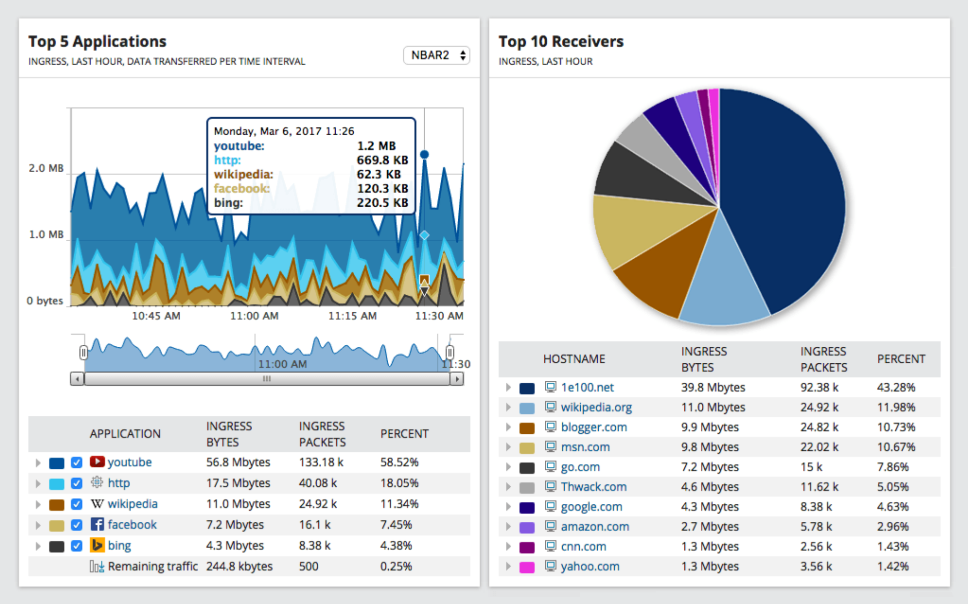
BARS: Local Robustness Certification for Deep Learning based Traffic Analysis Systems

Kai Wang, Zhiliang Wang, Dongqi Han, Wenqi Chen,

Jiahai Yang, Xingang Shi, Xia Yin

清華大學
Tsinghua University

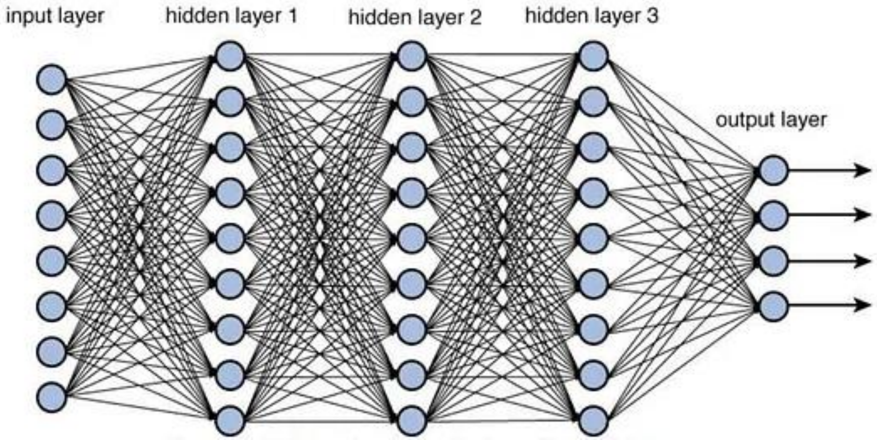
Traffic Analysis Meeting Deep Learning



Source: https://www.solarwinds.com/netflow-traffic-analyzer/use-cases/network-traffic-analysis

Traffic is an important data source for analyzing network activities and detecting cyberspace attack.

+



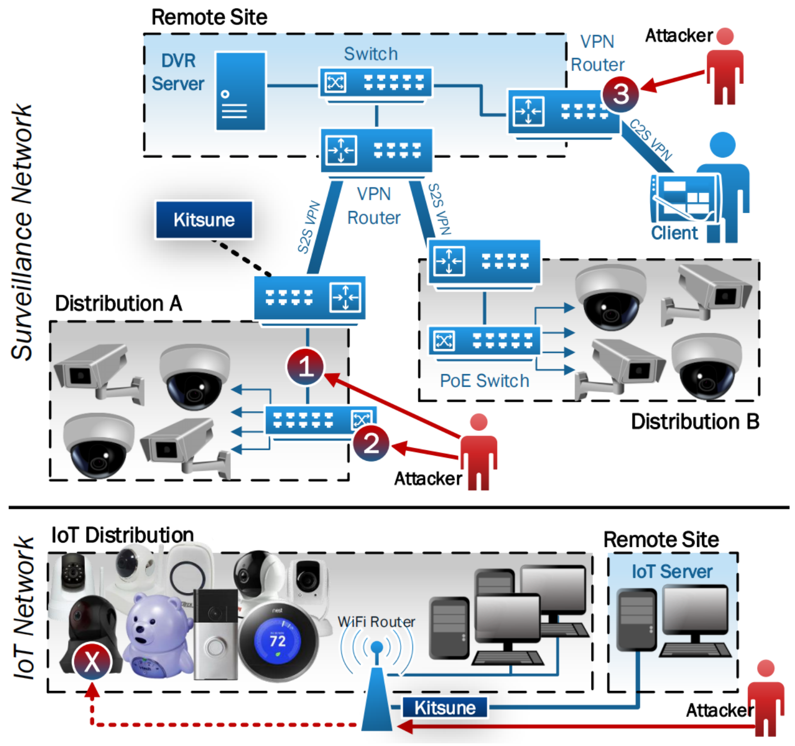
Source: https://towardsdatascience.com/training-deep-neural-networks-9fdb1964b964

Deep learning has been widely applied for data analysis.

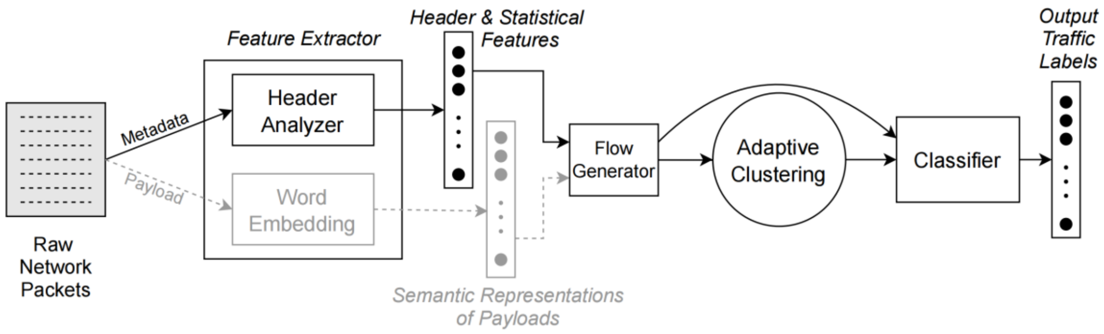
Can they combine?

DL-based Traffic Analysis Systems

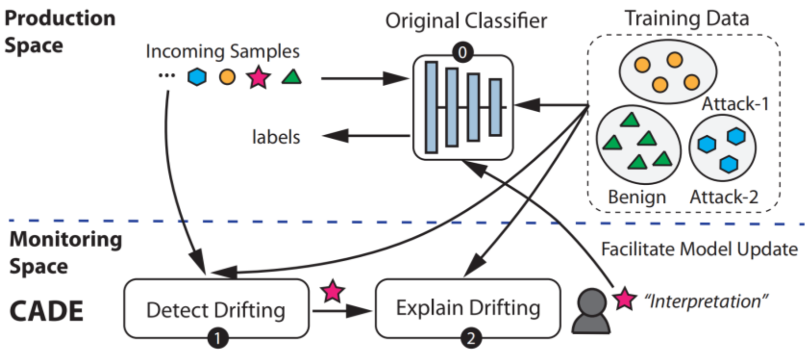
- Zero-positive NIDS (NDSS’18, CCS’19)
- Concept drift detection system (USENIX Security’21)
- Supervised multi-classification system (INFOCOM’21, CCS’18)



Kitsune (NDSS’18)



ACID (INFOCOM’21)



