

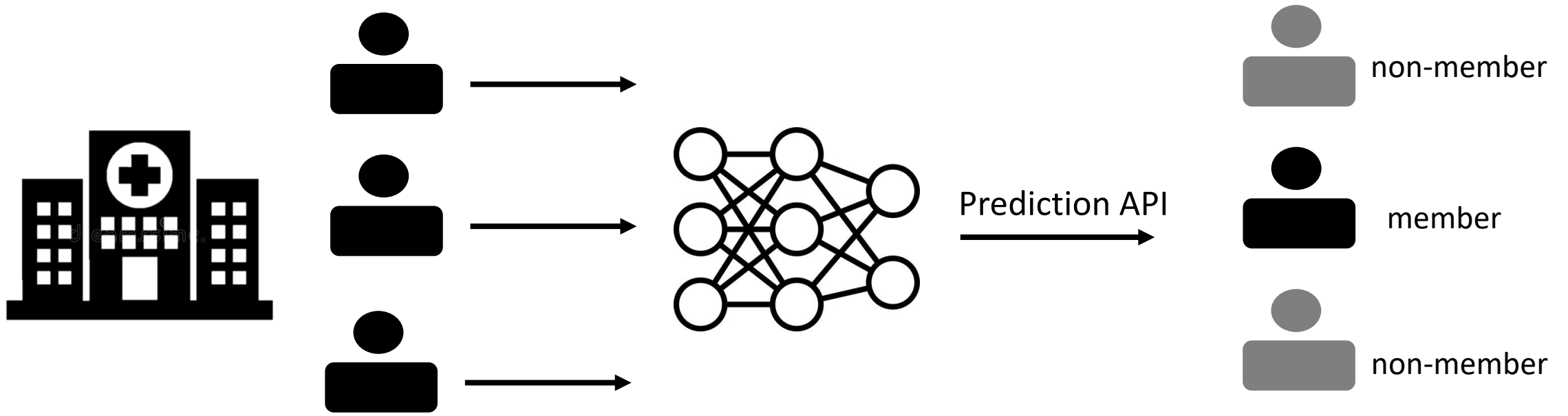
Overconfidence is a Dangerous Thing: Mitigating Membership Inference Attacks by Enforcing Less Confident Prediction

Zitao Chen, Karthik Pattabiraman



THE UNIVERSITY
OF BRITISH COLUMBIA

Membership Inference Attacks (MIAs)



Does the sensitive training set contain a target record?

MIAs as a privacy threat

MIAs as a privacy threat

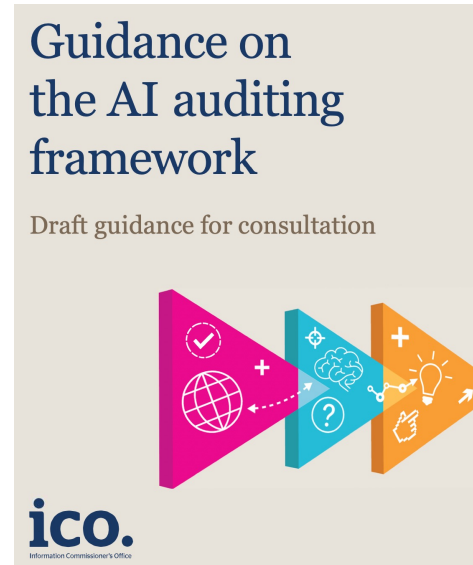


**Confidentiality
violation**

MIAs as a privacy threat

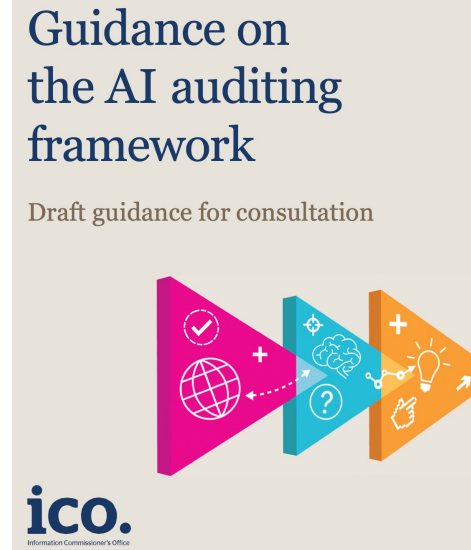


**Confidentiality
violation**



**Auditing
purpose**

MIAs as a privacy threat



Training Set



Caption: Living in the light with Ann Graham Lotz

Generated Image



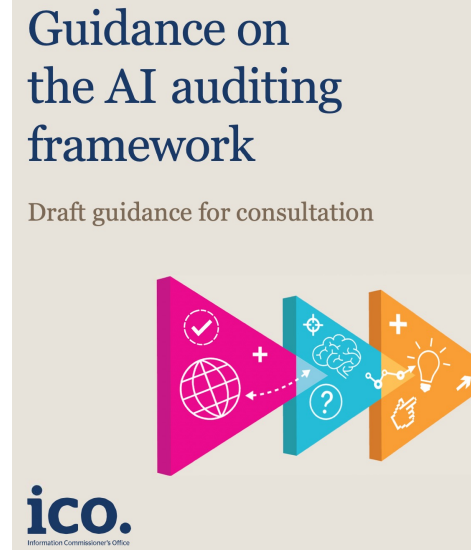
Prompt: Ann Graham Lotz

Confidentiality violation

Auditing purpose

Stepping stone for more powerful attack

MIAs as a privacy threat



Training Set



Caption: Living in the light with Ann Graham Lotz

Generated Image



Prompt: Ann Graham Lotz

Confidentiality violation

Auditing purpose

Stepping stone for more powerful attack

We need effective defense against MIAs!

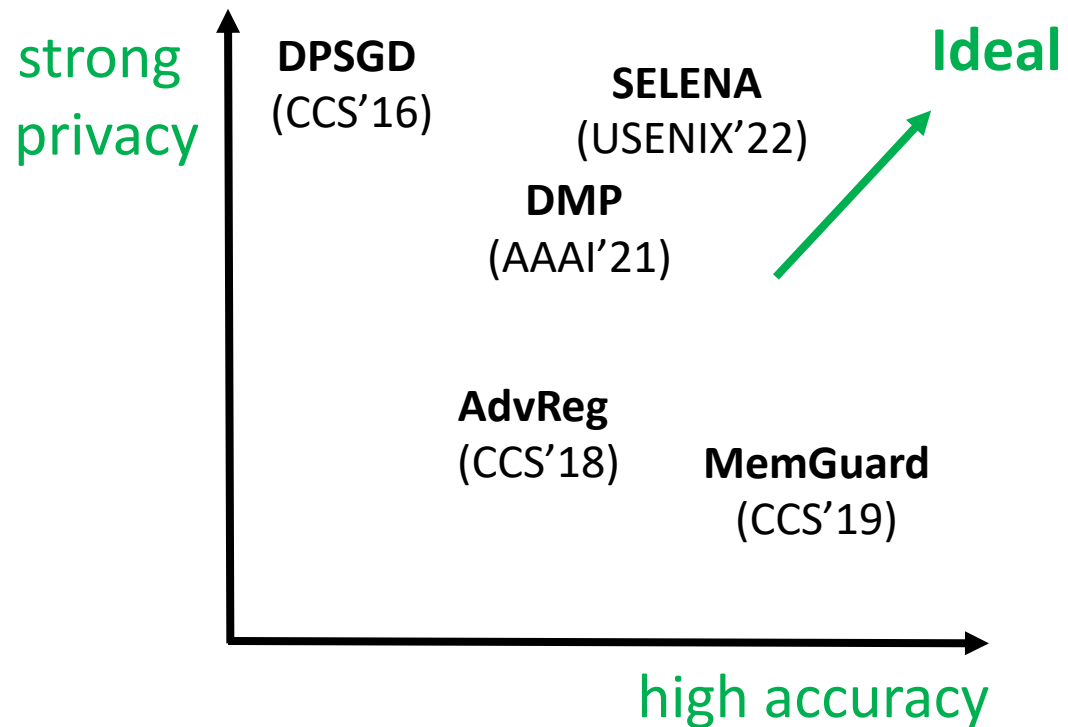
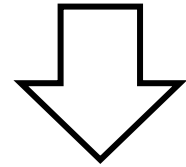
Existing defenses

