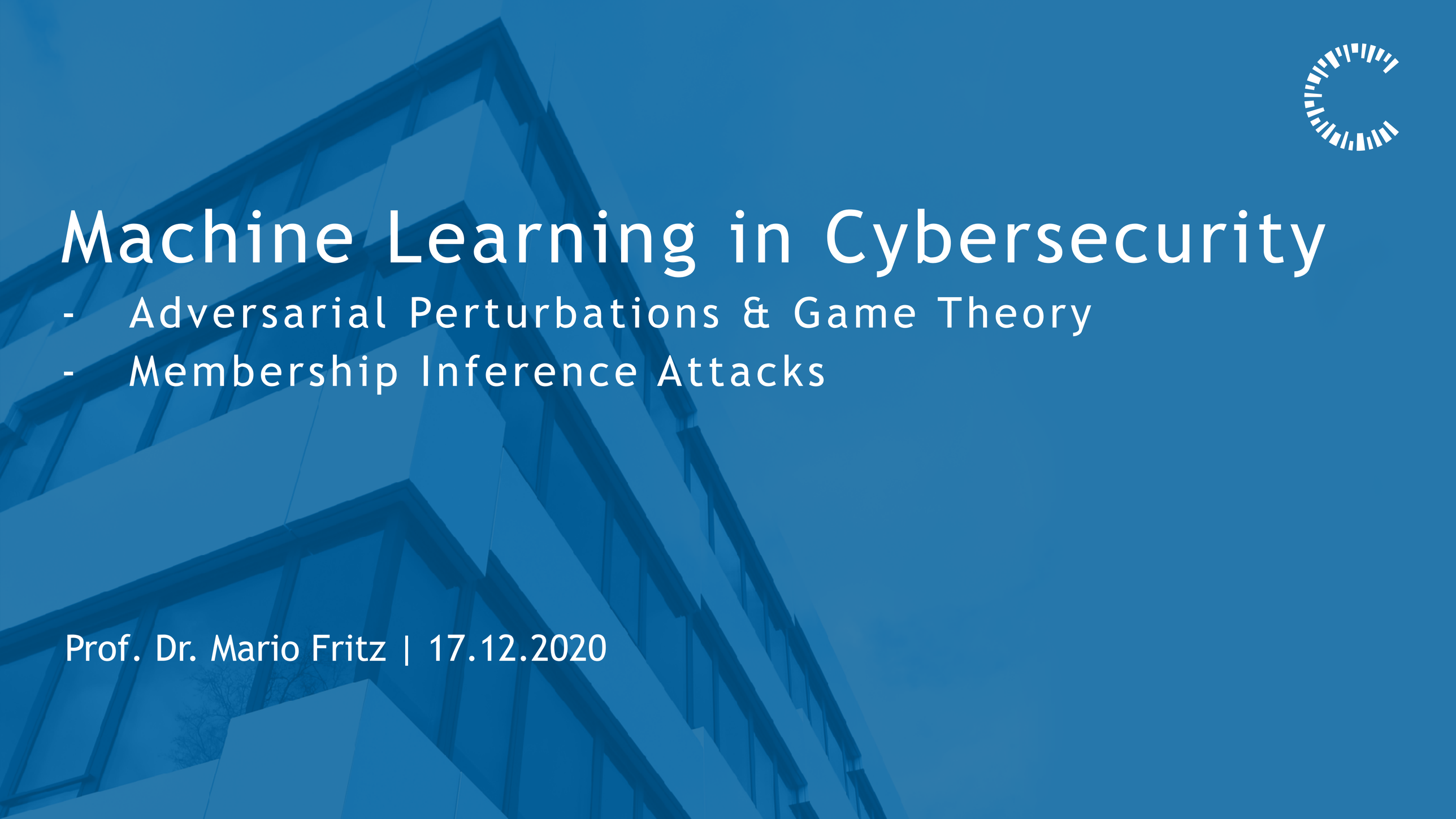


Machine Learning in Cybersecurity

- Adversarial Perturbations & Game Theory
- Membership Inference Attacks

Prof. Dr. Mario Fritz | 17.12.2020

Responsibility

- Act Responsible

- Differentiate between “what can be done” and “what should be done”

Main Conferences

- Security & Privacy
  - S&P (Oakland), CCS, Usenix Security, NDSS
- ML
  - Neurips (Dec 8th), ICML, KDD
    - Workshop on Security in Machine Learning
      - https://secml2018.github.io
    - Workshop on Privacy Preserving Machine Learning
      - https://ppml-workshop.github.io/ppml/
- AI
  - AAAI, IJCAI
- CSRankings.org