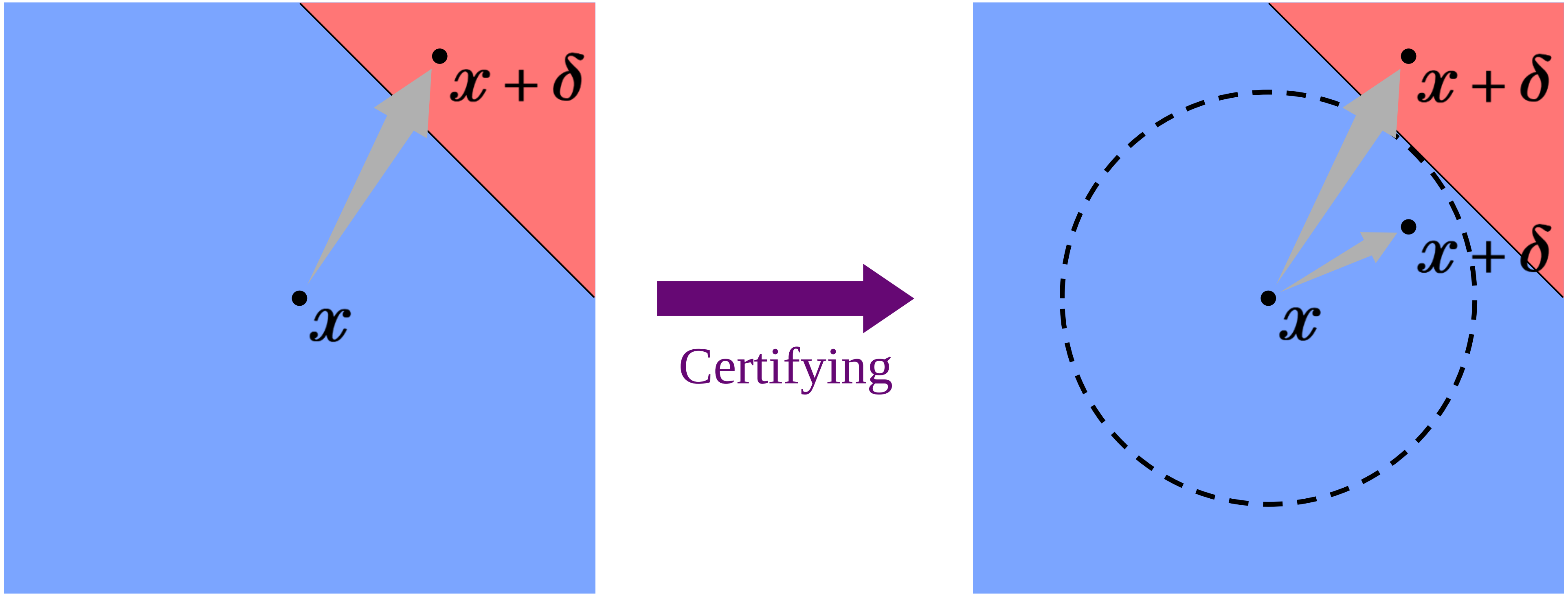
CADE (USENIX Security’21)

DL-based Traffic Analysis Systems

How does DL-based traffic analysis systems perform in practice?

They frequently suffer from adversarial attack due to the vulnerability of deep learning.

Adversarial Attack Meeting Robustness Certification



Vanilla randomized smoothing (ICML'19)

Adversarial Attack Meeting Robustness Certification

Can you give me a suitable robustness certification framework for DL-based traffic analysis systems?

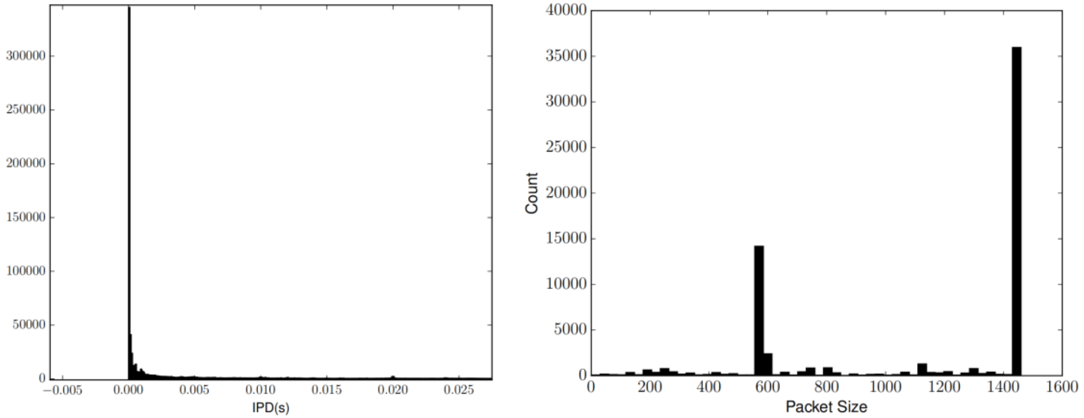
Unfortunately, existing robustness certification frameworks are not suitable for traffic analysis. We need to design a special one under the following three motivations.

Motivation I

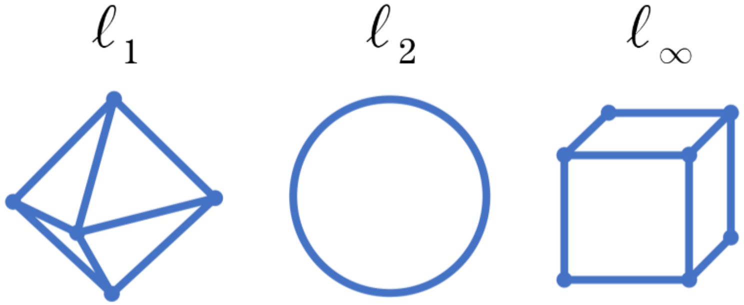
Traffic Analysis Meeting Existing Certification Methods

Highly heterogeneous features
(CCS'17, ICISSP'18)

ℓ_p robustness guarantee
(ICLR'21, ICLR'19, ICML'19, ICML'20)



Traffic features (CCS'17)



Norm ball (ICML'20)

We need dimension-heterogenous certification!

Motivation II

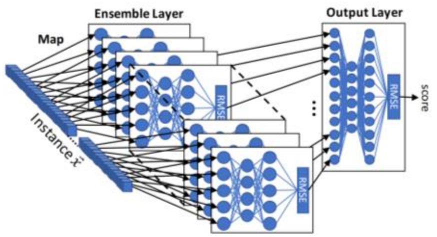
Traffic Analysis — Meeting — Existing Certification Methods

Varied model designs

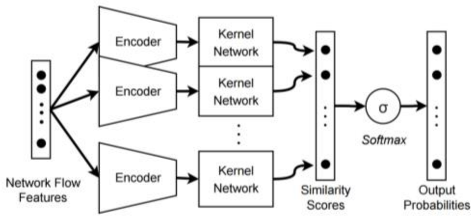
(NDSS’18, USENIX Security’21, INFOCOM’21)

Needing special designs

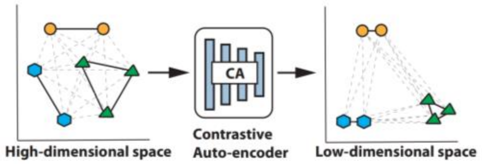
(ICLR’19, NeurIPS’20, NeurIPS’18)



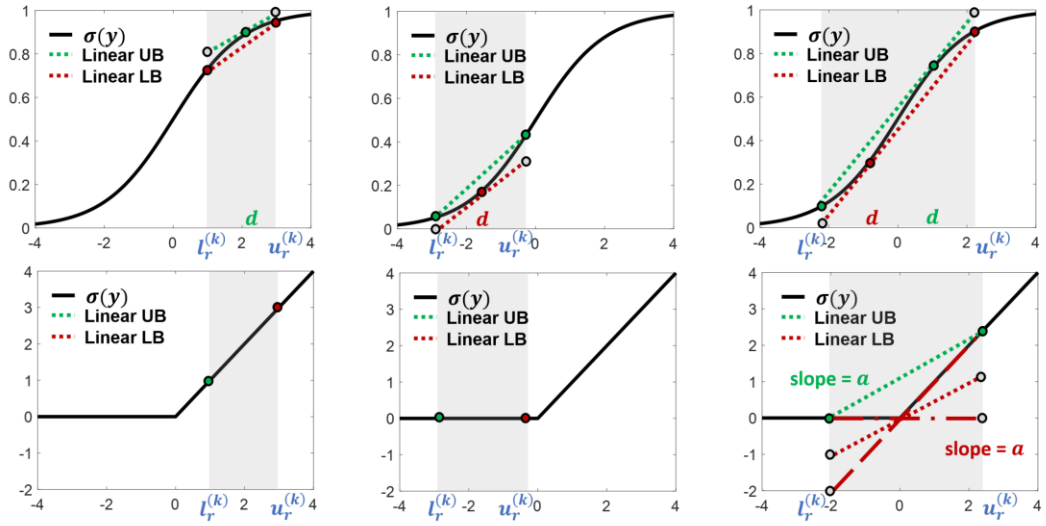
Kitsune (NDSS’18)



ACID (INFOCOM’21)



CADE (USENIX Security’21)



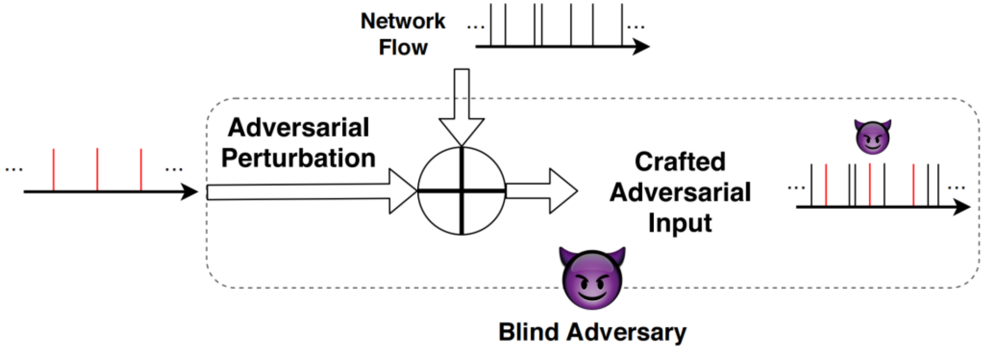
Special linear relaxation (NeurIPS’18)

We need universal certification!