Existing defenses

Poor privacy-utility trade off or requiring additional data

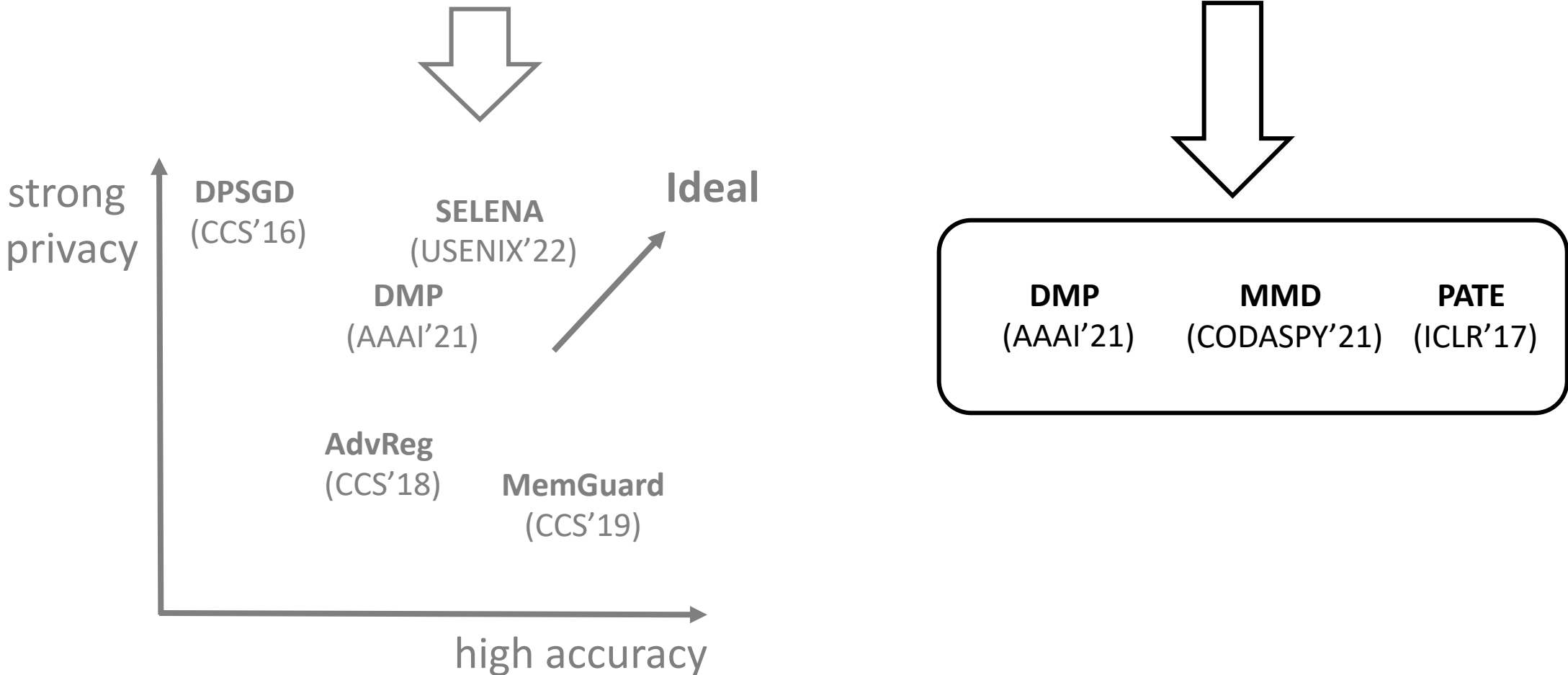
Existing defenses

Poor privacy-utility trade off or requiring additional data



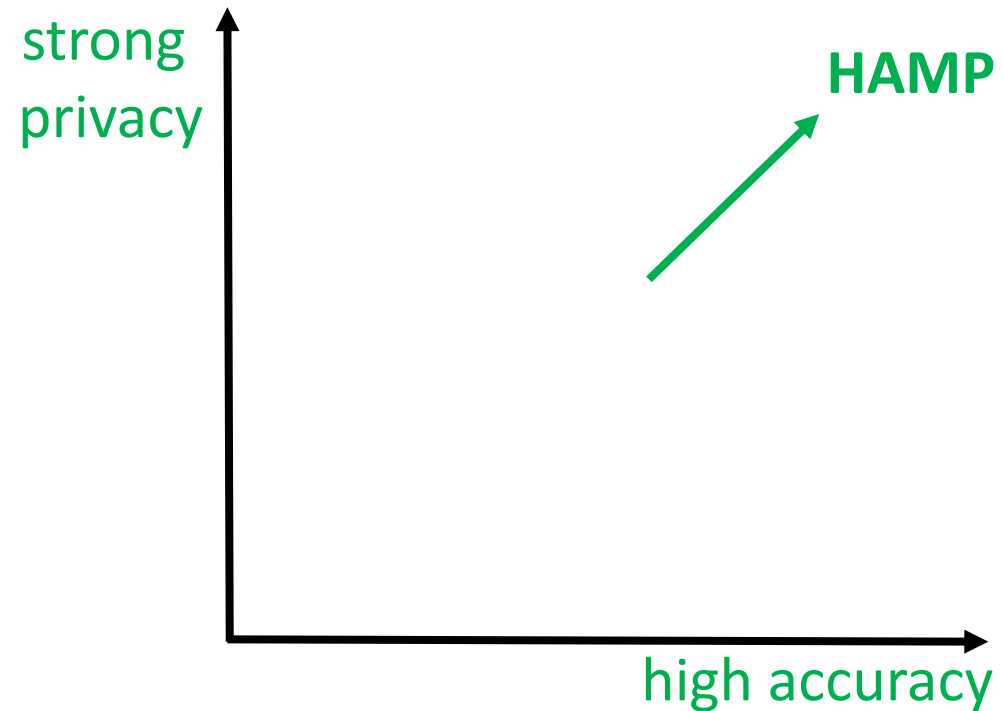
Existing defenses

Poor privacy-utility trade off or requiring additional data



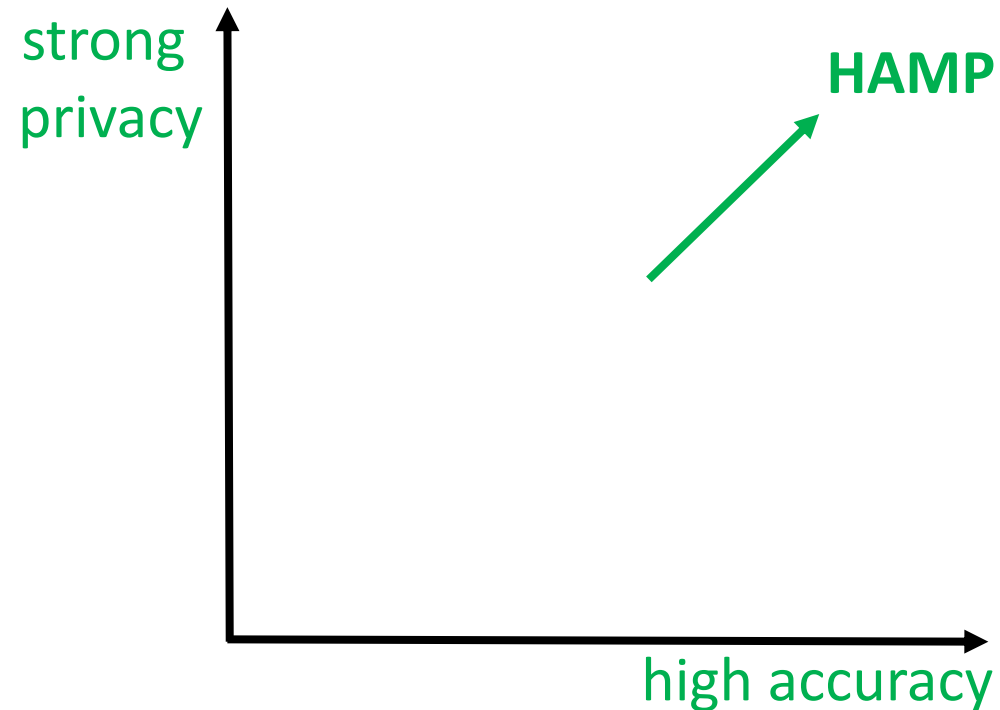
Our Work: HAMP

High Accuracy and Membership Privacy without additional data



Our Work: HAMP

A new way to combine **soft label training**, **training regularization** and **output modification** for **privacy-preserving training!**



Threat model

Adversary

- **Knowledge:**
 - Black-box adversary.
 - Half members and non-members.
 - Full defense knowledge.
- **Goal:** Membership inference

Threat model

Adversary

- **Knowledge:**
 - Black-box adversary.
 - Half members and non-members.
 - Full defense knowledge.
- **Goal:** Membership inference

Defender

- **Knowledge:**
 - The private dataset only.
- **Goal:** Model with high accuracy & membership privacy

Existing attacks

Diverse strategies

Existing attacks

Diverse strategies

Scaled-logit loss

Prediction entropy

Adv robustness

...

Existing attacks

Diverse strategies

Scaled-logit loss

Prediction entropy

Adv robustness

...



Common exploitation

Existing attacks

Diverse strategies

Scaled-logit loss

Prediction entropy

Adv robustness

...



Common exploitation

Exploit ML model's overconfident prediction on training samples

Existing attacks

Diverse strategies

Scaled-logit loss

Prediction entropy

Adv robustness

...



Common exploitation

Exploit ML model's overconfident prediction on training samples



Overconfident Prediction manifest as



Existing attacks

Diverse strategies

Scaled-logit loss

Prediction entropy

Adv robustness

...