Literature search

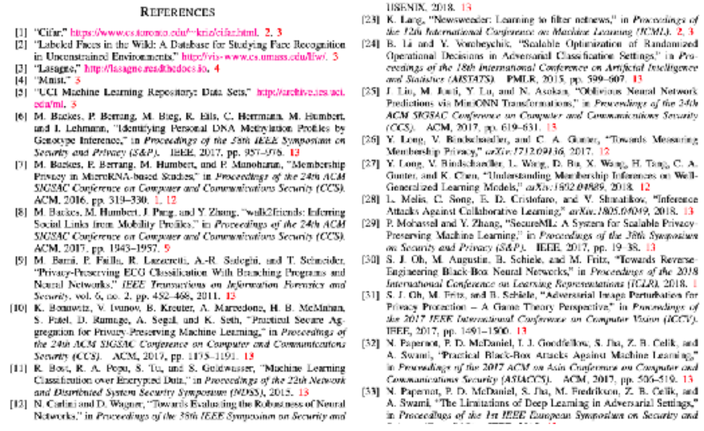
Past

Present

Future

Paper Bibliography

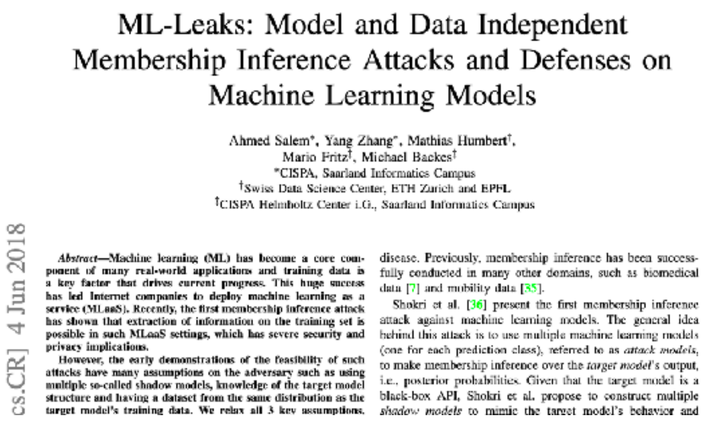
e.g. google scholar



REFERENCES

[1] "Cifar." https://www.cs.toronto.edu/~kriz/cifar.html. 2, 3
[2] "Labeled Faces in the Wild: A Database for Studying Face Recognition in Unconstrained Environments." http://vis-www.cs.umass.edu/lfw/. 3
[3] "Lasagne." http://lasagne.readthedocs.io. 4
[4] "Mnist." 3
[5] "UCI Machine Learning Repository: Data Sets." http://archive.ics.uci.edu/ml. 3
[6] M. Backes, P. Berrang, M. Bieg, R. Eils, C. Herrmann, M. Humbert, and I. Lehmann, "Identifying Personal DNA Methylation Profiles by Genotype Inference," in *Proceedings of the 38th IEEE Symposium on Security and Privacy (S&P)*. IEEE, 2017, pp. 957–976. 12
[7] M. Backes, P. Berrang, M. Humbert, and P. Manoharan, "Membership Privacy in MicroRNA-based Studies," in *Proceedings of the 24th ACM SIGSAC Conference on Computer and Communications Security (CCS)*. ACM, 2016, pp. 319–330. 1, 12
[8] M. Backes, M. Humbert, J. Pang, and Y. Zhang, "walk2friends: Inferring Social Links from Mobility Profiles," in *Proceedings of the 24th ACM SIGSAC Conference on Computer and Communications Security (CCS)*. ACM, 2017, pp. 1943–1957. 9
[9] M. Barni, P. Failla, R. Lazzeretti, A.-R. Sadeghi, and T. Schneider, "Privacy-Preserving ECG Classification With Branching Programs and Neural Networks," *IEEE Transactions on Information Forensics and Security*, vol. 6, no. 2, pp. 452–468, 2011. 12
[10] K. Bonawitz, V. Ivanov, B. Kreuter, A. Marcedone, H. B. McMahan, S. Patel, D. Ramage, A. Segal, and K. Seth, "Practical Secure Aggregation for Privacy-Preserving Machine Learning," in *Proceedings of the 24th ACM SIGSAC Conference on Computer and Communications Security (CCS)*. ACM, 2017, pp. 1175–1191. 13
[11] R. Bost, R. A. Popa, S. Tu, and S. Goldwasser, "Machine Learning Classification over Encrypted Data," in *Proceedings of the 22th Network and Distributed System Security Symposium (NDSS)*, 2015. 13
[12] N. Carlini and D. Wagner, "Towards Evaluating the Robustness of Neural Networks," in *Proceedings of the 38th IEEE Symposium on Security and* USENIX, 2018. 13
[23] K. Lang, "Newsweeder: Learning to filter netnews," in *Proceedings of the 12th International Conference on Machine Learning (ICML)*. 2, 3
[24] B. Li and Y. Vorobeychik, "Scalable Optimization of Randomized Operational Decisions in Adversarial Classification Settings," in *Proceedings of the 18th International Conference on Artificial Intelligence and Statistics (AISTATS)*. PMLR, 2015, pp. 599–607. 13
[25] J. Liu, M. Juuti, Y. Lu, and N. Asokan, "Oblivious Neural Network Predictions via MiniONN Transformations," in *Proceedings of the 24th ACM SIGSAC Conference on Computer and Communications Security (CCS)*. ACM, 2017, pp. 619–631. 13
[26] Y. Long, V. Bindschaedler, and C. A. Gunter, "Towards Measuring Membership Privacy," *arXiv:1712.09136*, 2017. 12
[27] Y. Long, V. Bindschaedler, L. Wang, D. Bu, X. Wang, H. Tang, C. A. Gunter, and K. Chen, "Understanding Membership Inferences on Well-Generalized Learning Models," *arXiv:1802.04889*, 2018. 12
[28] L. Melis, C. Song, E. D. Cristofaro, and V. Shmatikov, "Inference Attacks Against Collaborative Learning," *arXiv:1805.04049*, 2018. 13
[29] P. Mohassel and Y. Zhang, "SecureML: A System for Scalable Privacy-Preserving Machine Learning," in *Proceedings of the 38th Symposium on Security and Privacy (S&P)*. IEEE, 2017, pp. 19–38. 13
[30] S. J. Oh, M. Augustin, B. Schiele, and M. Fritz, "Towards Reverse-Engineering Black-Box Neural Networks," in *Proceedings of the 2018 International Conference on Learning Representations (ICLR)*, 2018. 1
[31] S. J. Oh, M. Fritz, and B. Schiele, "Adversarial Image Perturbation for Privacy Protection – A Game Theory Perspective," in *Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV)*. IEEE, 2017, pp. 1491–1500. 13
[32] N. Papernot, P. D. McDaniel, I. J. Goodfellow, S. Jha, Z. B. Celik, and A. Swami, "Practical Black-Box Attacks Against Machine Learning," in *Proceedings of the 2017 ACM on Asia Conference on Computer and Communications Security (ASIACCS)*. ACM, 2017, pp. 506–519. 13
[33] N. Papernot, P. D. McDaniel, S. Jha, M. Fredrikson, Z. B. Celik, and A. Swami, "The Limitations of Deep Learning in Adversarial Settings," in *Proceedings of the 1st IEEE European Symposium on Security and*

paper



cs.CR] 4 Jun 2018

ML-Leaks: Model and Data Independent Membership Inference Attacks and Defenses on Machine Learning Models

Ahmed Salem*, Yang Zhang*, Mathias Humbert†,
Mario Fritz‡, Michael Backes*
*CISPA, Saarland Informatics Campus
†Swiss Data Science Center, ETH Zurich and EPFL
‡CISPA Helmholtz Center i.G., Saarland Informatics Campus

Abstract—Machine learning (ML) has become a core component of many real-world applications and training data is a key factor that drives current progress. This huge success has led Internet companies to deploy machine learning as a service (MLaaS). Recently, the first membership inference attack has shown that extraction of information on the training set is possible in such MLaaS settings, which has severe security and privacy implications.

However, the early demonstrations of the feasibility of such attacks have many assumptions on the adversary such as using multiple so-called shadow models, knowledge of the target model structure and having a dataset from the same distribution as the target model's training data. We relax all 3 key assumptions,

disease. Previously, membership inference has been successfully conducted in many other domains, such as biomedical data [7] and mobility data [35].