Motivation III

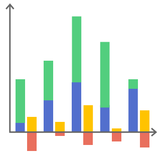
Traffic Analysis **Meeting** Existing Certification Methods

Adversarial operating environments
(USENIX Security'21, INFOCOM'20)



Blind adversarial perturbations
(USENIX Security'21)

No real-time certification
(CCS'21, ICLR'21, NeurIPS'21)

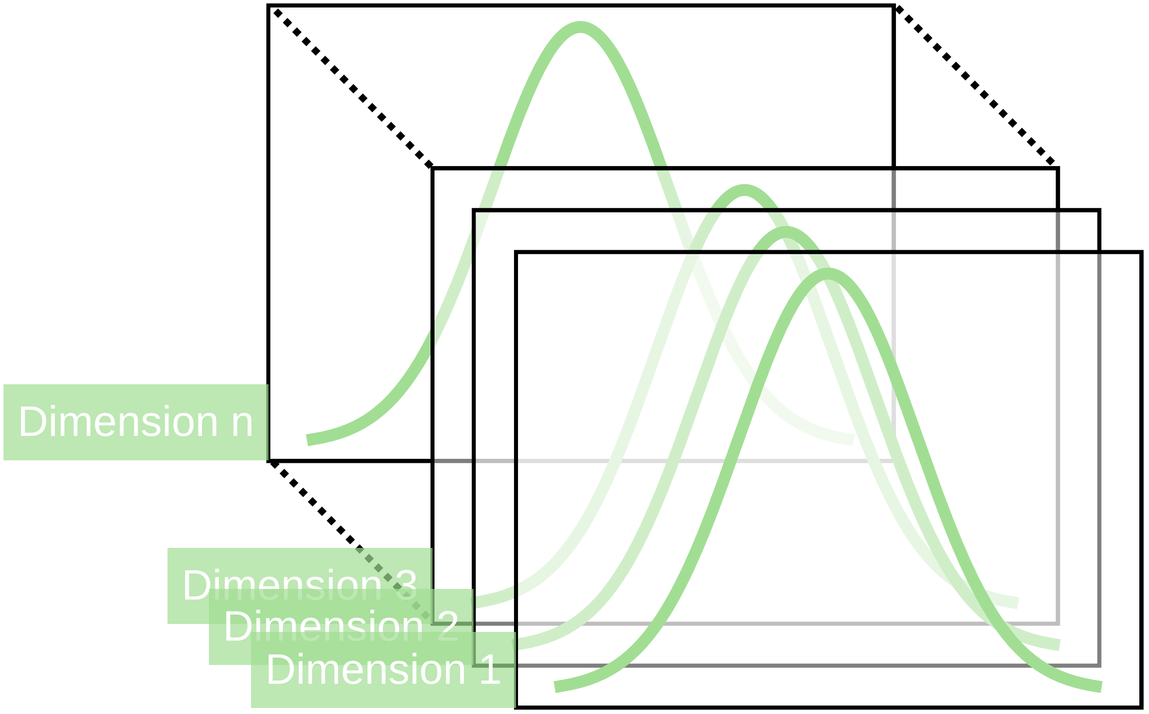


Independent of data distribution (CCS'21)

Low efficiency (ICLR'21, NeurIPS'21)

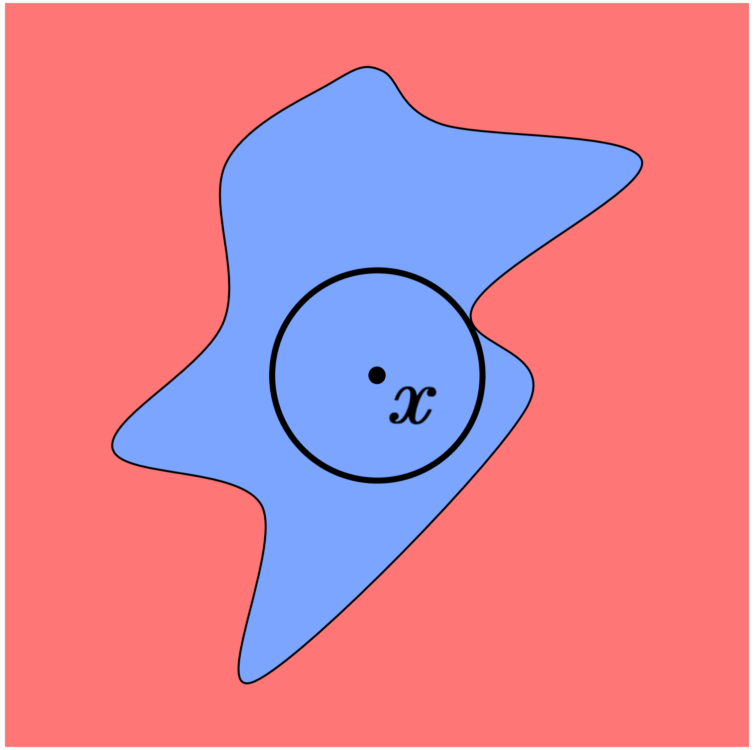
We need real-time certification!

Classical Randomized Smoothing



Classical isotropic noise

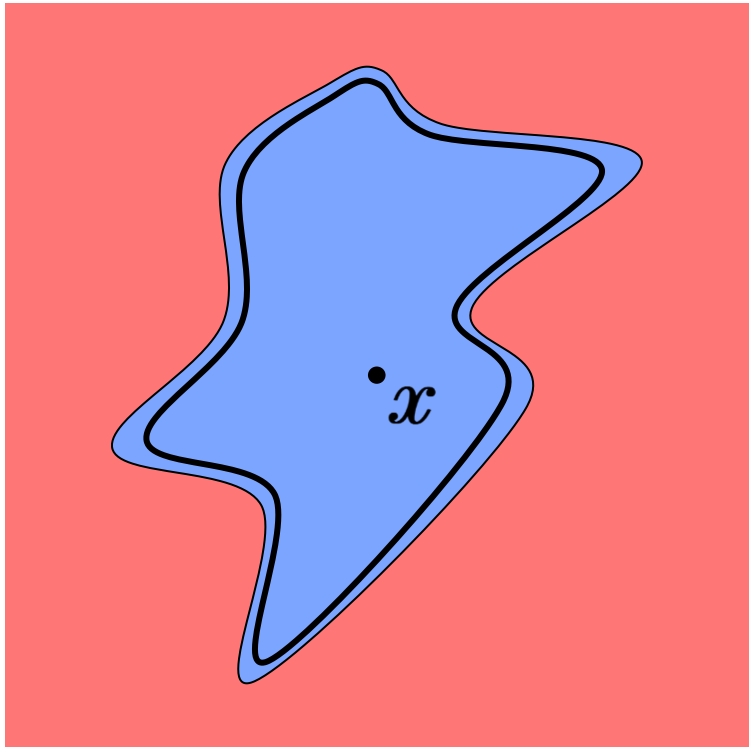
Classical Randomized Smoothing



Local robustness region

Classical isotropic noise is not suitable for highly heterogeneous features!

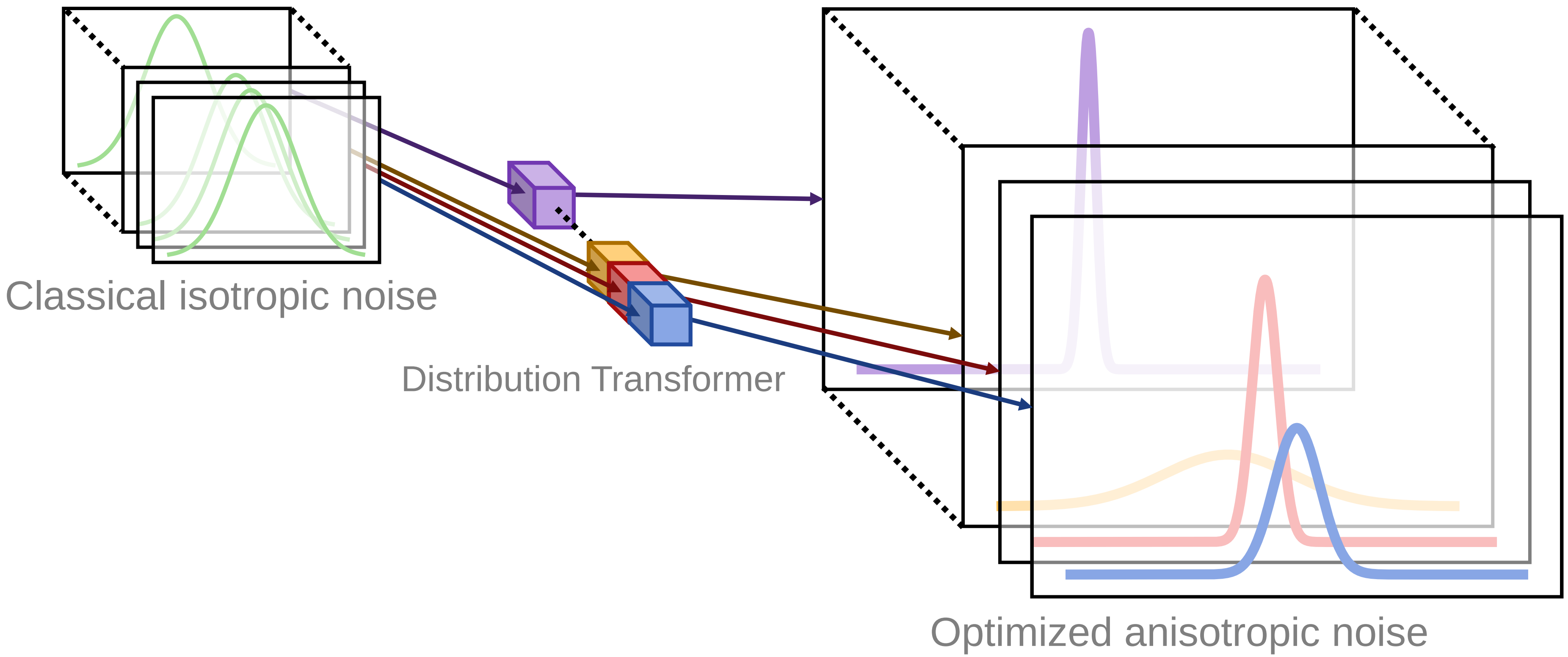
Optimized Randomized Smoothing



Local robustness region

We need adaptive smoothing noise!