Common exploitation

Exploit ML model's overconfident prediction on training samples

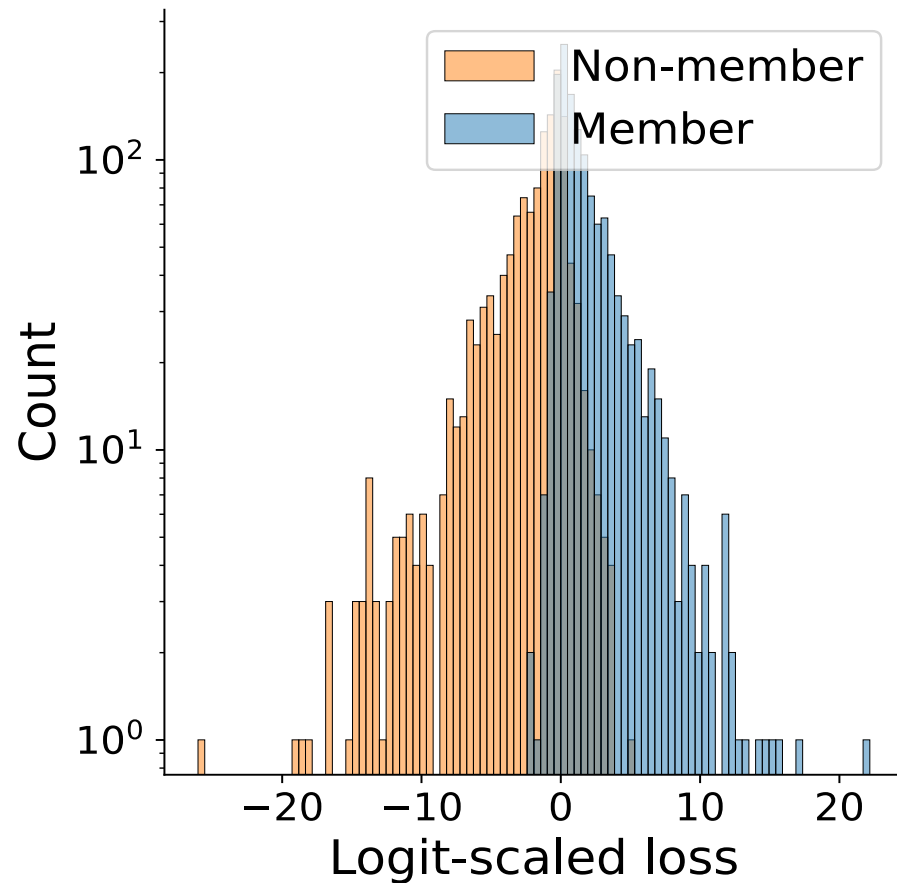
**Overconfident
Prediction**

manifest as

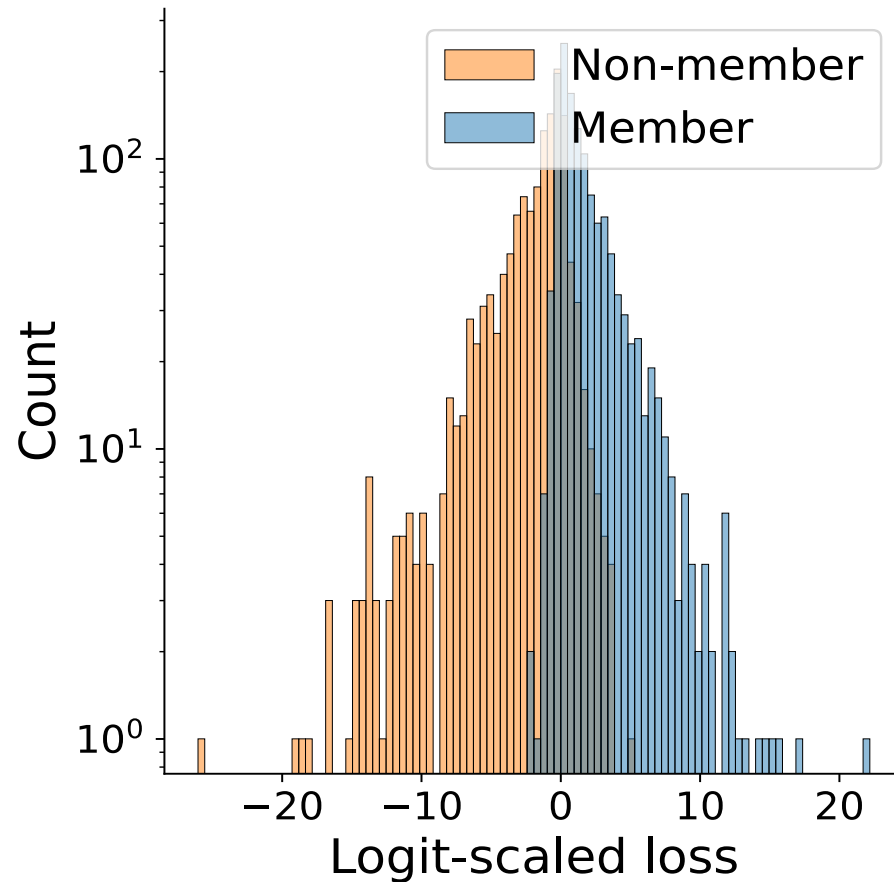
- High logit-scaled loss;
- Low prediction entropy;
- High robustness to adv perturbations;
- ...

Example: Overconfident prediction via logit-scaled loss

Example: Overconfident prediction via logit-scaled loss

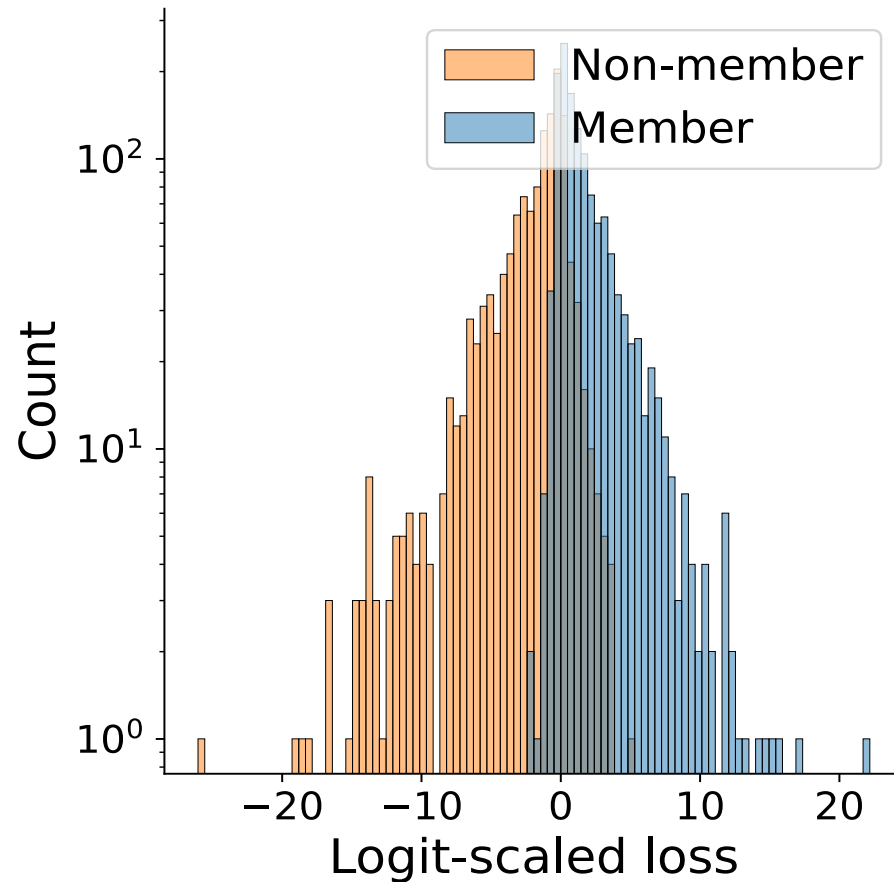


Example: Overconfident prediction via logit-scaled loss



Member samples with
high scaled loss

Example: Overconfident prediction via logit-scaled loss



Member samples with
high scaled loss
↓
**Due to overly high prediction
confidence**

Defense principle



MIAAs exploit ML model's overconfident prediction on training samples



Mitigating ML model's overconfident prediction on training samples without jeopardizing model accuracy

What leads to overconfident prediction?

What leads to overconfident prediction?

- ❑ Training with **one-hot hard label**.

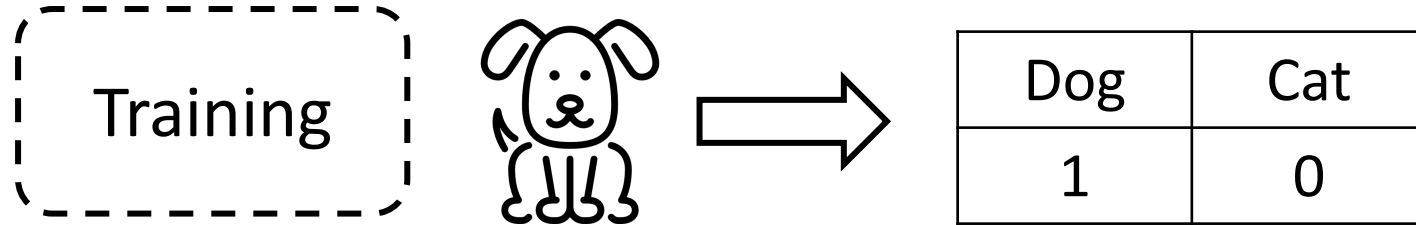
What leads to overconfident prediction?

- ❑ Training with **one-hot hard label**.

Training

What leads to overconfident prediction?

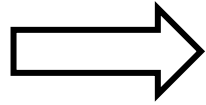
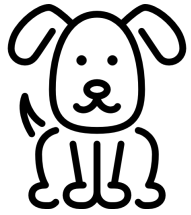
- ❑ Training with **one-hot hard label**.



What leads to overconfident prediction?

- Training with **one-hot hard label**.

Training

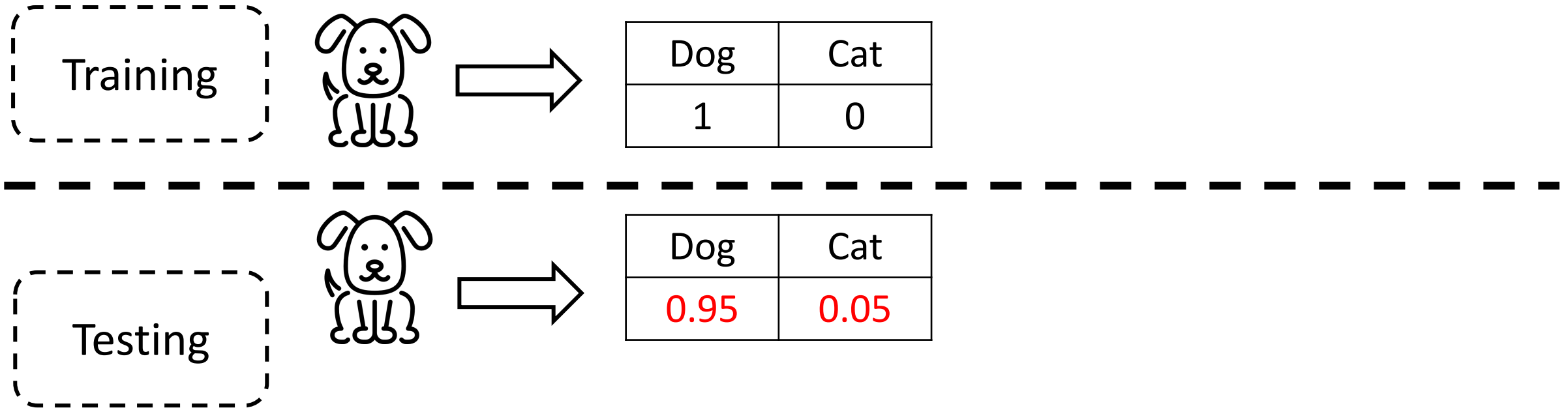


Dog	Cat
1	0

Testing

What leads to overconfident prediction?