Shokri et al. [36] present the first membership inference attack against machine learning models. The general idea behind this attack is to use multiple machine learning models (one for each prediction class), referred to as *attack models*, to make membership inference over the *target model*'s output, i.e., posterior probabilities. Given that the target model is a black-box API, Shokri et al. propose to construct multiple *shadow models* to mimic the target model's behavior and

MI-leaks: Model and data independent membership inference attacks and defenses on machine learning models
A Salem, Y Zhang, M Humbert, P Berrang… - arXiv preprint arXiv …, 2018 - arxiv.org
Machine learning (ML) has become a core component of many real-world applications and training data is a key factor that drives current progress. This huge success has led Internet companies to deploy machine learning as a service (MLaaS). Recently, the first membership …
☆ Cited by 43 Related articles All 8 versions

Privacy-preserving machine learning through data obfuscation
T Zhang, Z He, RB Lee - arXiv preprint arXiv:1807.01860, 2018 - arxiv.org
As machine learning becomes a practice and commodity, numerous cloud-based services and frameworks are provided to help customers develop and deploy machine learning applications. While it is prevalent to outsource model training and serving tasks in the cloud …
☆ Cited by 10 Related articles All 3 versions

Mlcapsule: Guarded offline deployment of machine learning as a service
L Hanzlik, Y Zhang, K Grosse, A Salem… - arXiv preprint arXiv …, 2018 - arxiv.org
With the widespread use of machine learning (ML) techniques, ML as a service has become increasingly popular. In this setting, an ML model resides on a server and users can query the model with their data via an API. However, if the user's input is sensitive, sending it to the …
☆ Cited by 15 Related articles All 5 versions

Comprehensive privacy analysis of deep learning: Passive and active white-box inference attacks against centralized and federated learning
M Nasr, R Shokri, A Houmansadr - 2019 IEEE Symposium on …, 2019 - ieeexplore.ieee.org
Deep neural networks are susceptible to various inference attacks as they remember information about their training data. We design white-box inference attacks to perform a comprehensive privacy analysis of deep learning models. We measure the privacy leakage …
☆ Cited by 6 Related articles

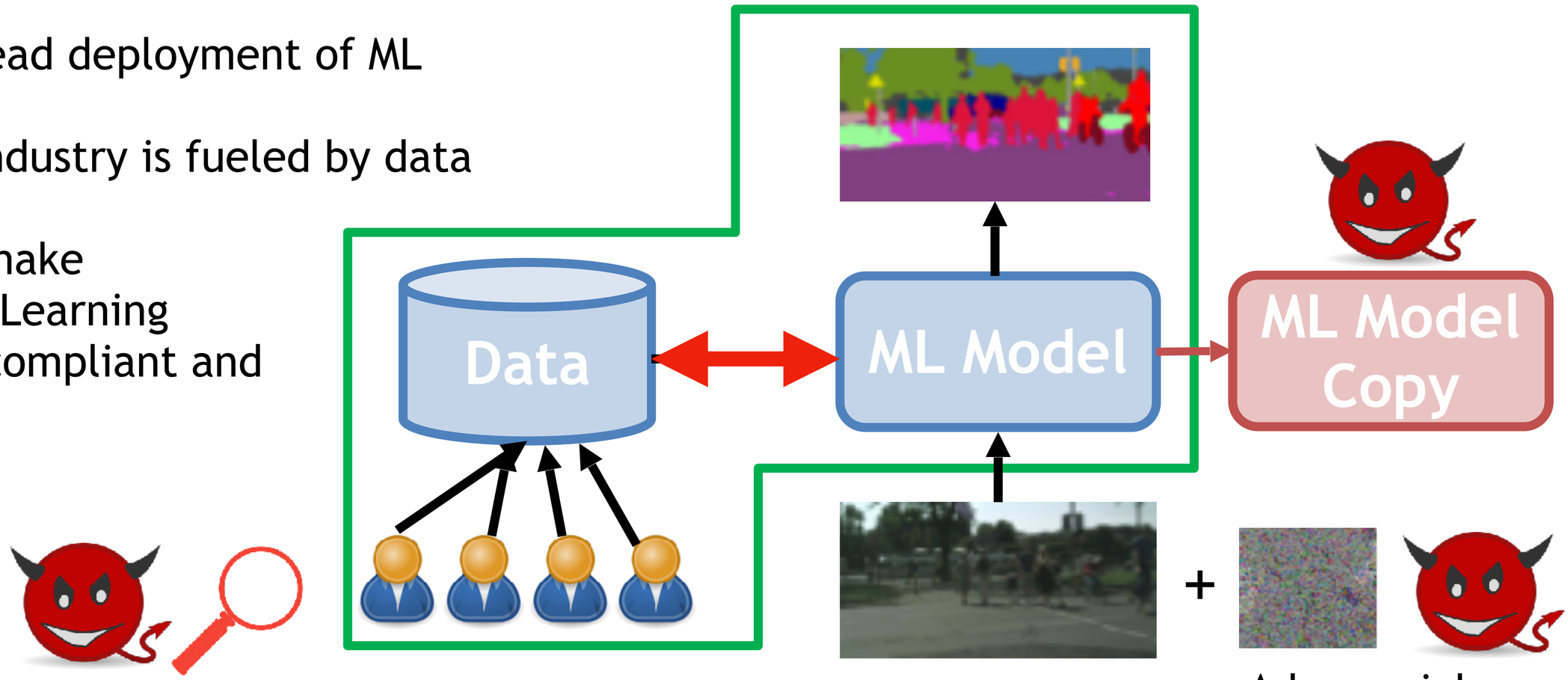
Knockoff nets: Stealing functionality of black-box models
T Orekondy, B Schiele, M Fritz - Proceedings of the IEEE …, 2019 - openaccess.thecvf.com
Abstract Machine Learning (ML) models are increasingly deployed in the wild to perform a wide range of tasks. In this work, we ask to what extent can an adversary steal functionality of such "victim" models based solely on blackbox interactions: image in, predictions out. In …
☆ Cited by 13 Related articles All 7 versions

Attributing fake images to GANs: Analyzing fingerprints in generated images
N Yu, L Davis, M Fritz - arXiv preprint arXiv:1811.08180, 2018 - arxiv.org
Research in computer graphics has seen in pursuit of realistic image generation for a long time. Recent advances in machine learning with deep generative models have shown increasing success of closing the realism gap by using data-driven and learned …
☆ Cited by 8 Related articles All 2 versions