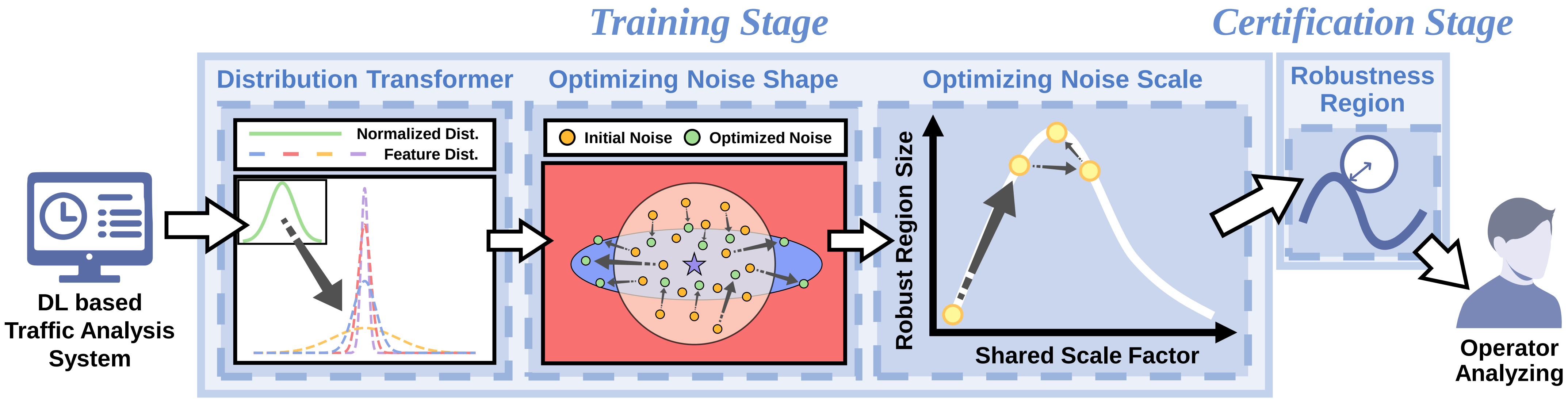
Optimized Randomized Smoothing



Transform classical isotropic noise to optimized anisotropic noise.

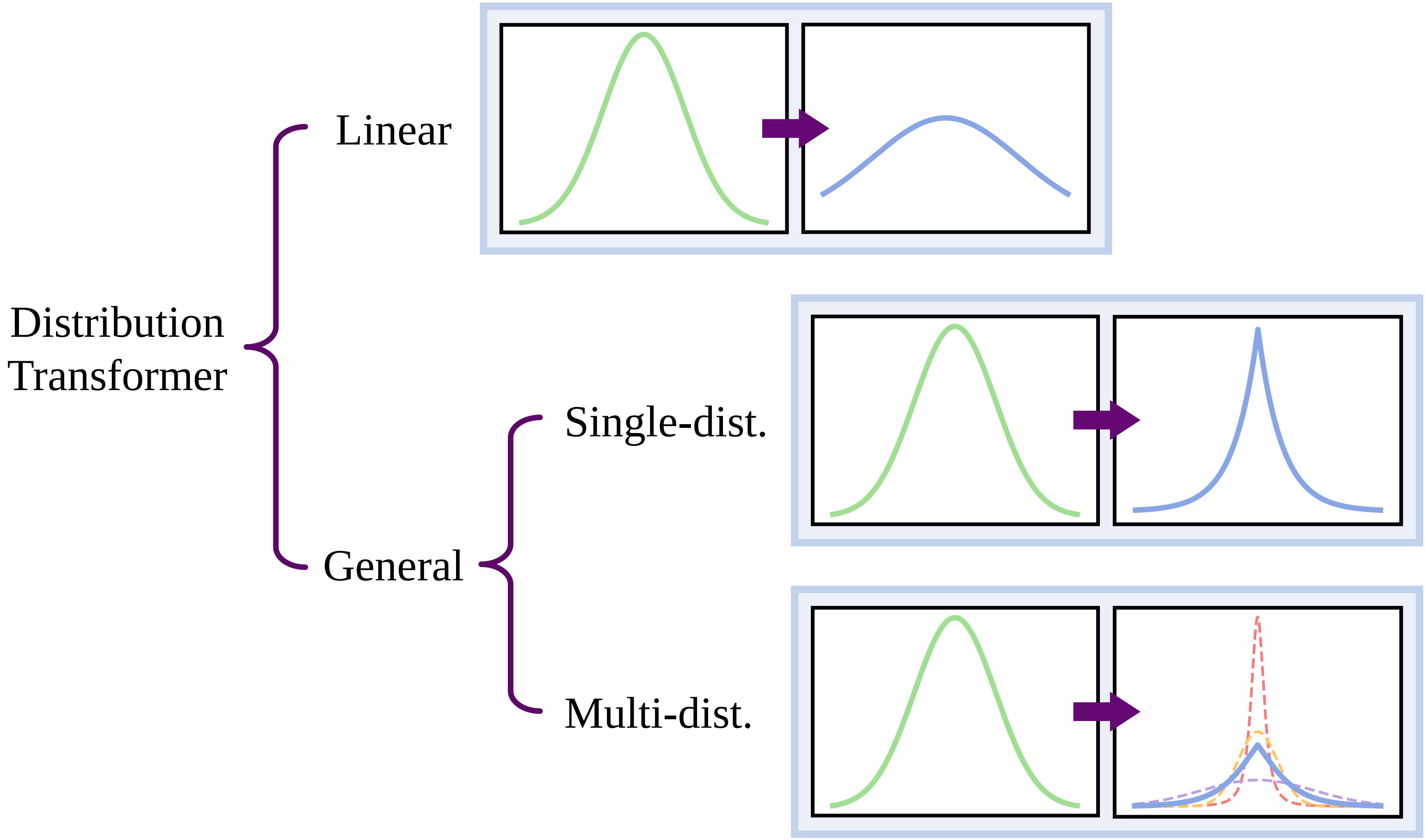
Overview

BARS (Boundary-Adaptive Randomized Smoothing)

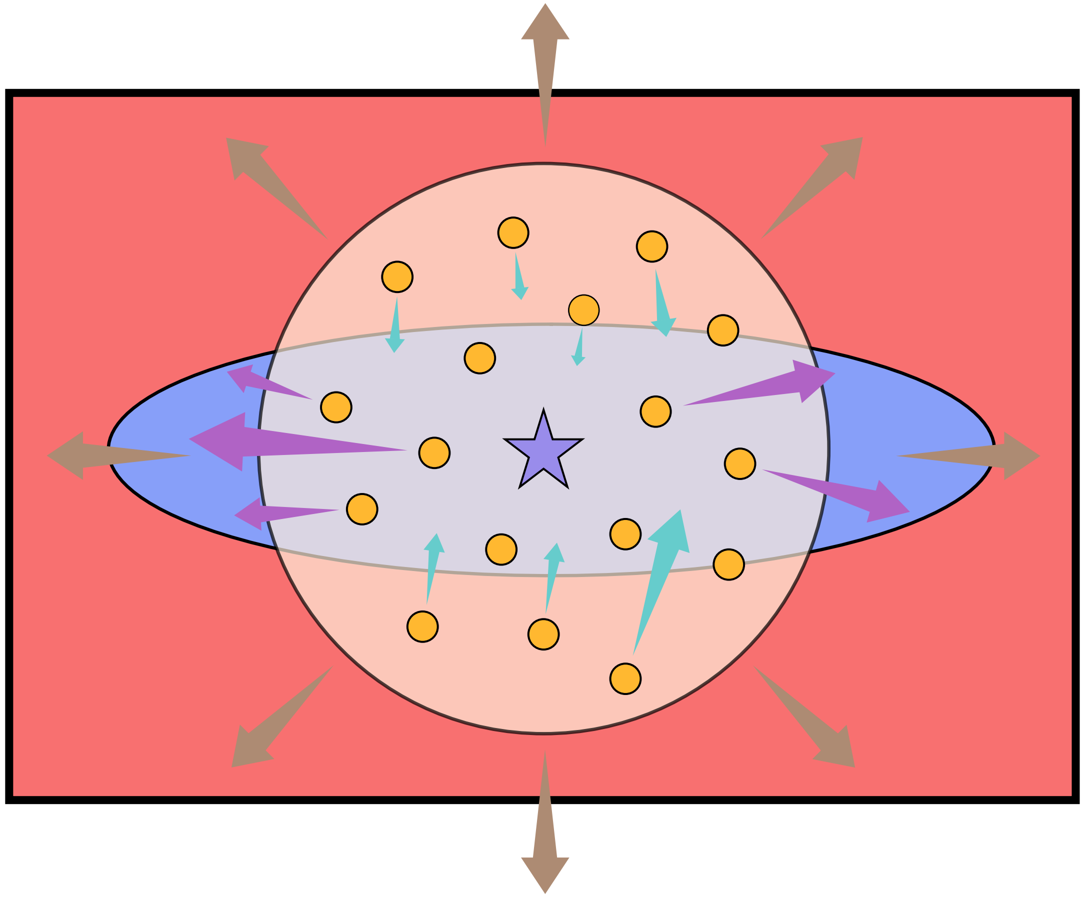


- Dimension-heterogenous smoothing √
- Assuming nothing about model designs √
- Efficient implementation in parallel √