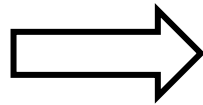
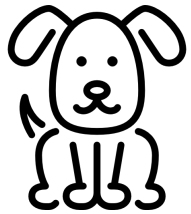
- Training with **one-hot hard label**.



What leads to overconfident prediction?

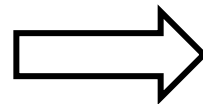
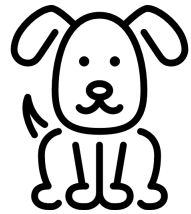
- Training with **one-hot hard label**.

Training

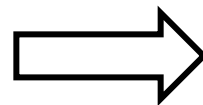
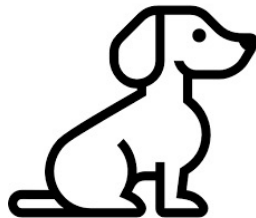


Dog	Cat
1	0

Testing



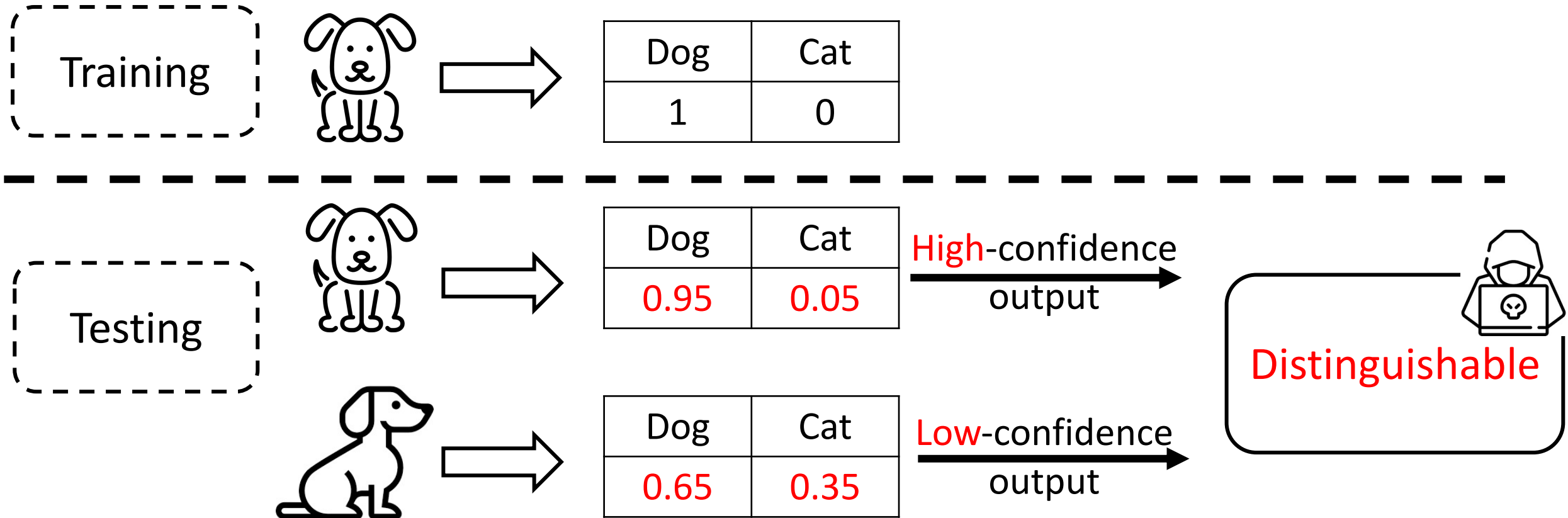
Dog	Cat
0.95	0.05



Dog	Cat
0.65	0.35

What leads to overconfident prediction?

- Training with **one-hot hard label**.



HAMP

Training-time defense

Testing-time defense

HAMP

Training-time defense

Testing-time defense



Produce high-utility models with
strong membership privacy

HAMP

Training-time defense



Produce high-utility models with
strong membership privacy

Testing-time defense



Gain higher privacy without
degrading accuracy

HAMP - Training-time defense

High-entropy soft labels

HAMP - Training-time defense

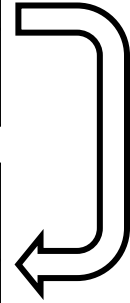
High-entropy soft labels

Original hard
label

Dog	Cat
1	0

High-entropy
soft label

Dog	Cat
0.7	0.3



HAMP - Training-time defense

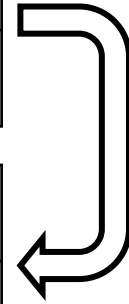
High-entropy soft labels

Original hard
label

Dog	Cat
1	0

High-entropy
soft label

Dog	Cat
0.7	0.3



Explicitly enforce the model to
make less confident prediction

HAMP - Training-time defense

High-entropy soft labels

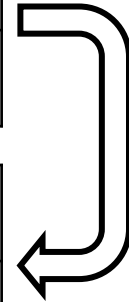
Entropy-based regularization

Original hard
label

Dog	Cat
1	0

High-entropy
soft label

Dog	Cat
0.7	0.3



Explicitly enforce the model to
make less confident prediction

HAMP - Training-time defense

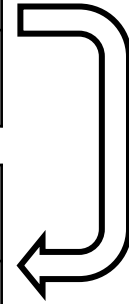
High-entropy soft labels

Original hard label

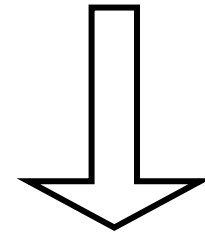
Dog	Cat
1	0

High-entropy soft label

Dog	Cat
0.7	0.3



Entropy-based regularization



Penalize low-entropy predictions

Explicitly enforce the model to make less confident prediction

HAMP - Training-time defense

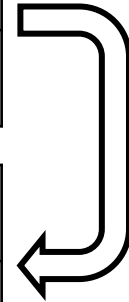
High-entropy soft labels

Original hard label

Dog	Cat
1	0

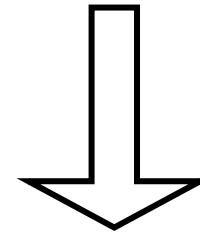
High-entropy soft label

Dog	Cat
0.7	0.3



Explicitly enforce the model to make less confident prediction

Entropy-based regularization



Penalize low-entropy predictions

Regularize the prediction confidence level

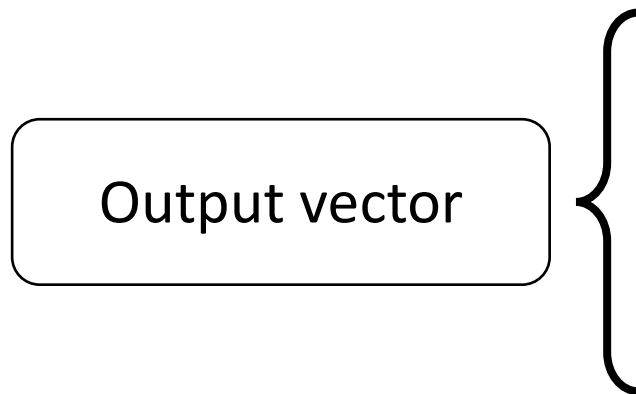
HAMP – Testing-time defense

HAMP – Testing-time defense

- ❑ Modify **all** output vectors → **low** confidence outputs.

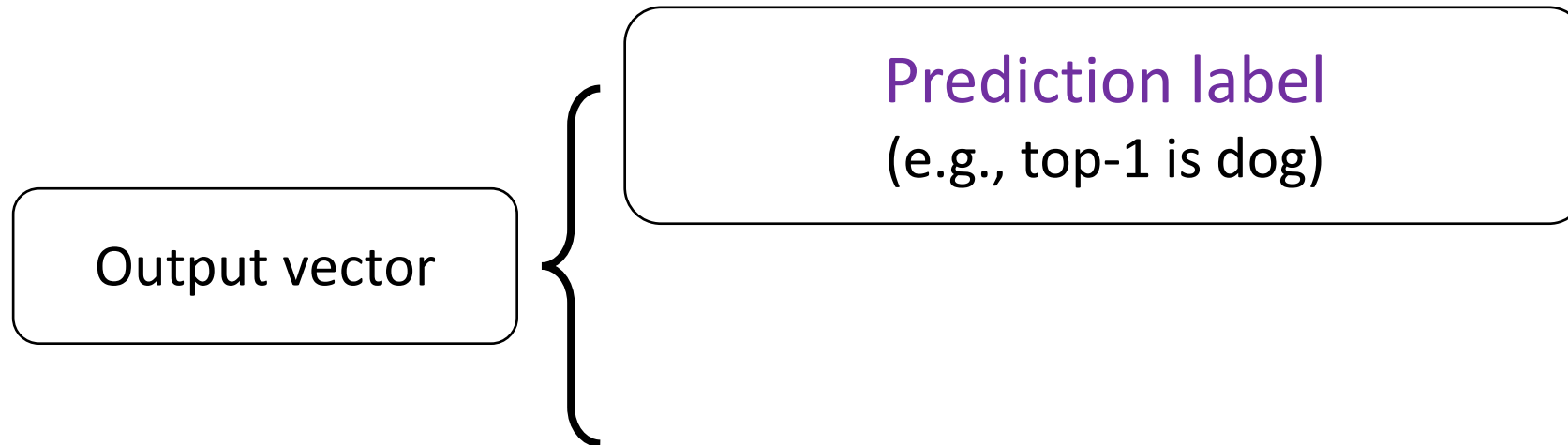
HAMP – Testing-time defense

- ❑ Modify **all** output vectors → **low** confidence outputs.