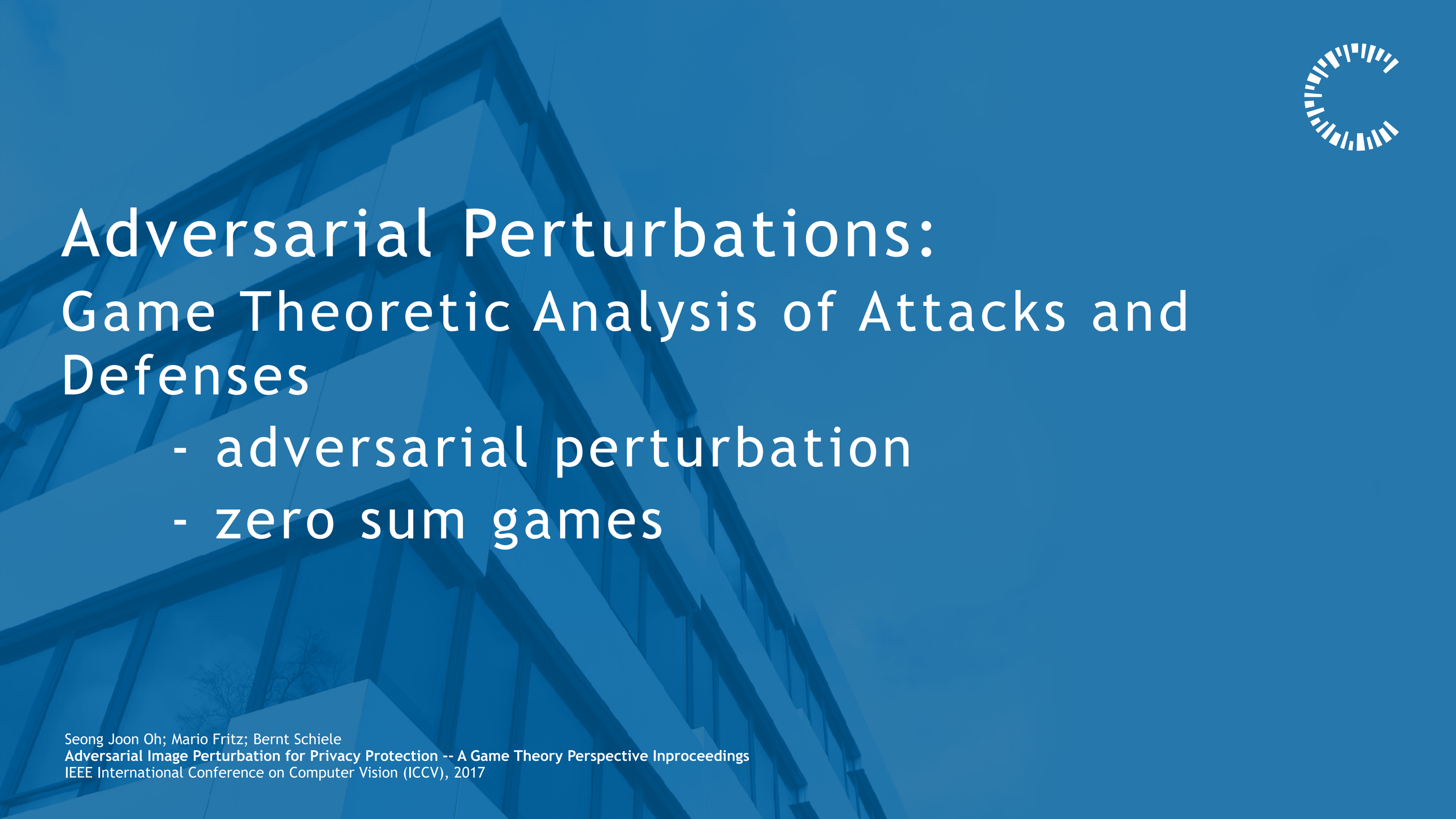
High-fidelity extraction of neural network models
M Jagielski, N Carlini, D Berthelot, A Kurakin… - arXiv preprint arXiv …, 2019 - arxiv.org
Model extraction allows an adversary to steal a copy of a remotely deployed machine learning model given access to its predictions. Adversaries are motivated to mount such attacks for a variety of reasons, ranging from reducing their computational costs, to …
☆ Cited by 3 All 3 versions

Attacks on ML and Privacy

- Widespread deployment of ML
- Future industry is fueled by data
- How to make Machine Learning privacy compliant and secure?



- **Membership Inference**
- Linkability Attack



Adversarial Perturbations:
Game Theoretic Analysis of Attacks and Defenses

- adversarial perturbation
- zero sum games

Seong Joon Oh; Mario Fritz; Bernt Schiele
Adversarial Image Perturbation for Privacy Protection -- A Game Theory Perspective Inproceedings
IEEE International Conference on Computer Vision (ICCV), 2017

Attacking for the good

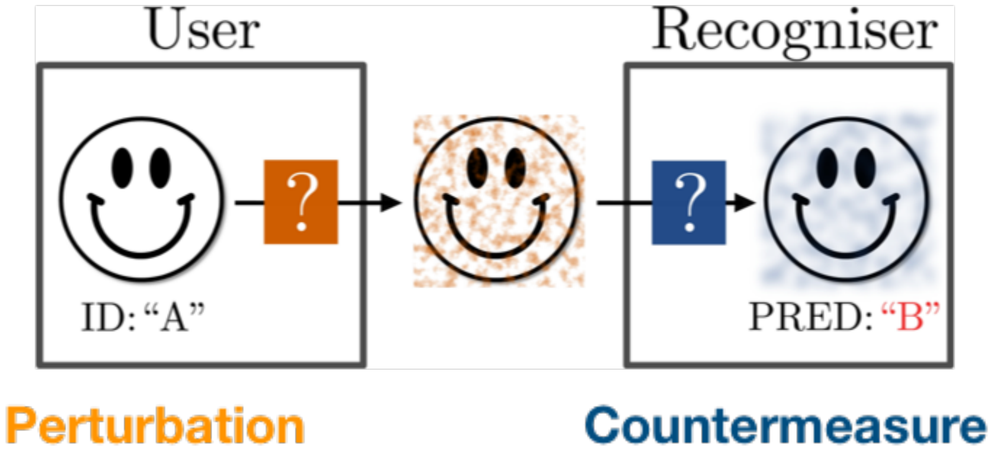
- Adversarial perturbations as privacy protection

- Pros
  - Imperceptible
  - Effective (traditional obfuscation techniques are not
- Cons
  - Model specific