Distribution Transformer



Optimizing Noise Shape



$$\mathcal{L} = \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}_{\varepsilon_n \sim D_n}\left(\mathcal{L}_w\left(\boldsymbol{x}^{(i)}, \varepsilon_n\right) + \mathcal{L}_c\left(\boldsymbol{x}^{(i)}, \varepsilon_n\right)\right) + \lambda\Lambda(\Theta),$$

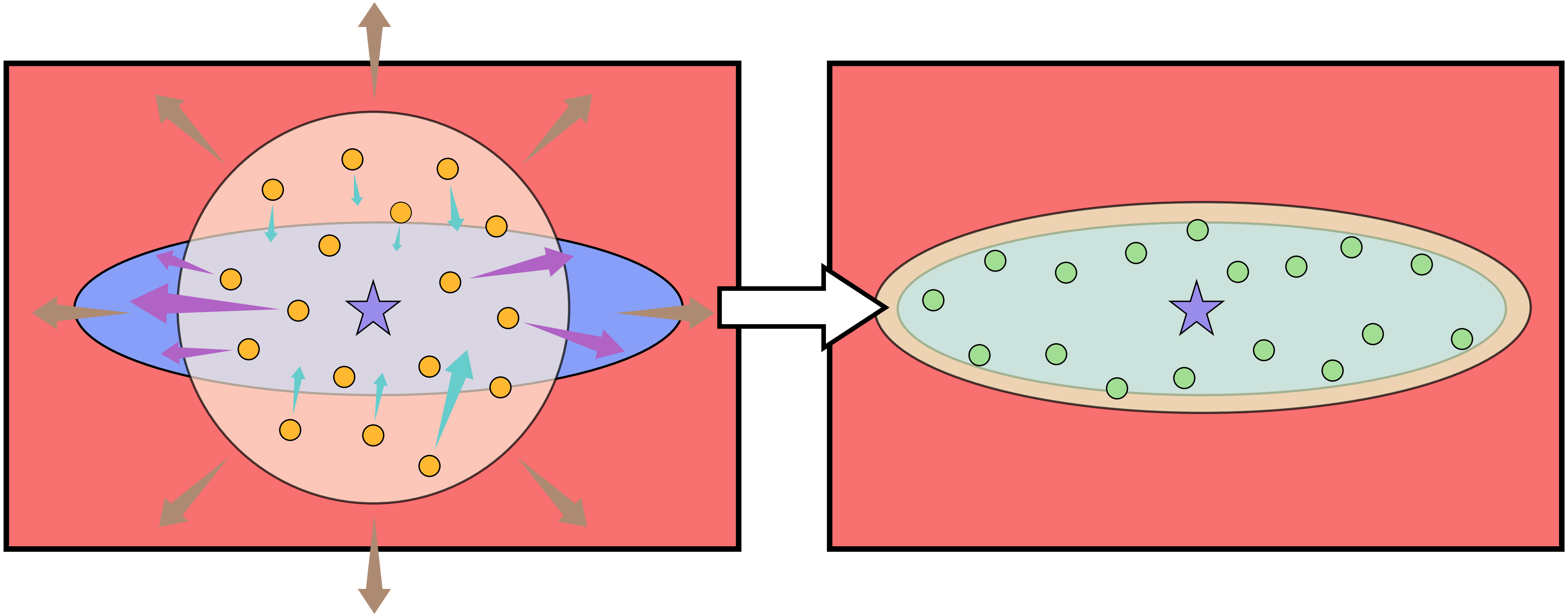
$$\mathcal{L}_w\left(\boldsymbol{x}^{(i)}, \varepsilon_n\right) = \mathbb{I}\left\{f\left(\boldsymbol{x}^{(i)} + \Psi(\varepsilon_n)\right) \neq f\left(\boldsymbol{x}^{(i)}\right)\right\} \cdot \mathcal{L}_{C,w}\left(s\left(\boldsymbol{x}^{(i)} + \Psi(\varepsilon_n)\right), f\left(\boldsymbol{x}^{(i)}\right)\right),$$

$$\mathcal{L}_c\left(\boldsymbol{x}^{(i)}, \varepsilon_n\right) = \mathbb{I}\left\{f\left(\boldsymbol{x}^{(i)} + \Psi(\varepsilon_n)\right) = f\left(\boldsymbol{x}^{(i)}\right)\right\} \cdot \mathcal{L}_{C,c}\left(s\left(\boldsymbol{x}^{(i)} + \Psi(\varepsilon_n)\right), f\left(\boldsymbol{x}^{(i)}\right)\right),$$

$$\Lambda(\Theta) = \sum_{w \in \Theta}\log\left(1 + e^{-w}\right).$$

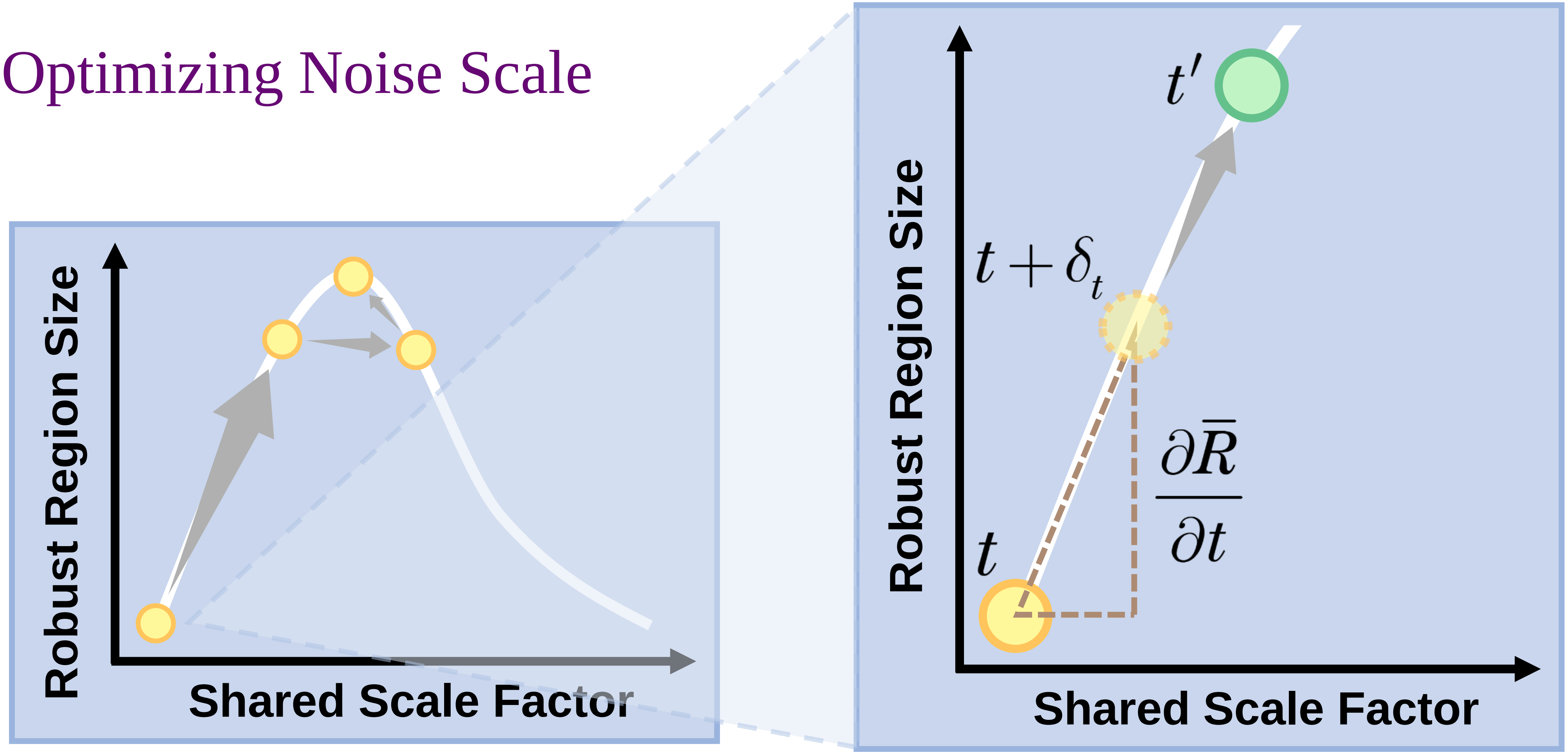
Move noised samples close to classification boundary.

Optimizing Noise Shape



Move noised samples close to classification boundary.