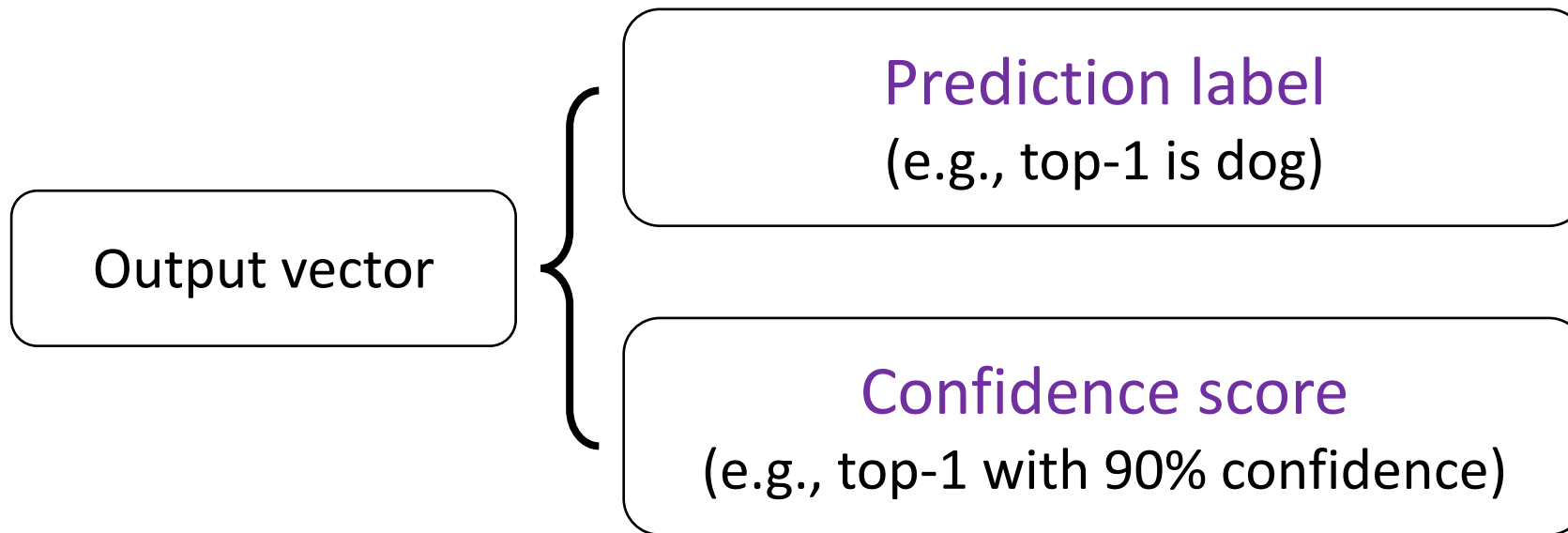
HAMP – Testing-time defense

- ❑ Modify **all** output vectors → **low** confidence outputs.



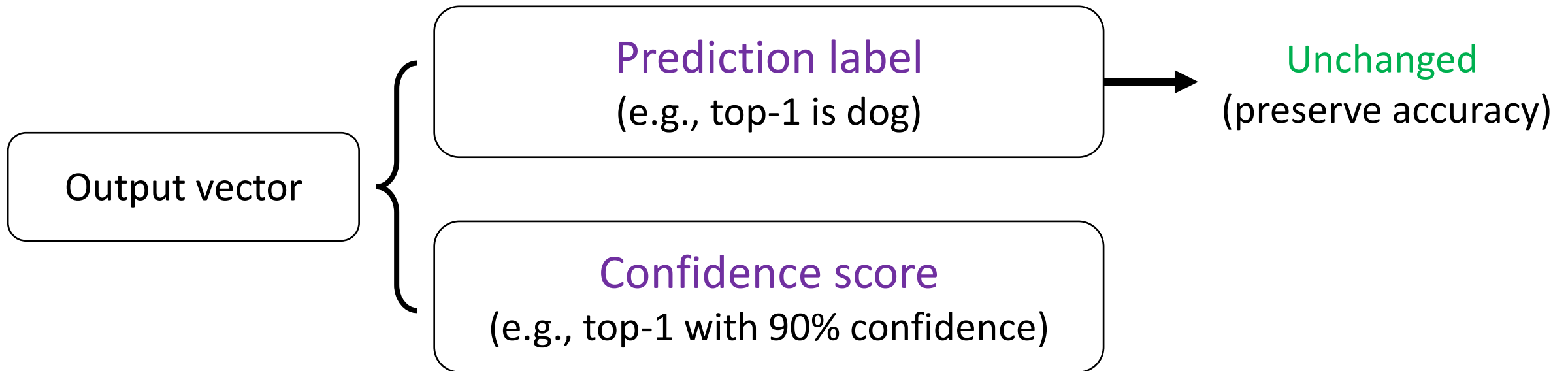
HAMP – Testing-time defense

- ❑ Modify **all** output vectors → **low** confidence outputs.



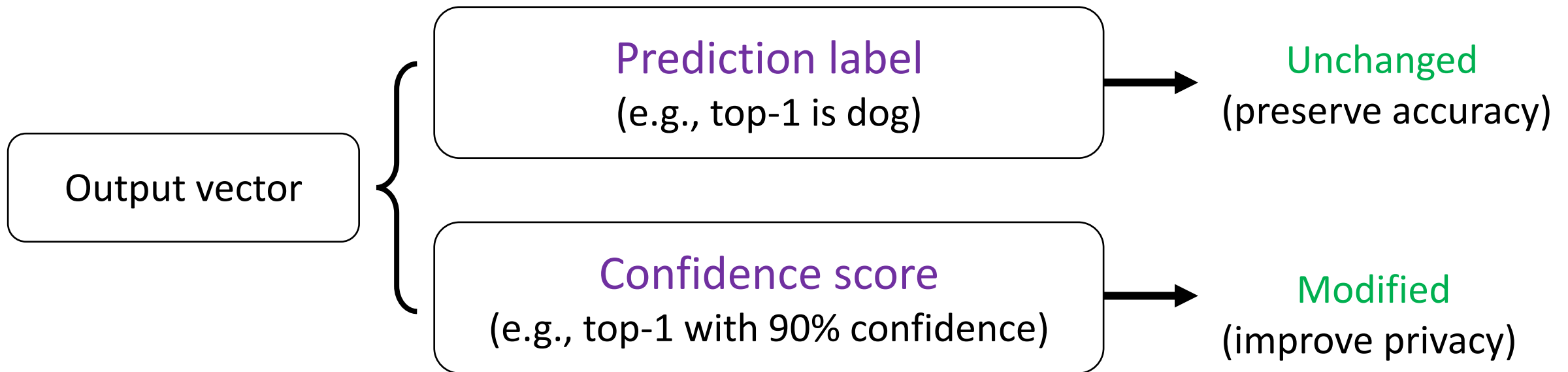
HAMP – Testing-time defense

- ❑ Modify **all** output vectors → **low** confidence outputs.



HAMP – Testing-time defense

- ❑ Modify **all** output vectors → **low** confidence outputs.



HAMP – Testing-time defense

- ❑ Modify **all** output vectors → **low** confidence outputs.
- ❑ How to obtain low confidence outputs?

HAMP – Testing-time defense

- ❑ Modify **all** output vectors → **low** confidence outputs.
- ❑ How to obtain low confidence outputs?
 - **Utilize random samples as (highly probable) non-members.**

HAMP – Testing-time defense

- ❑ Modify **all** output vectors → **low** confidence outputs.
- ❑ How to obtain low confidence outputs?
 - **Utilize random samples as (highly probable) non-members.**

Simple ✓
(optimization-free)

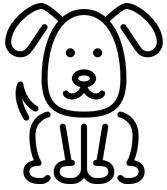
HAMP – Testing-time defense

- ❑ Modify **all** output vectors → **low** confidence outputs.
- ❑ How to obtain low confidence outputs?
 - **Utilize random samples as (highly probable) non-members.**

Simple ✓
(optimization-free)

Effective ✓
(improve membership privacy)

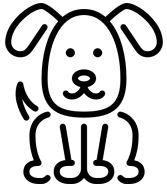
Output modification with random samples



Dog	Cat
0.85	0.15

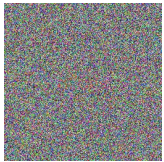
member sample

Output modification with random samples



Dog	Cat
0.85	0.15

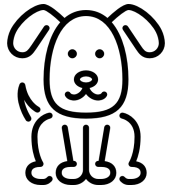
member sample



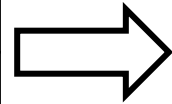
Dog	Cat
0.45	0.55

random sample

Output modification with random samples

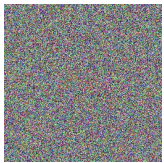


Dog	Cat
0.85	0.15



Keep prediction label
(top-1 → dog)