- Game Theoretic analysis can help in attack/ defense/situation

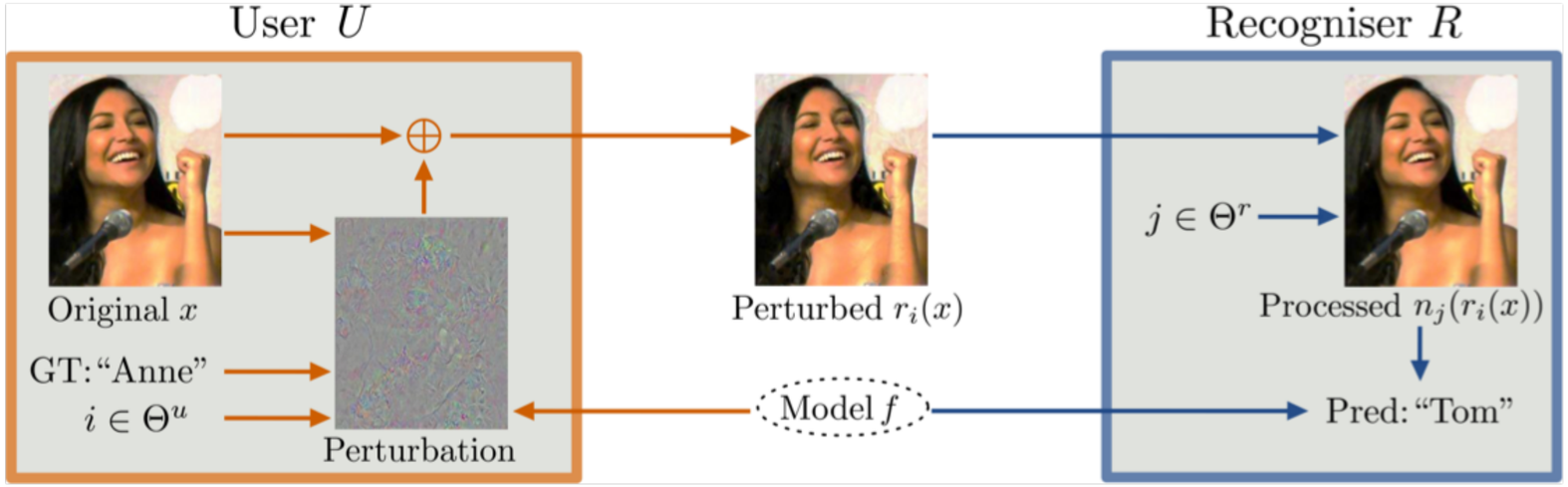
Adversarial Image Perturbations

Motivation: Game Theoretic Analysis



- Perturbation techniques and counter measures rapidly evolving
- Counter measures: blur, crop, jpg encoding
- Goal:
  - Analysis of action space
  - Upper bound on recognition rate

Model: Zero-Sum Game



- Two players (User, Recogniser), antagonistic goals.
- Finite set of actions
- Strategy is sampled following an unknown distribution.

- Principled formulation of opponent-independent guarantee.
- Assumption: Known, fixed recognition model (white box)

Model: Zero-Sum Game

- User (U) strategy

Defense-specific AIPs, Diverse set of AIP algorithms, AIP generation module weights (infinite space), …

- Recogniser (R) strategy

Defense mechanism, Neural net architectures, Neural net weights (infinite space), …

$i \in \Theta^u$

AIP 1

AIP 2

AIP 3

$j \in \Theta^r$

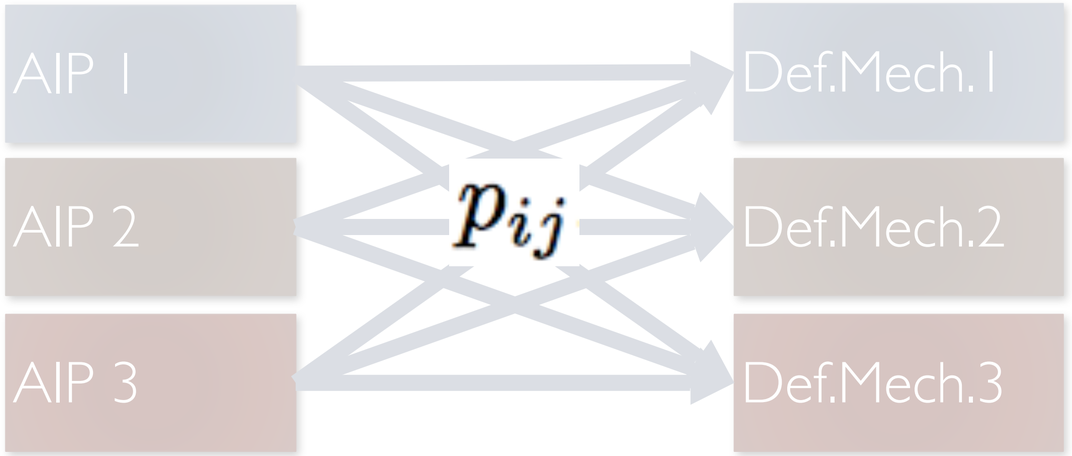
Def.Mech.1

Def.Mech.2

Def.Mech.3

Model: Zero-Sum Game

- Reward for R, p_{ij} , is the recognition rate when U plays i and R plays j:

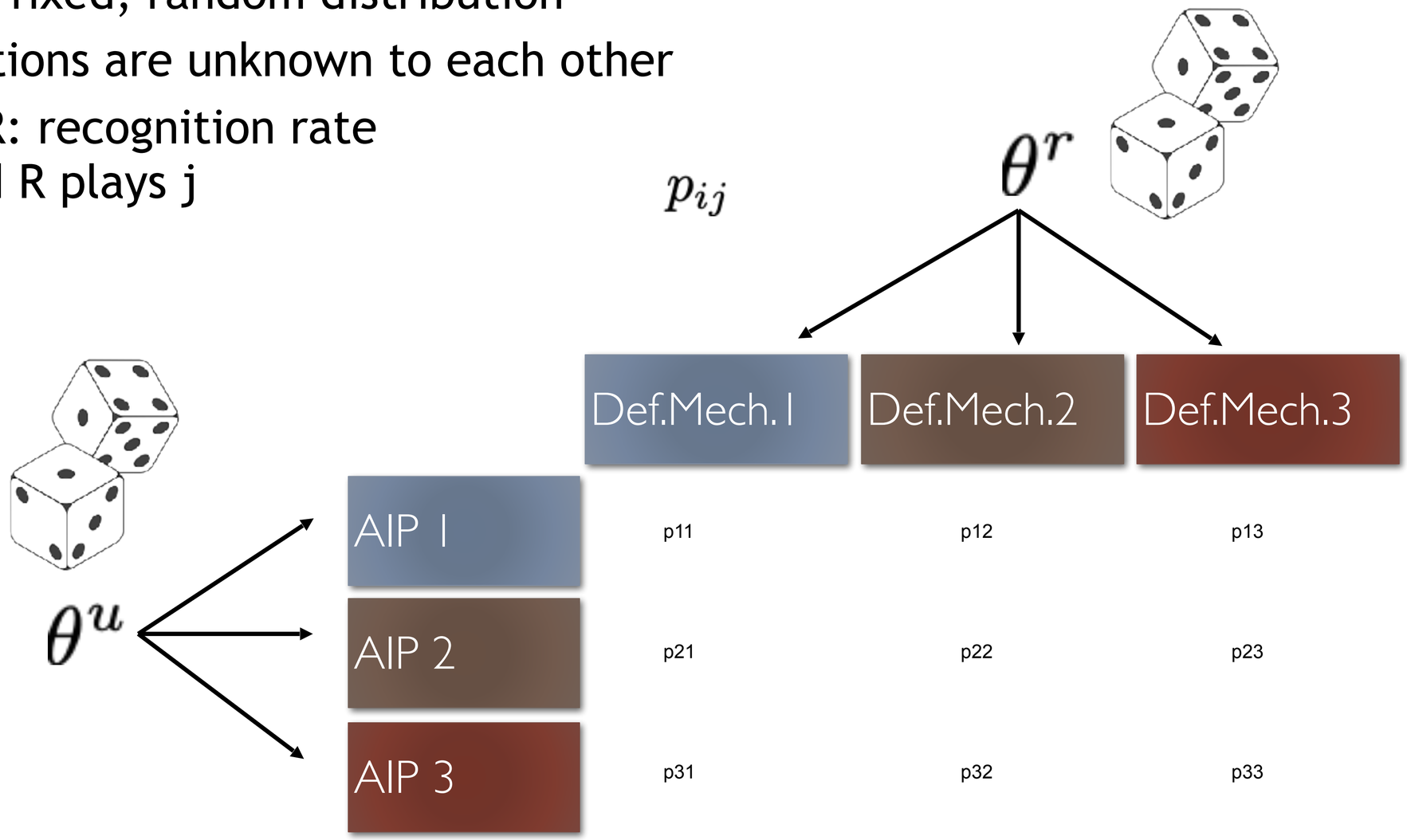


- Reward for U is the mis-recognition rate (ZSG):

$$1 - p_{ij}$$

Model: Zero-Sum Game

- Strategy is a fixed, random distribution
- The distributions are unknown to each other
- Reward for R: recognition rate
  U plays i and R plays j



Model: Zero-Sum Game

- R's expected reward:

$$\sum_{i,j} \theta_i^u \theta_j^r p_{ij}$$

- U's "optimal" strategy (also value of the game v):

$$\arg\min_{\theta^u} \max_{\theta^r} \sum_{i,j} \theta_i^u \theta_j^r p_{ij} \quad \text{s.t. } \theta^u, \theta^r \text{ are distributions.}$$

- If U plays optimal strategy, recognition rates is upper bounded by v

- Bound independent of R's strategy

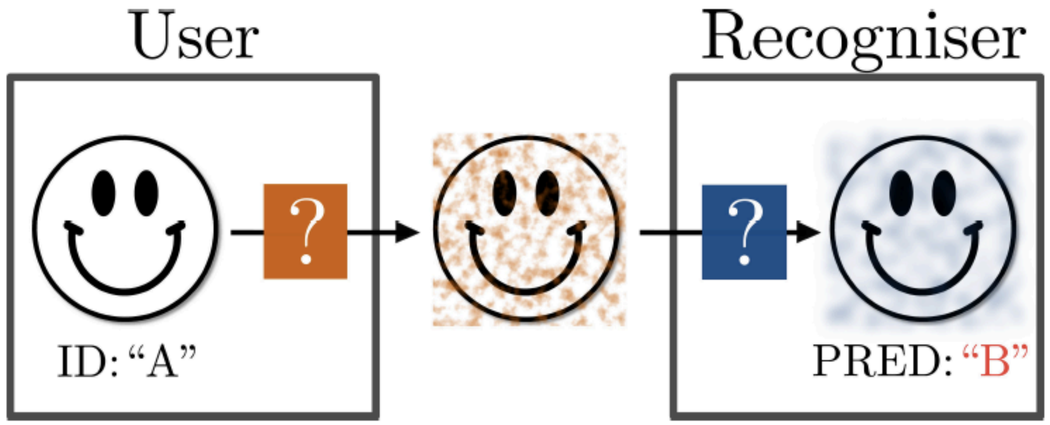
Actions: Attack-Defense-Game

- Attack:
  - Adversarial perturbations make the recognizer fail

- Defense against attack:
  - Adversarial training: train with perturbed examples

- Attack against defense:
  - Adversarial perturbations of the adversarially trained classifier

Case study: Action space



- User actions:

  AIP variants:

  {**GAMAN**, **GAMAN**$_{/T}$, **GAMAN**$_{/N}$, **GAMAN**$_{/B}$, **GAMAN**$_{/C}$, **GAMAN**$_{/TNBC}$}.

- Recogniser actions:

  Defense measures:

  Translation, Noise, Blur, Crop.

  [Graese *et al.* ICMLA'16]

  { f, f_T, f_N, f_B, f_C, f_{TNBC}}.

Results

User Θ^u	Recogniser Θ^r Proc	T	N	B	C	TNBC
GAMAN	4.0	6.6	15.0	22.2	16.7	9.9
/T	2.5	2.3	11.6	18.5	7.2	4.9
/N	5.8	7.6	4.6	23.6	16.6	9.1
/B	0.4	0.8	8.6	5.8	3.1	1.4
/C	2.6	2.2	11.8	18.1	3.4	4.3
/TNBC	0.7	0.9	5.2	9.5	3.2	2.0

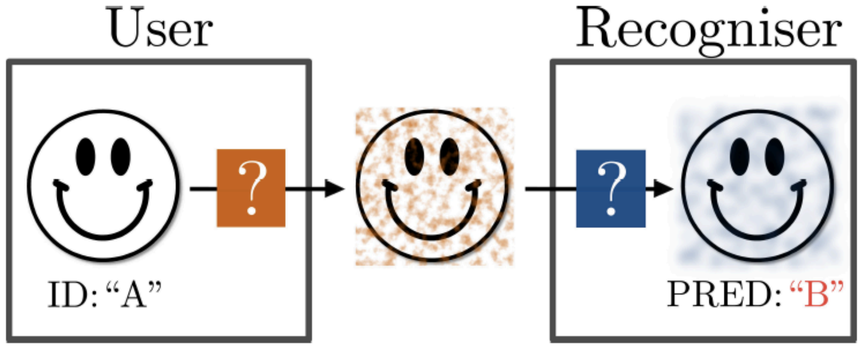
recogniser pay-off table

- Reference: Clean image: 91.1%, No-image: 0.8%)
- Best deterministic: 8.6% (GAMAN/B: 100%) (fN: 100%)
- Best randomised: guarantees < 7.4% expected recognition rate
  - Optimal U strategy: (GAMAN/B:61%, GAMAN/TNBC:39%).
  - Optimal R strategy: (fN:52%, fN:48%).