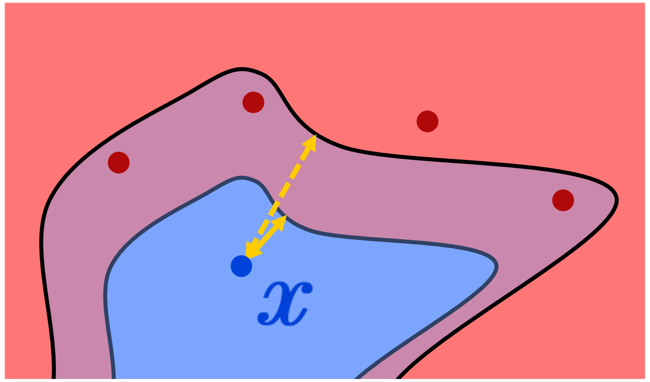
Optimizing Noise Scale



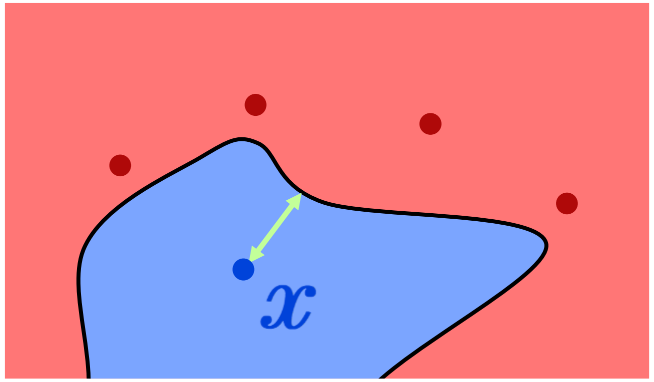
Optimize the shared scale factor for tight robustness guarantee.

Application Case I

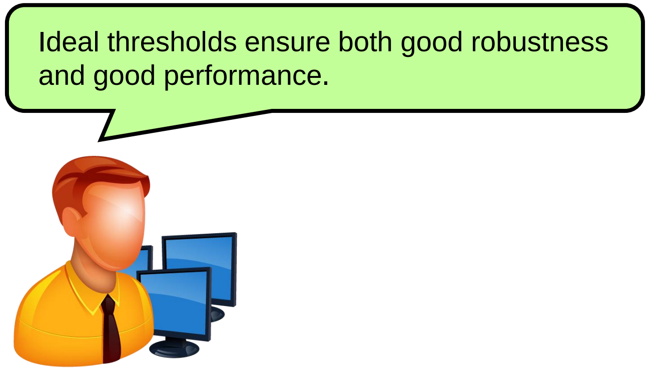
Quantitatively Evaluating Robustness (Detection Threshold)

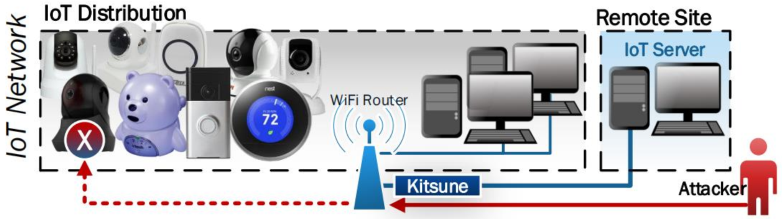


Too small or too large threshold

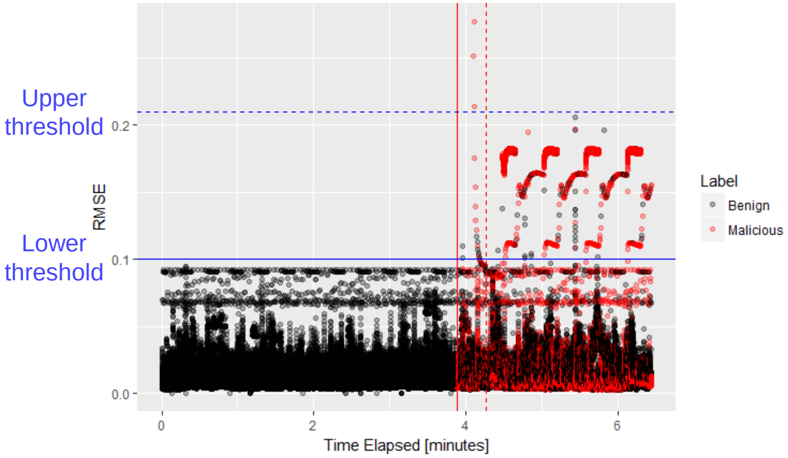


Ideal threshold

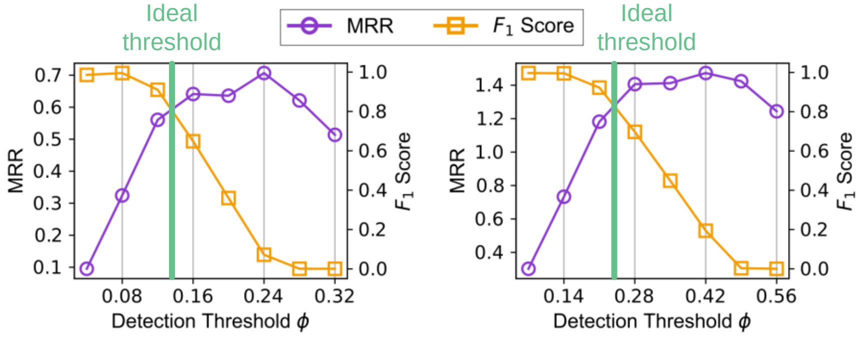




Kitsune (NDSS'18)



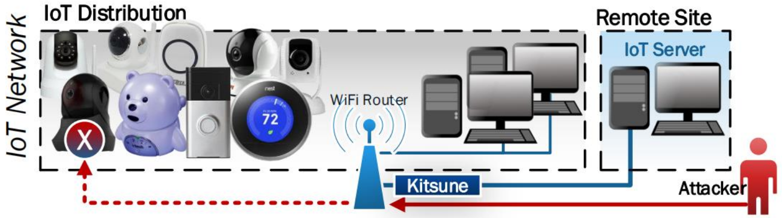
Kitsune (NDSS'18)



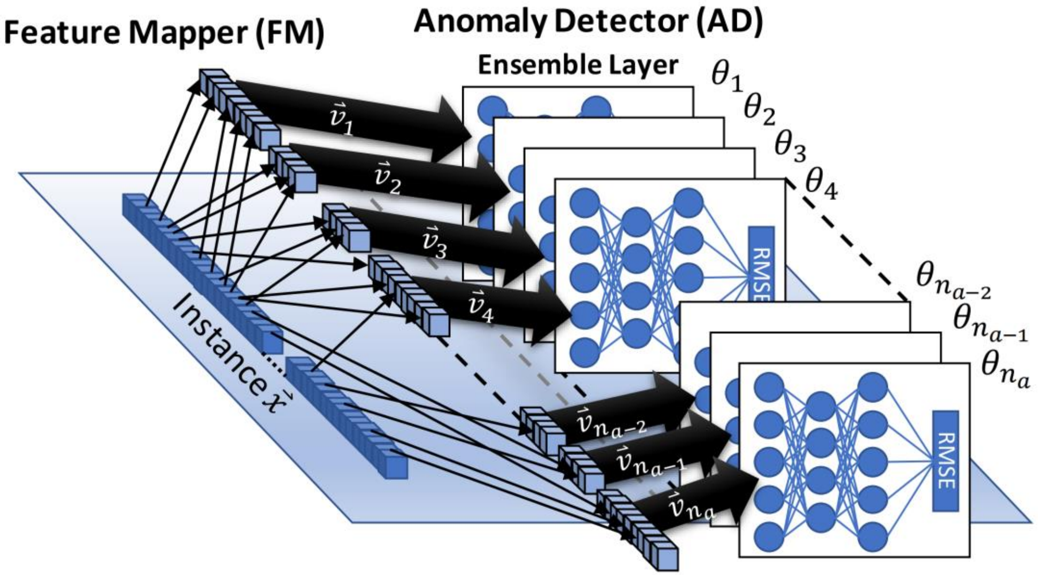
(a) 32 AEs ($m = 7$). (b) 16 AEs ($m = 16$).

Application Case I

Quantitatively Evaluating Robustness (AE Number)



Kitsune (NDSS'18)



Kitsune (NDSS'18)

Mean robustness radius (Robustness)

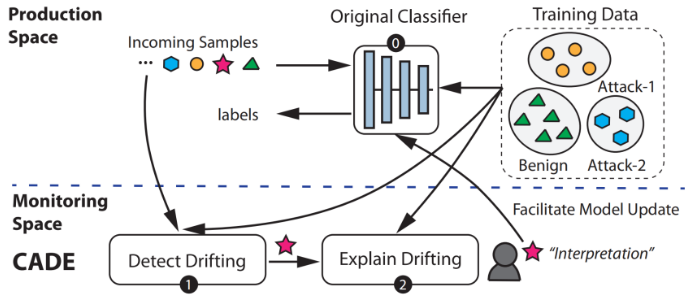
Coefficient of variation for robustness radius (Fitting capability)

Ideal AE number

AE Number	m	MRR	CVR	F_1 Score
1	100	3.4749	0.1409	0.9796
2	80	4.5540	0.2525	0.9793
4	75	4.2375	0.3740	0.9797
8	43	2.3316	0.6326	0.9806
16	16	0.9923	0.6729	0.9806
32	7	0.4628	0.8025	0.9802
64	2	2.5844	0.3210	0.9784
100	1	2.2312	0.2712	0.9782