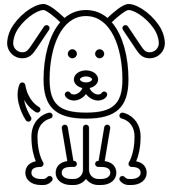
member sample



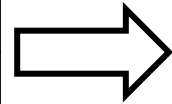
Dog	Cat
0.45	0.55

random sample

Output modification with random samples

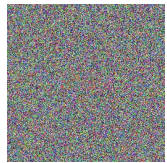


Dog	Cat
0.85	0.15

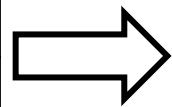


Keep prediction label
(top-1 → dog)

member sample



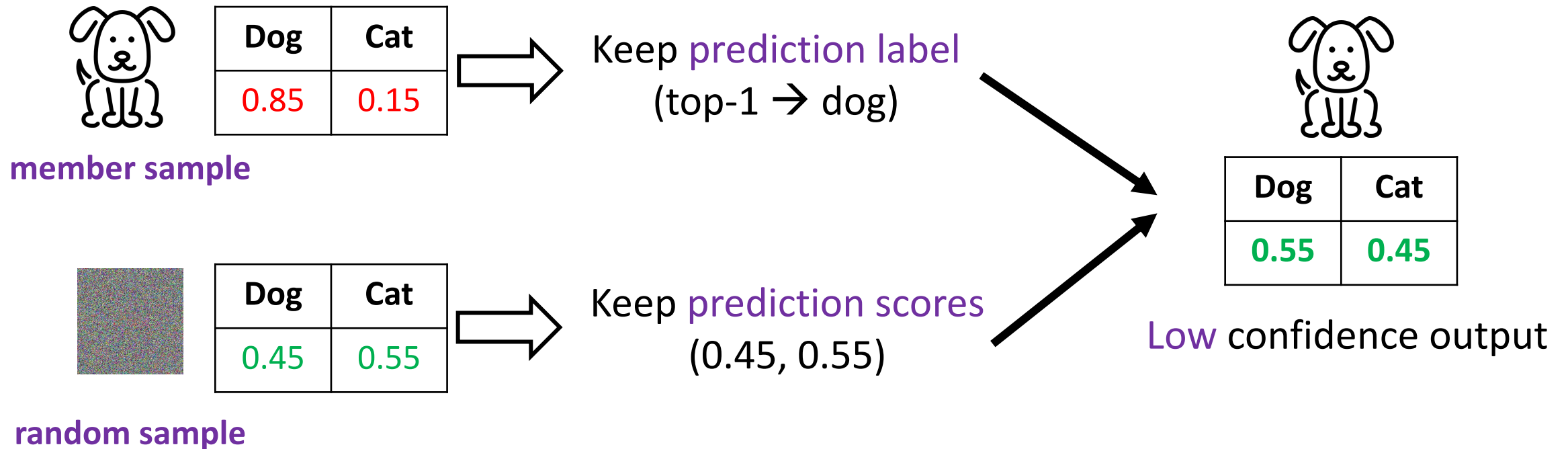
Dog	Cat
0.45	0.55



Keep prediction scores
(0.45, 0.55)

random sample

Output modification with random samples



Evaluation

Evaluation

5 datasets

Purchase100

Texas100

Location30

CIFAR10

CIFAR100

Evaluation

5 datasets

Purchase100
Texas100
Location30
CIFAR10
CIFAR100

9 attacks

NN-based
Loss-based
Entropy-based
Modified-entropy-based
Confidence-based
Likelihood-ratio attack (LiRA)

Correctness-based
Boundary-based
Augmentation-based

Evaluation

5 datasets

Purchase100
Texas100
Location30
CIFAR10
CIFAR100

9 attacks

NN-based
Loss-based
Entropy-based
Modified-entropy-based
Confidence-based
Likelihood-ratio attack (LiRA)
Correctness-based
Boundary-based
Augmentation-based

7 defenses

AdvReg (CCS'18)
MemGuard (CCS'19)
DMP (AAAI'21)
SELENA (USENIX'22)
Early stopping (USENIX'21)
Label Smoothing (CVPR'16)
DPSGD (CCS'16)

Evaluation

5 datasets

Purchase100
Texas100
Location30
CIFAR10
CIFAR100

9 attacks

NN-based
Loss-based
Entropy-based
Modified-entropy-based
Confidence-based
Likelihood-ratio attack (LiRA)

Correctness-based
Boundary-based
Augmentation-based

7 defenses

AdvReg (CCS'18)
MemGuard (CCS'19)
DMP (AAAI'21)
SELENA (USENIX'22)
Early stopping (USENIX'21)
Label Smoothing (CVPR'16)
DPSGD (CCS'16)

HAMP configuration

α for high-entropy soft labels γ for regularization strength

Refer to the paper for details

Evaluation

5 datasets

Purchase100
Texas100
Location30
CIFAR10
CIFAR100

9 attacks

NN-based
Loss-based
Entropy-based
Modified-entropy-based
Confidence-based
Likelihood-ratio attack (LiRA)

Correctness-based
Boundary-based
Augmentation-based

7 defenses

AdvReg (CCS'18)
MemGuard (CCS'19)
DMP (AAAI'21)
SELENA (USENIX'22)
Early stopping (USENIX'21)
Label Smoothing (CVPR'16)
DPSGD (CCS'16)

HAMP configuration

α for high-entropy soft labels γ for regularization strength

Refer to the paper for details

2 metrics

TPR @ 0.1% FPR
TNR @ 0.1% FNR

Evaluation

5 datasets

- Purchase100
- Texas100
- Location30
- CIFAR10
- CIFAR100

9 attacks

- NN-based
- Loss-based
- Entropy-based
- Modified-entropy-based
- Confidence-based
- Likelihood-ratio attack (LiRA)**
- Correctness-based
- Boundary-based
- Augmentation-based

7 defenses

- AdvReg (CCS'18)
- MemGuard (CCS'19)
- DMP (AAAI'21)
- SELENA (USENIX'22)
- Early stopping (USENIX'21)
- Label Smoothing (CVPR'16)
- DPSGD (CCS'16)

