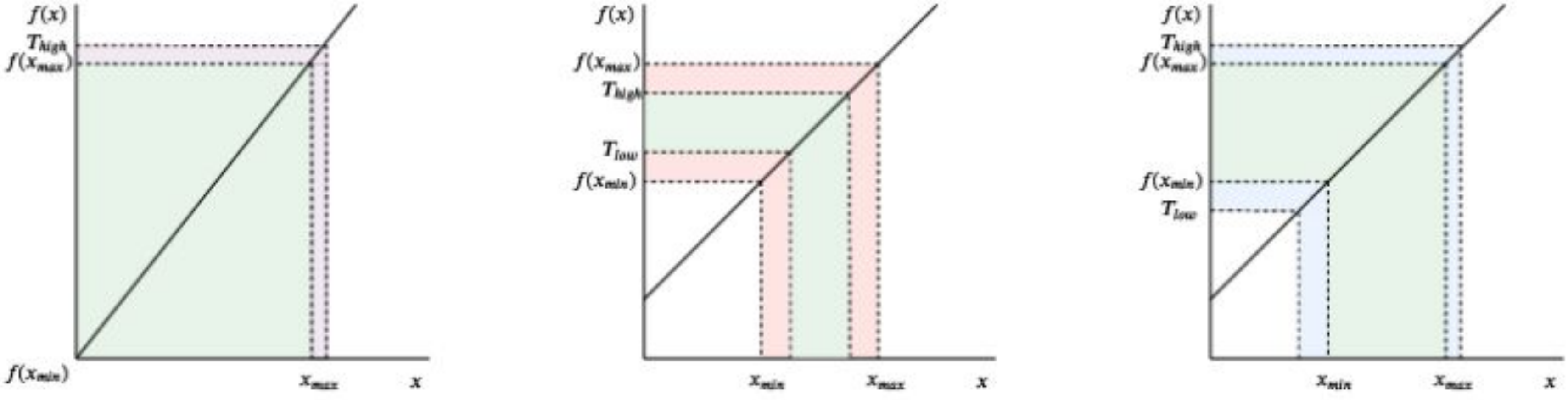


For ***ε***-ball around the data point **X' = X±*ε*** enforce that upper bound of f(**X'**) ≥ **T**

BOSCH-FORSCHUNGSSTIFTUNG
im Stifterverband

UNIVERSITY OF CAMBRIDGE

Certifiable Taboo Trap (CTT) more generally



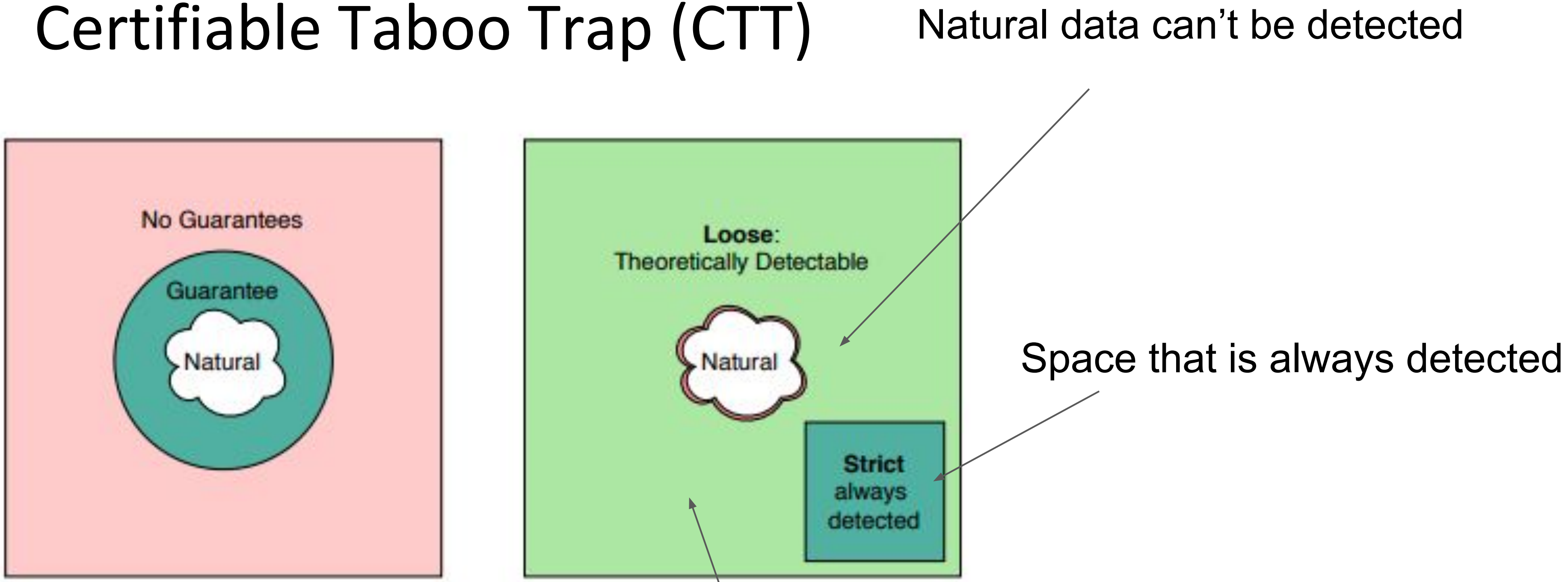
(a) Original Taboo Trap. (b) False positives (in red). (c) Undetectable range (in blue).