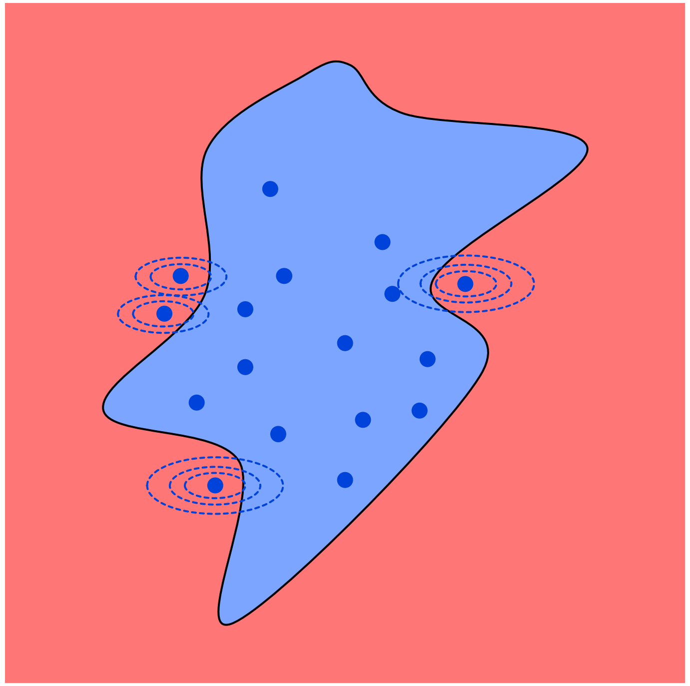
Application Case II

Reducing False Alarms



CADE (USENIX Security'21)

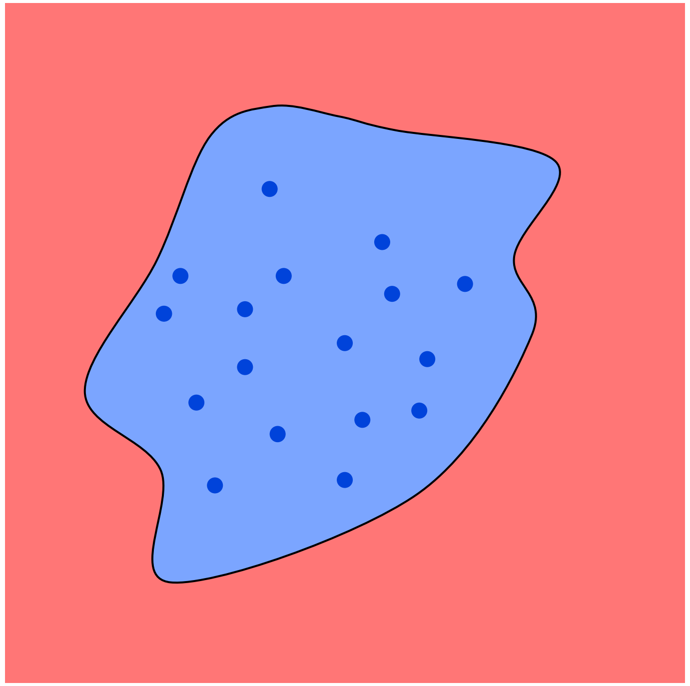
Vanilla CADE



FPR				FNR
Benign	SSH-Bruteforce	DoS-Hulk	Total	Total
0.0495	0.0418	0.0110	0.0350	0.0000

Retraining

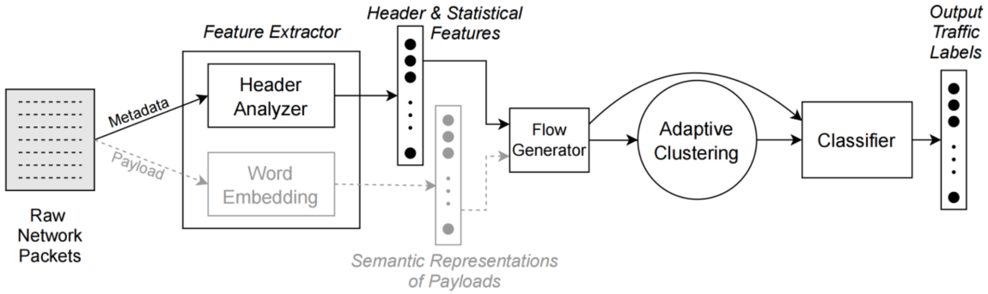
Noise data augmentation retraining



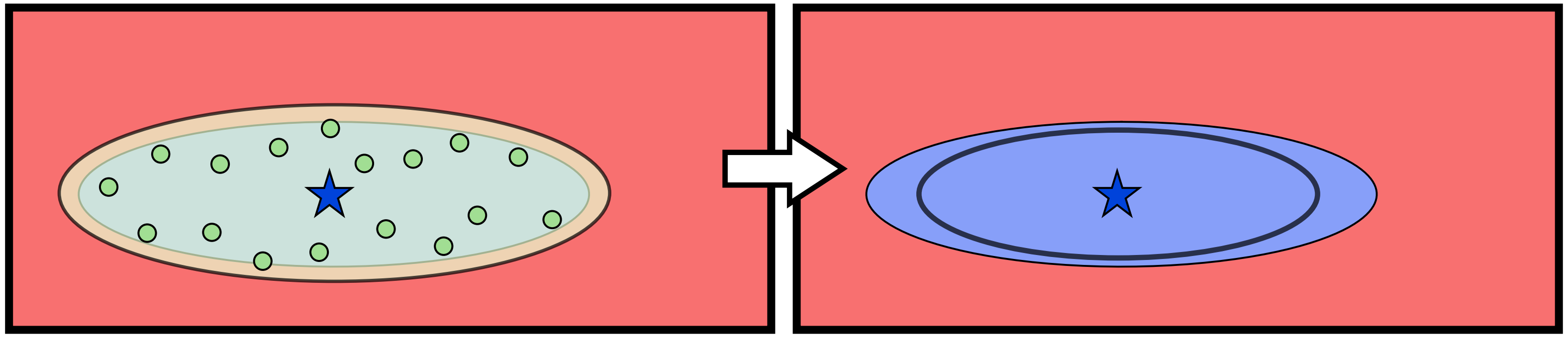
FPR				FNR
Benign	SSH-Bruteforce	DoS-Hulk	Total	Total
0.0283	0.0128	0.0066	0.0190	0.0000

Application Case III

Evasion Attack Awareness

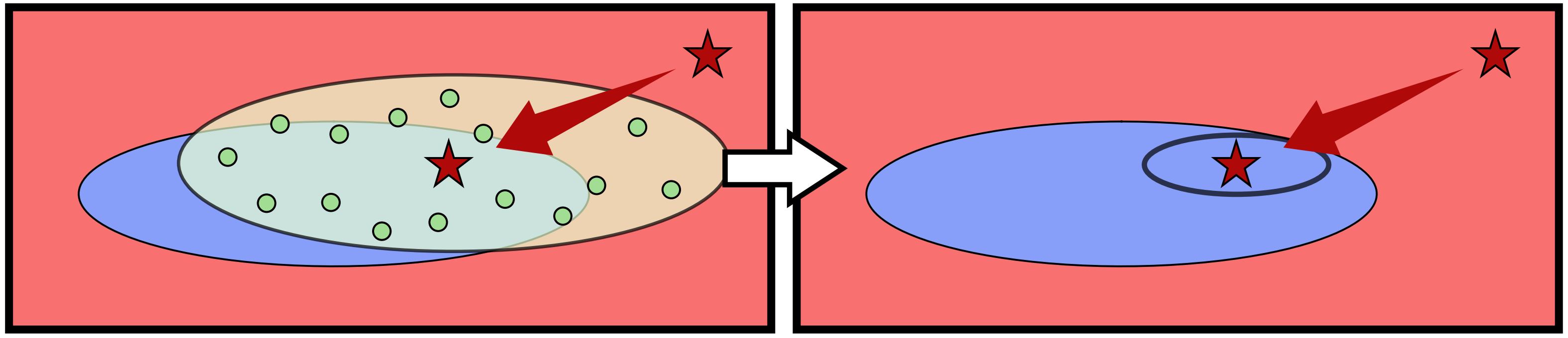


ACID (INFOCOM'21)



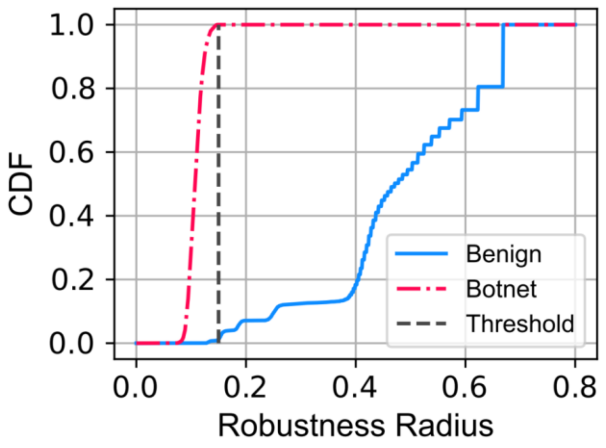
Smoothed clean sample

Clean sample robustness region



Smoothed evasion sample

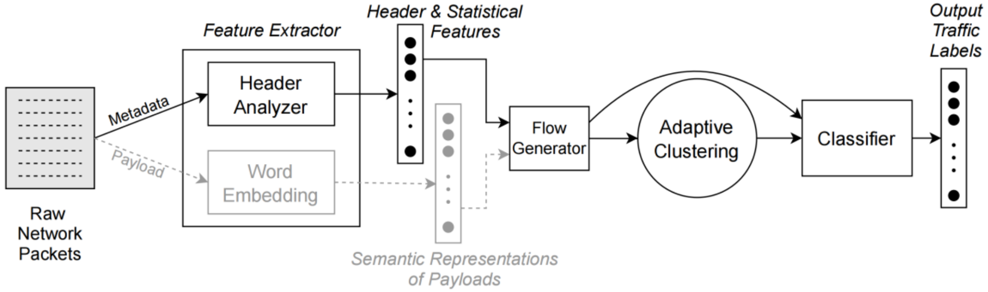
Evasion sample robustness region



Method	Precision	Recall	F_1 Score
V.R.S.	0.6861	0.8380	0.7544
BARS-L	0.9819	0.9181	0.9489
BARS-G	0.9455	1.0000	0.9720