HAMP configuration

α for high-entropy soft labels γ for regularization strength

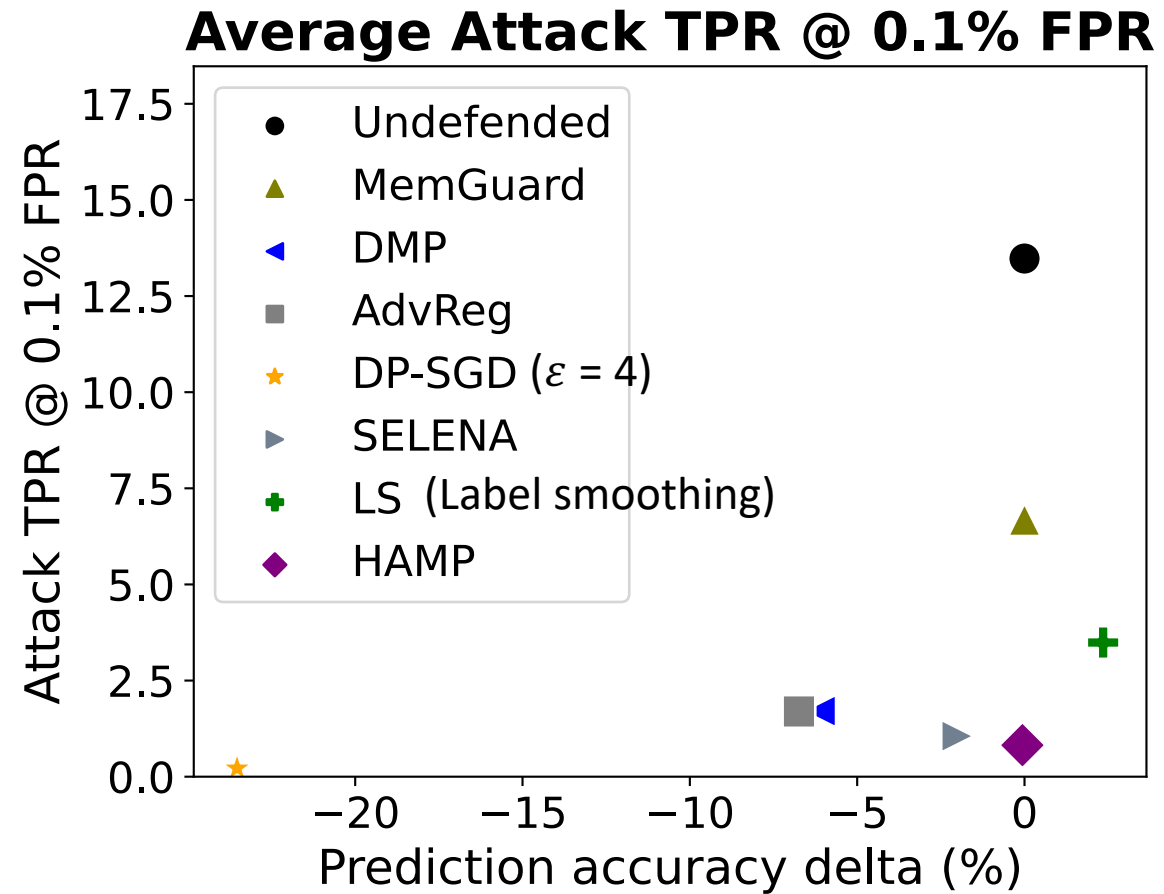
Refer to the paper for details

2 metrics

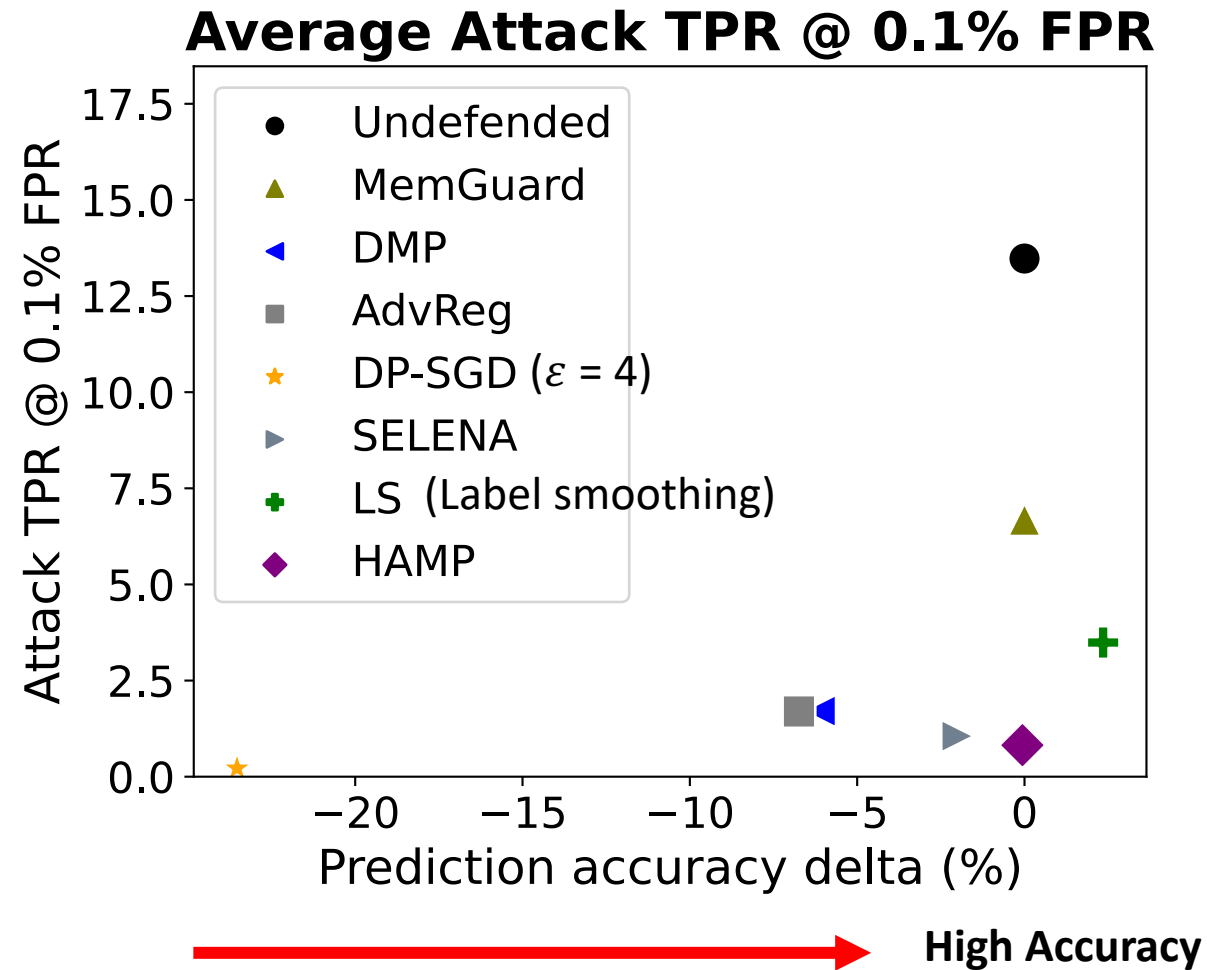
- TPR @ 0.1% FPR
- TNR @ 0.1% FNR

Artifact https://github.com/DependableSystemsLab/MIA_defense_HAMP

Key results

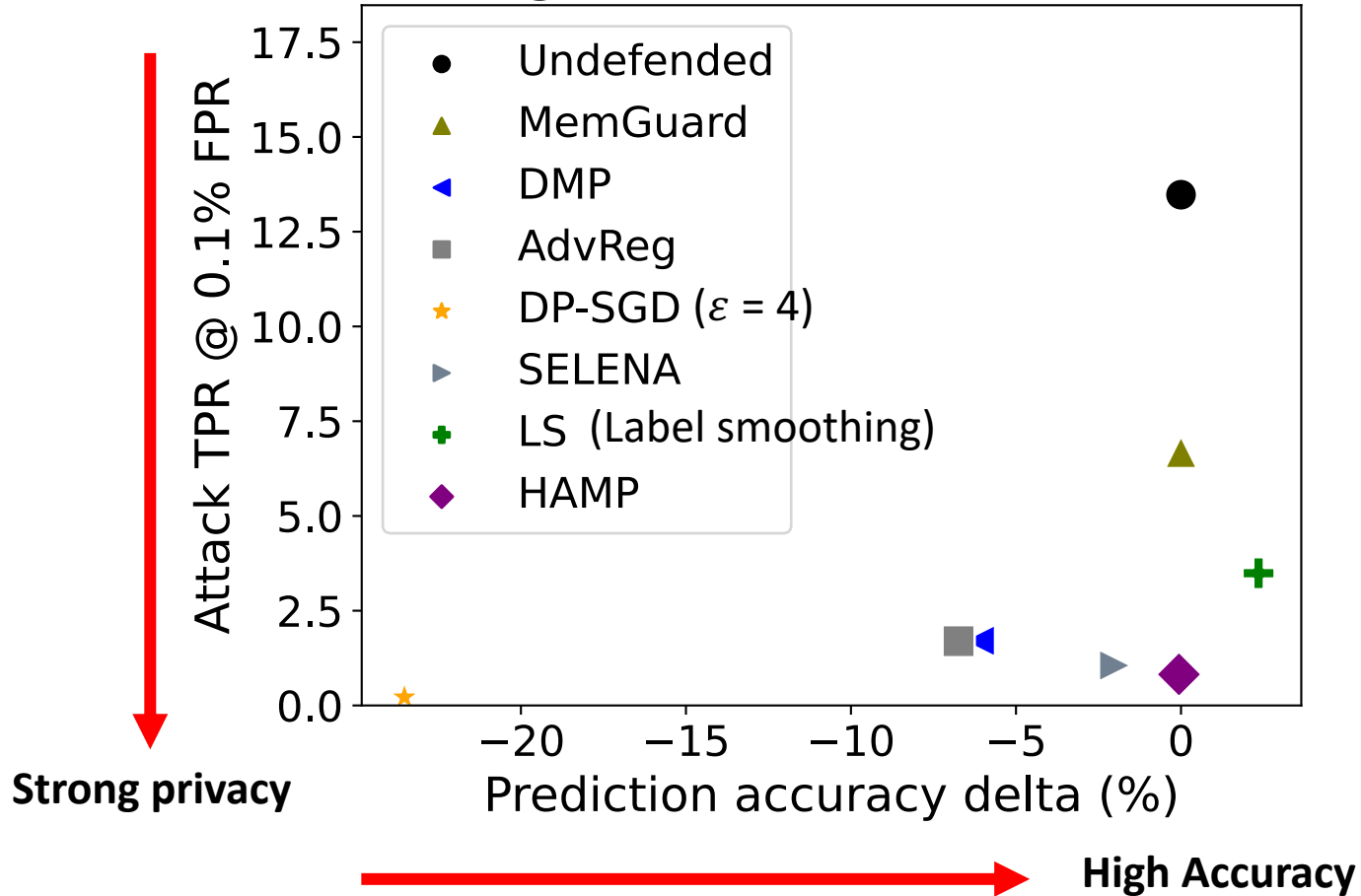


Key results

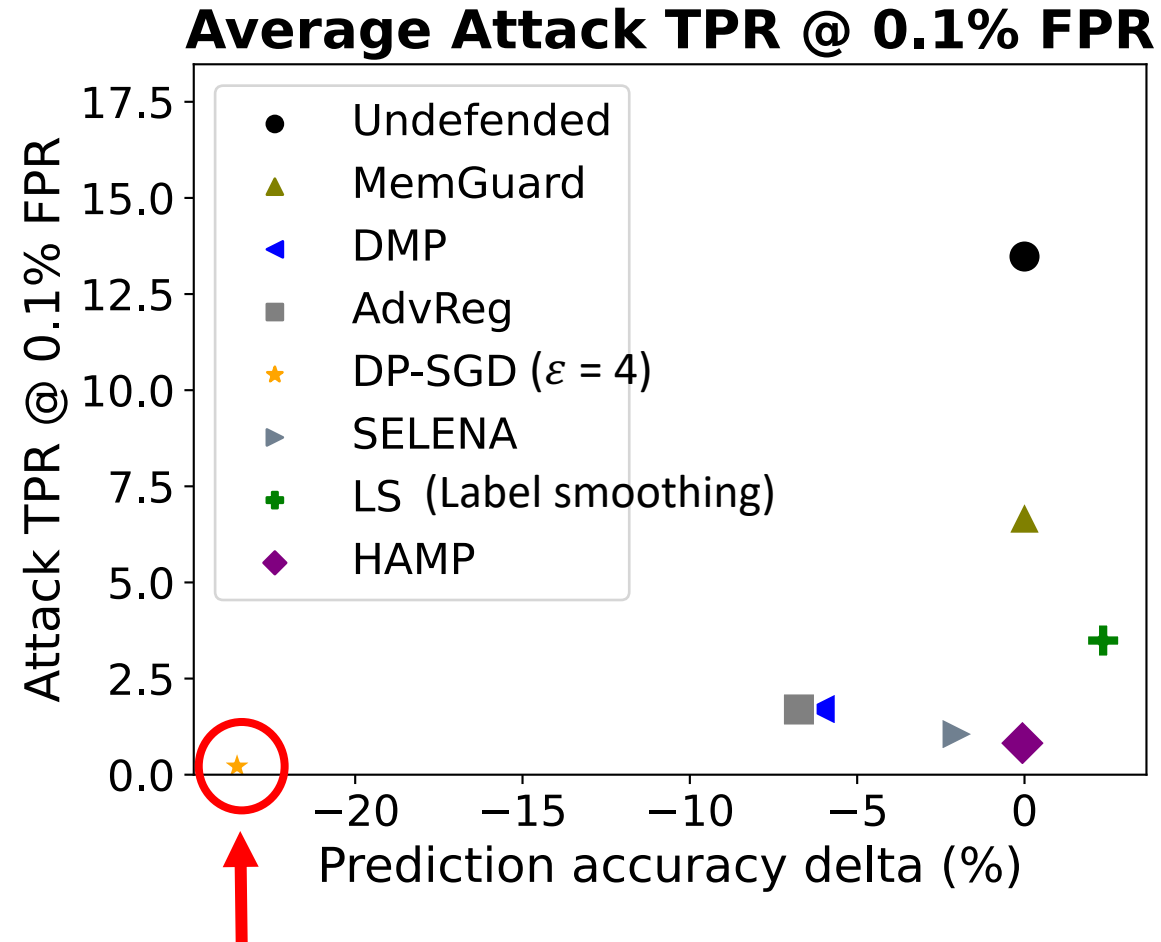


Key results

Average Attack TPR @ 0.1% FPR

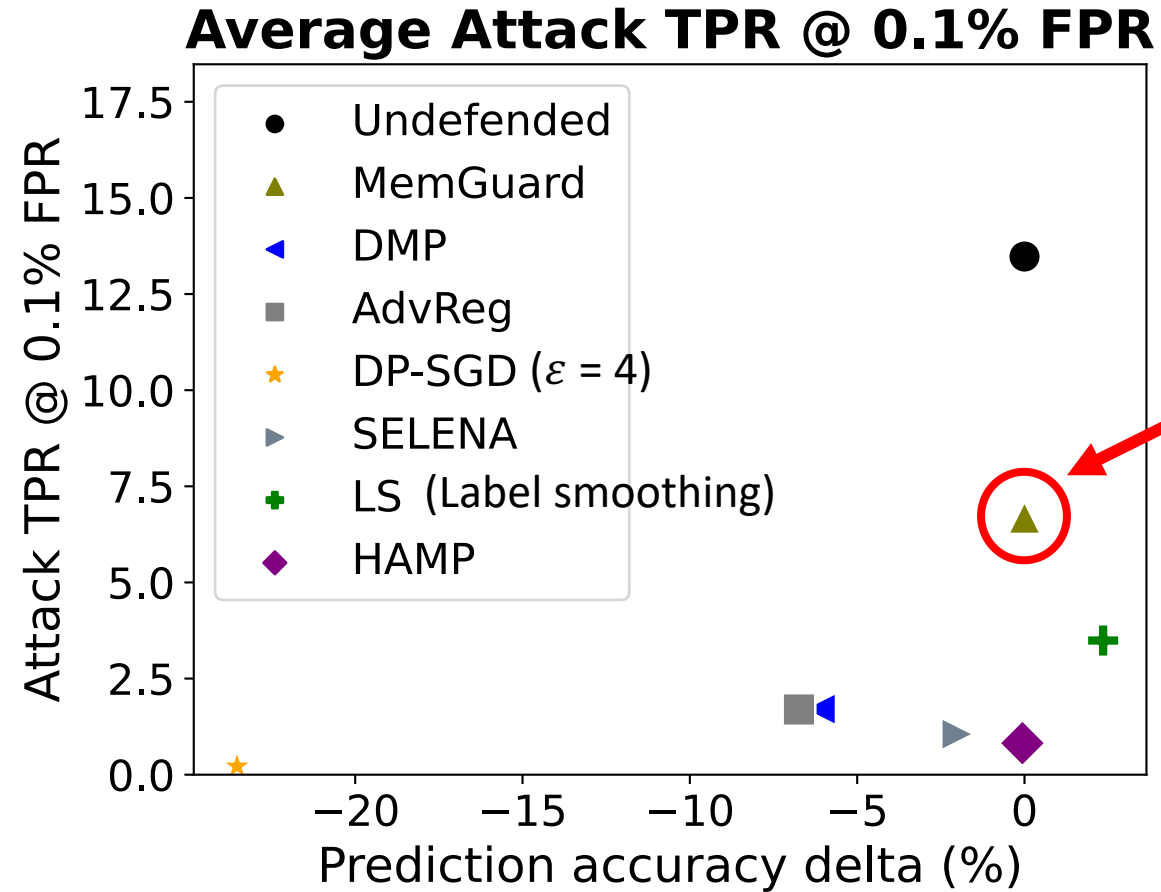


Key results



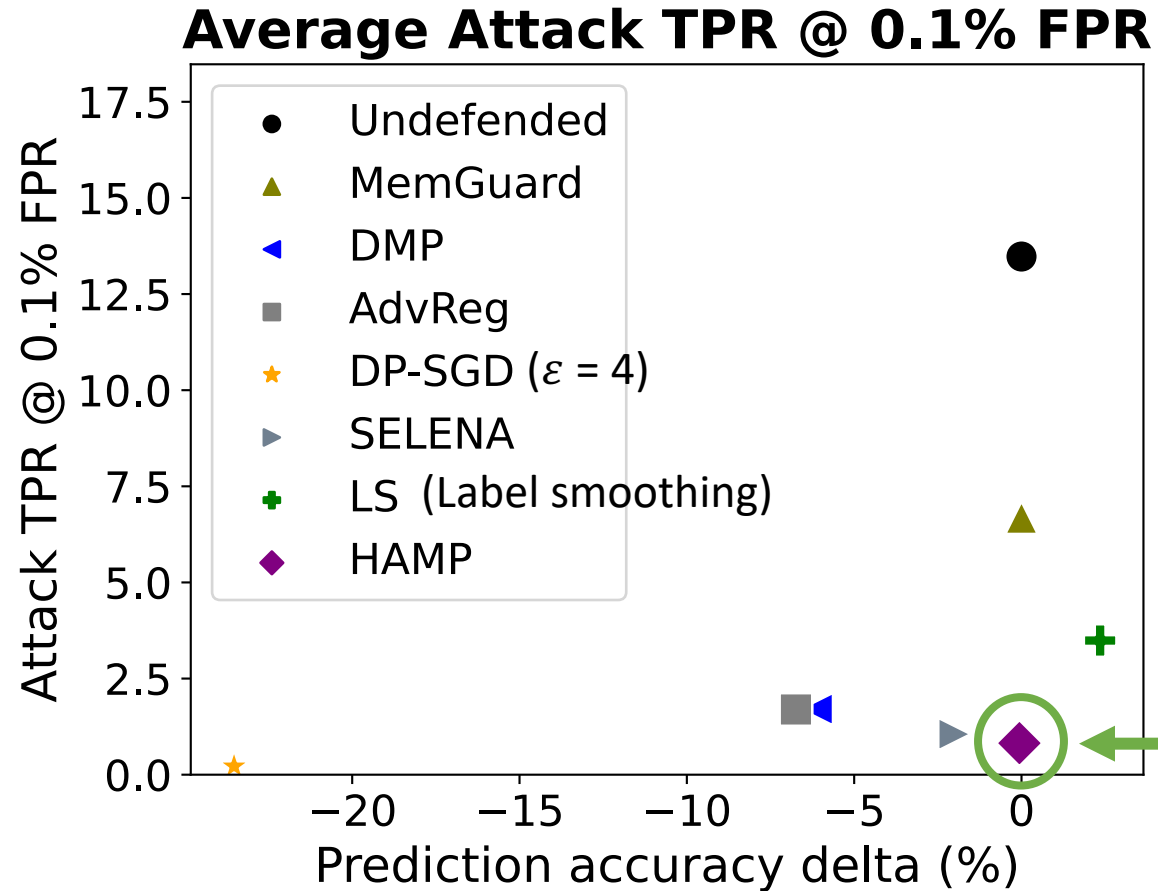
DPSGD: **Strong** privacy, but **low** accuracy

Key results



MemGuard: **High** accuracy,
but **poor** privacy

Key results



HAMP
Strong privacy: attack TPR \downarrow 94%
High accuracy: 0.46% accuracy drop

Summary



How to mitigate **membership inference attacks** with strong privacy protection and low accuracy drop?



Summary



How to mitigate **membership inference attacks** with strong privacy protection and low accuracy drop?



Mitigating ML model's overconfident prediction on training samples without jeopardizing model accuracy.



Summary



How to mitigate **membership inference attacks** with strong privacy protection and low accuracy drop?



Mitigating ML model's overconfident prediction on training samples without jeopardizing model accuracy.



HAMP: A new way to combine soft label training, training regularization and output modification for privacy-preserving training!

Paper



Code



zitaoc@ece.ubc.ca

Ablation study

TABLE II: Ablation study on different components of HAMP:

①: High-entropy soft labels; ②: Entropy-based regularizer;
③: Testing-time output modification.

Defense component	Training accuracy	Testing accuracy	Attack TPR @0.1% FPR	Attack TNR @0.1% FNR
None (undefended)	99.36	80.85	14.37	14.62
①	94.58	81.75	4.76	4.22
②	98.06	81.10	3.39	4.19
③	99.36	80.85	8.51	5.34
① + ②	91.12	81.15	1.86	1.07
① + ③	94.58	81.75	0.82	1.23
② + ③	98.06	81.10	2.90	3.76
① + ② + ③ (full defense)	91.12	81.15	0.40	0.44

Full defense has the **best privacy protection** with **competitive accuracy**

Comparison with DPSGD