Conclusion

- AIP as an effective anonymisation tool for person recognition
- Robust AIP variants.

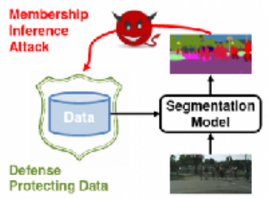
- Game theoretical framework:
  - makes assumptions explicit
  - accounts for the lack of knowledge on the deployed recogniser.
  - analysis tool for a design space
- “Selective Confusion” possible



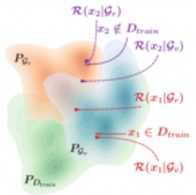
Membership Inference Attacks

Privacy Attacks and Privacy Preserving Machine Learning

Salem; Zhang; Humbert; **Fritz**; Backes
ML-**Leaks**: Model and Data Independent Membership Inference Attacks and Defenses on Machine Learning Models
NDSS'19

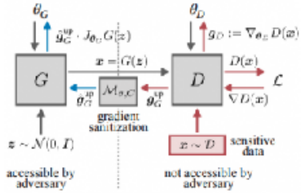


He; Rahimian; Schiele; **Fritz**
Segmentations-**Leak**: Membership Inference Attacks and Defenses in Semantic Image Segmentation
ECCV'20



Chen; Yu; Zhang; **Fritz**
GAN-**Leaks**: A Taxonomy of Membership Inference Attacks against GANs
CCS'20

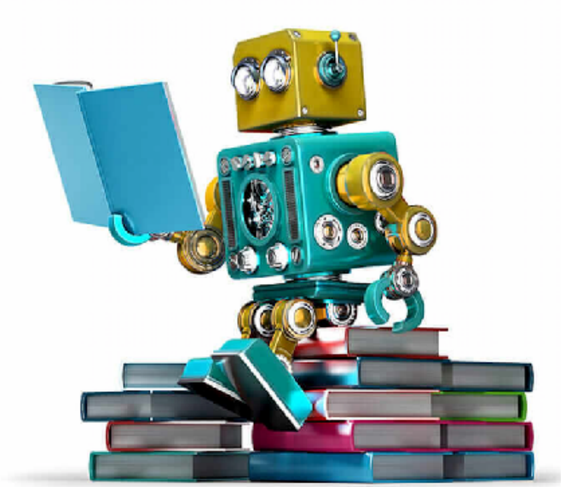
Salem; Bhattacharyya; Backes; **Fritz**; Zhang
Updates-**Leak**: Data Set Inference and Reconstruction Attacks in Online Learning
USENIX Security'20

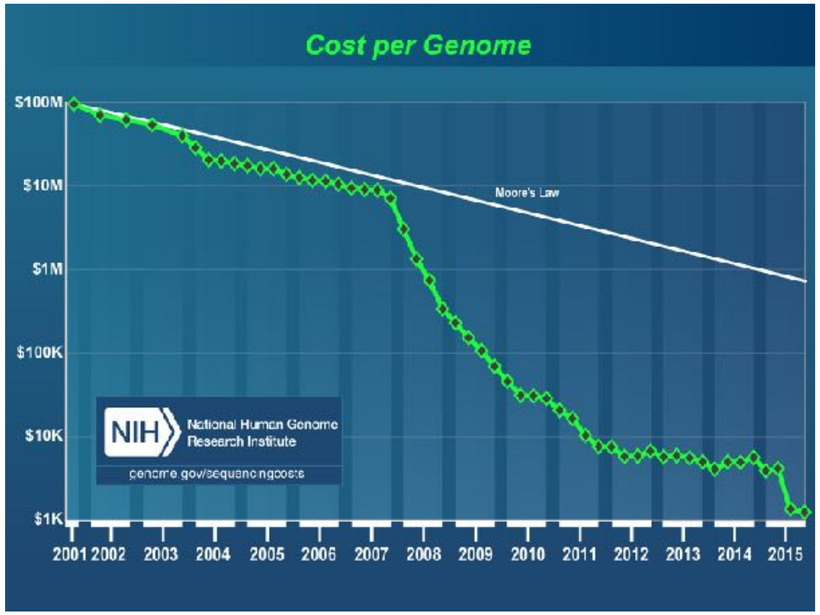


Chen; Orekondy; **Fritz**
GS-WGAN: A Gradient-Sanitized Approach for Learning Differentially Private Generators
NeurIPS'20

New Technologies

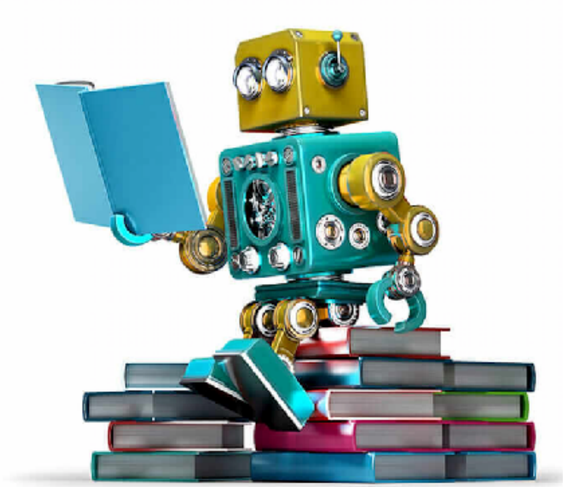
CISPA
HELMHOLTZ CENTER FOR INFORMATION SECURITY

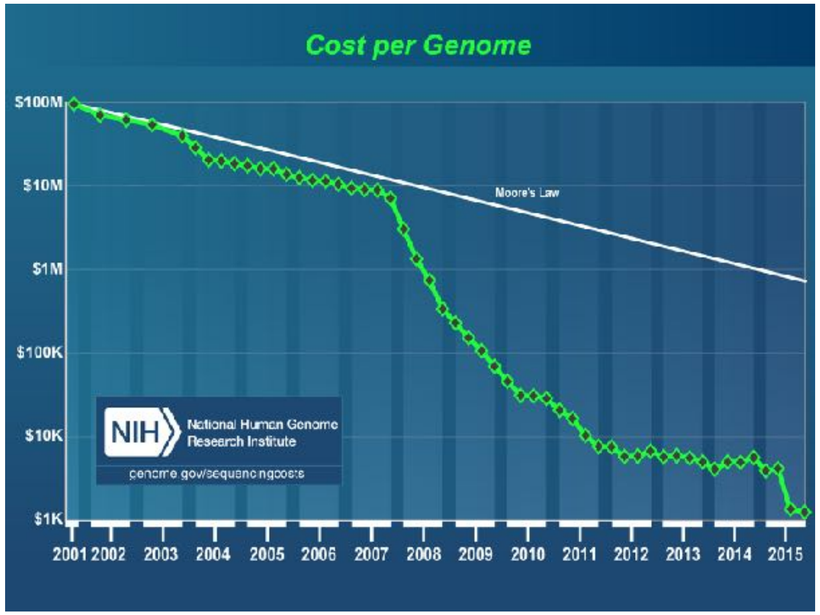




New Privacy Challenges!