- Easily quantifiable space that is either *False positive* or *Undetectable*
- Allows for easy certification!

BOSCH-FORSCHUNGSSTIFTUNG
im Stifterverband

UNIVERSITY OF CAMBRIDGE

Certifiable Taboo Trap (CTT)



(a) Certifiable Robustness with IBP

(b) Certifiable Detection with CTT

Space can theoretically be detected

CTT with MNIST

		Baseline	AdvTrain	Ensemble	PCL	MagNet			CTT-lite			CTT-loose			CTT-strict		
Attack	Param	Acc	Acc	Acc	Acc	Det_{l_1}	Det_{l_2}	$Det_{l_1 \| l_2}$	Acc	Det	l_2	Acc	Det	l_2	Acc	Det	l_2
No Attack		99.1	99.5	99.5	99.3	1.75	1.93	2.93	99.1	1.9	-	98.5	1.6	-	98.9	1.1	-
FGSM	$\epsilon = 0.1$	66.7	73.0	96.3	96.5	54.49	54.59	54.80	70.9	1.4	2.08	25.0	100.0	1.98	61.1	100.0	1.99
	$\epsilon = 0.2$	25.7	52.7	52.8	77.9	85.20	85.31	85.31	21.9	1.0	4.14	15.0	100.0	3.89	32.7	100.0	3.90
BIM	$\epsilon = 0.1$	49.4	62.0	88.5	92.1	80.82	24.90	80.92	44.2	1.0	1.13	0.0	100.0	0.38	0.15	100.0	0.75
	$\epsilon = 0.15$	15.4	18.7	73.6	77.3	88.37	37.14	88.47	4.2	0.8	1.48	0.0	100.0	0.50	2.0	100.0	0.97
PGD	$\epsilon = 0.1$	59.4	62.7	82.8	93.9	83.78	77.96	83.78	51.0	1.2	1.50	1.0	100.0	1.24	13.4	100.0	1.35
	$\epsilon = 0.2$	1.83	31.9	41.0	80.2	98.27	98.27	98.27	0.0	1.1	2.73	0.0	100.0	2.43	0.9	100.0	2.53

- CTT can detect strong attackers with MNIST
- CTT outperforms other methods with comparable false positives

CTT with Cifar10

Attack	Param	Baseline Acc	AdvTrain Acc	Ensemble Acc	PCL Acc	MagNet Det_{l_1}	MagNet Det_{l_2}	MagNet $Det_{l_1 \| l_2}$	CTT-loose Acc	Det	l_2	Acc	Det	l_2	CTT-strict Acc	Det	l_2
No Attack		89.1	84.5	90.6	91.9	6.40	6.61	8.13	86.2	3.4	-	86.3	6.4	-	86.1	3.0	-
FGSM	$\epsilon = 0.02$	33.6	44.3	61.7	78.5	7.80	6.64	9.55	18.6	95.7	1.07	16.8	98.5	1.08	16.1	96.4	1.06
	$\epsilon = 0.04$	22.4	31.0	46.2	69.9	11.53	8.38	13.27	7.6	93.6	2.00	7.2	94.2	2.01	6.0	93.1	2.06
BIM	$\epsilon = 0.01$	13.5	22.6	46.6	74.5	6.98	6.52	8.61	0.5	9.0	0.15	0.0	14.1	0.16	1.1	10.9	0.16
	$\epsilon = 0.02$	1.5	7.8	31.0	57.3	6.64	6.52	8.50	0.0	14.2	0.21	0.0	25.9	0.20	0.0	17.2	0.21
PGD	$\epsilon = 0.01$	24.0	24.3	48.4	75.7	7.10	6.52	8.73	0.1	10.4	0.34	2.9	24.3	0.34	2.0	16.6	0.34
	$\epsilon = 0.02$	2.9	7.8	30.4	48.5	6.98	6.52	8.85	0.0	40.8	0.65	0.0	70.3	0.65	0.0	49.9	0.65

- CTT can detect some strong attackers with Cifar10
- CTT outperforms some other methods with comparable false positives

Towards more usable detection schemes

- Lesson from system security: **every system breaks**
- Manipulation must be expected and detected
- Recovery should be easy
- Diversity is paramount
- Detection and defence mechanisms can and should be used together
- Robust situational awareness is the missing link

Towards more usable detection schemes

- CTT can use **different keys** by using different neurons detection
    - If one model is compromised others are not affected
- CTT is simple and fast
    - It can run on **any hardware** that can run the network
- CTT can be used to enforce **strict detection** of specific data regions

Thank you very much for listening!

Please do not hesitate to reach out in case there are any questions at

ilia.shumailov@cl.cam.ac.uk