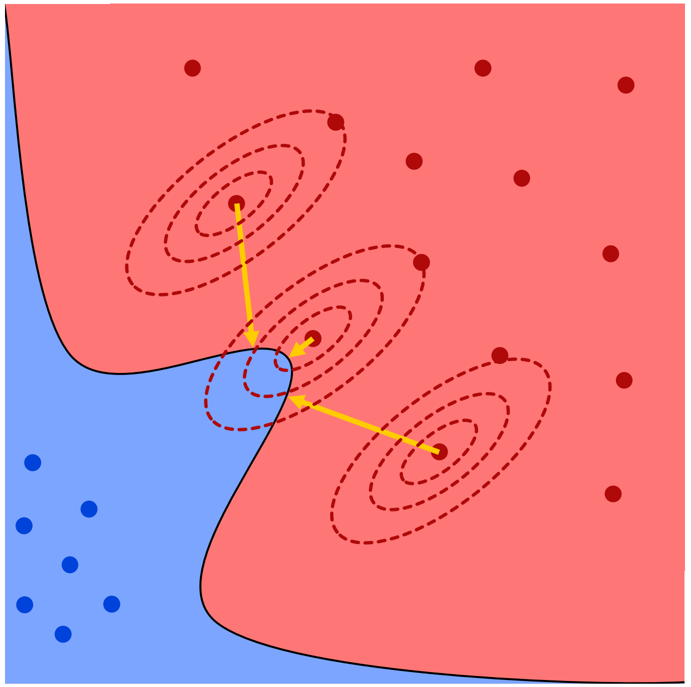
Application Case IV

Evasion Attack Defense



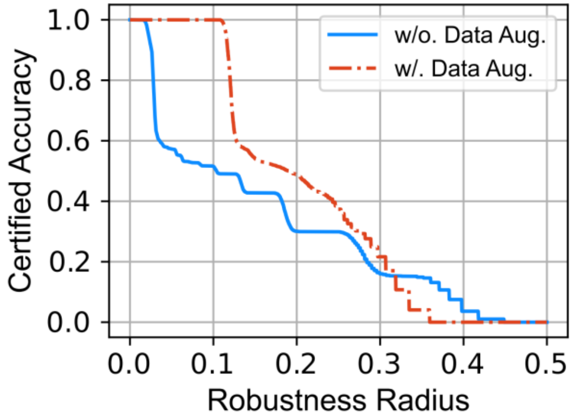
ACID (INFOCOM'21)

Vanilla ACID

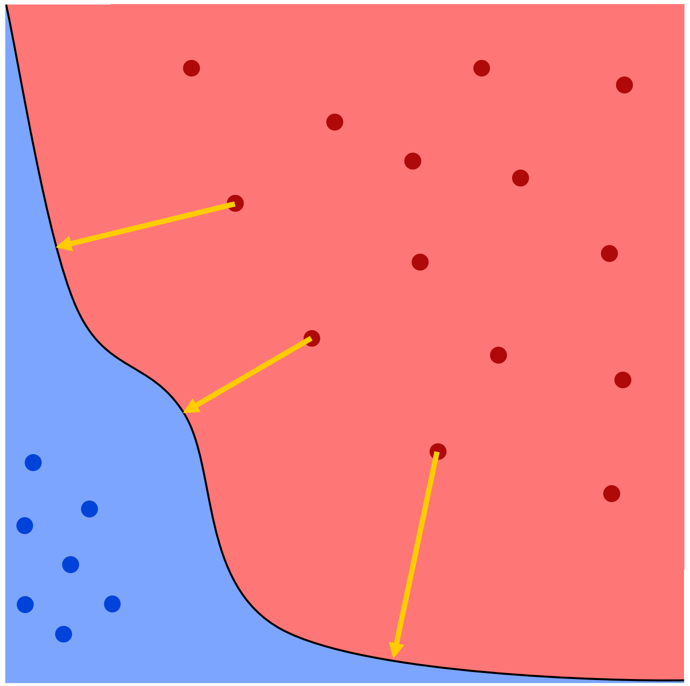


Evasion Success Rate		
Random	PGD	B.A.P.
0.3069	1.0000	1.0000

Retraining



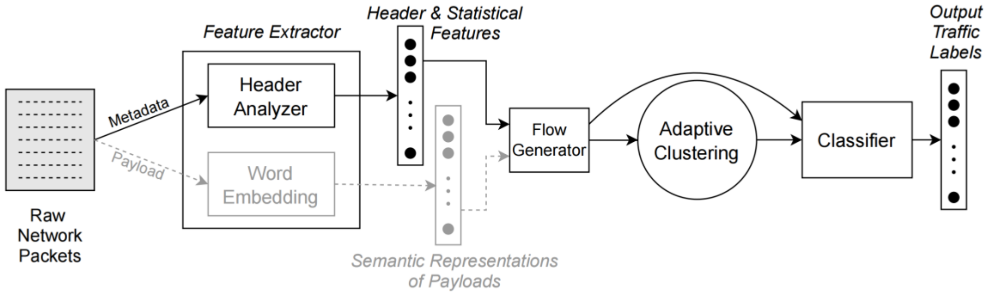
Noise Data Augmentation Retraining



Evasion Success Rate		
Random	PGD	B.A.P.
0.2024	0.4006	0.8475

Application Case V

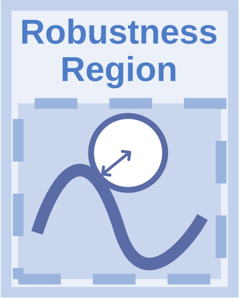
Explaining Attack Detection



ACID (INFOCOM'21)

Feature	Radius	Description
Init Fwd Win Byts	5.1728×10^{-2}	Total number of bytes sent in initial window in forward direction.
Fwd IAT Max	9.3542×10^{-2}	Maximum time between two packets sent in forward direction.
Mean Radius	1.8561×10^{-1}	Mean robustness radius in all dimensions.

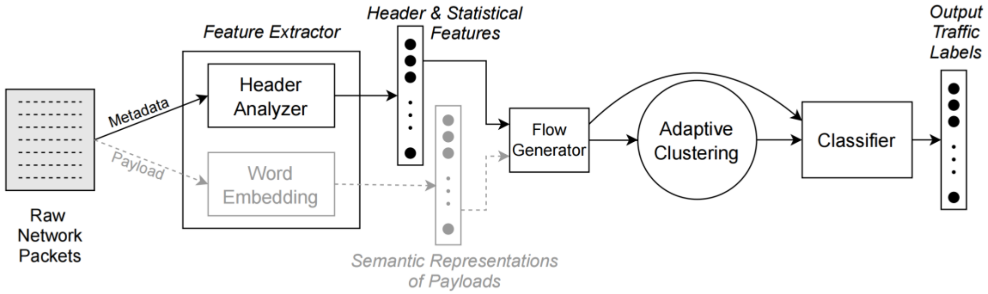
Certification Stage



Classification results are sensitive to these weakly robust features. They are important!

Application Case V

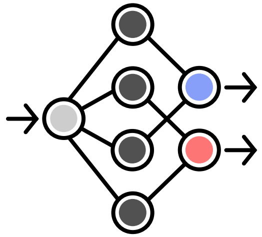
Explaining Attack Detection



ACID (INFOCOM'21)

How is its fidelity?

Please replace the values of important features with random numbers.



F_1	F_2	F_3	…	F_n
0.1	0.5	0.9	…	0.1
0.3	0.6	0.9	…	0.2
0.2	0.8	0.8	…	0.4
0.5	0.5	0.7	…	0.2

Original features

F_1	F_2	F_3	…	F_n
0.1	0.5	ξ	…	0.1
0.3	0.6	ξ	…	0.2
0.2	0.8	ξ	…	0.4
0.5	0.5	ξ	…	0.2

Replaced values

Metric	Vanilla	Random	BARS-L	BARS-G
Precision	1.0000	0.9928	0.9423	0.9064
Recall	1.0000	0.9040	0.7918	0.7707
F_1 Score	1.0000	0.9397	0.8605	0.8330

Performance under random feature values

Summary

We propose a general robustness certification framework for DL-based traffic analyzers.

- Dimension-heterogeneous, universal, real-time

We show how to apply the framework to five domain-specific problems of traffic analysis.

- Quantitatively evaluating robustness, reducing false alarms, evasion attack awareness, evasion attack defense, explaining attack detection

We implement the framework on three practical DL-based traffic analyzers.

- Zero-positive NIDS, concept drift detection system, supervised multi-classification system

https://github.com/KaiWangGitHub/BARS

Thank you!

BARS: Local Robustness Certification for Deep Learning based Traffic Analysis Systems

Presenter: Kai Wang
k-wang20@mails.tsinghua.edu.cn

清華大學
Tsinghua University