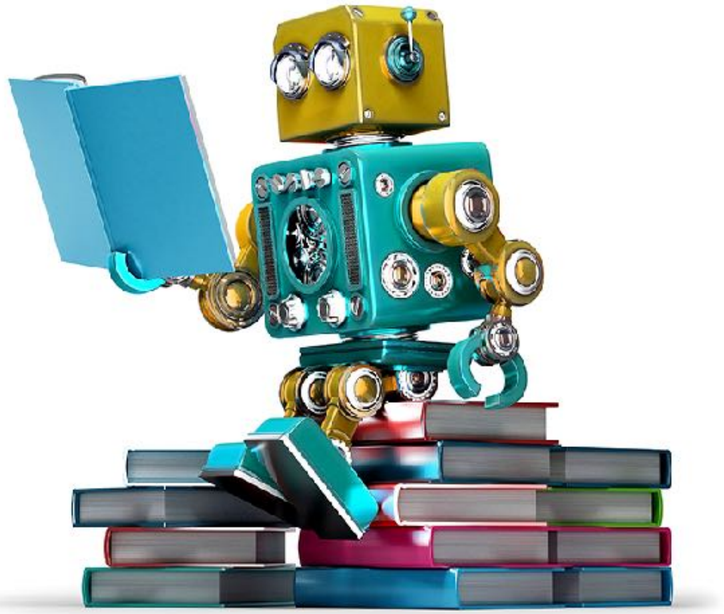
CISPA
HELMHOLTZ CENTER FOR INFORMATION SECURITY

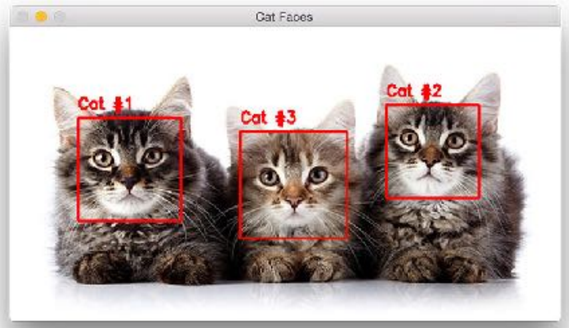




Era of Machine Learning

CISPA
HELMHOLTZ CENTER FOR INFORMATION SECURITY

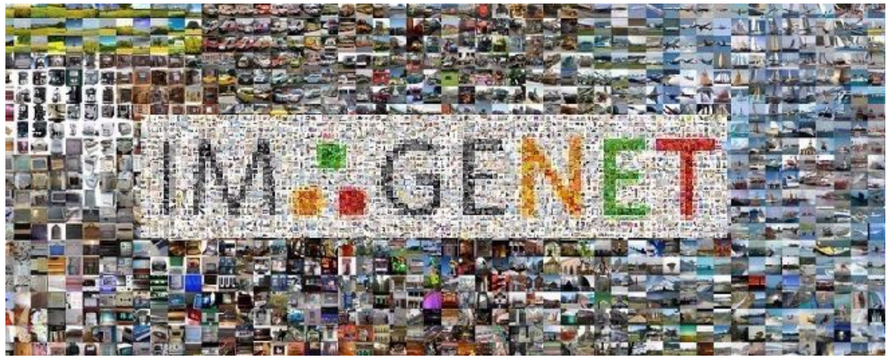




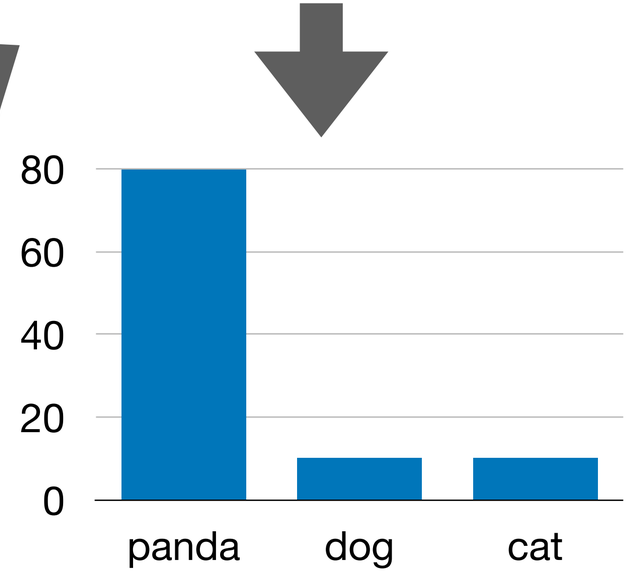
- ML models are often trained on sensitive data
  - Biomedical data
  - Keyboard input
  - Image / Language

How About Privacy?

Machine Learning Pipeline

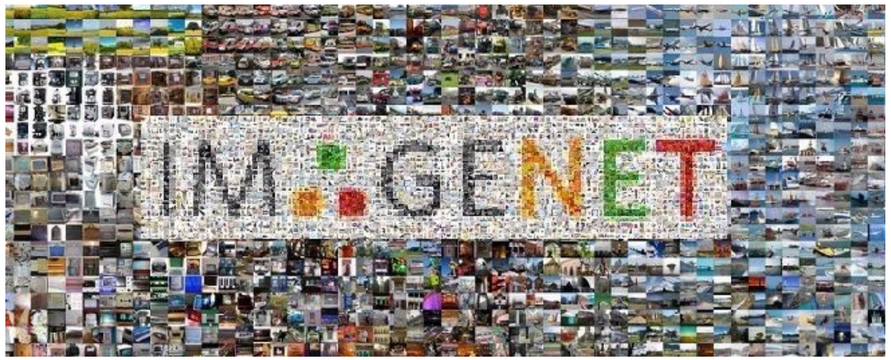


ML Model

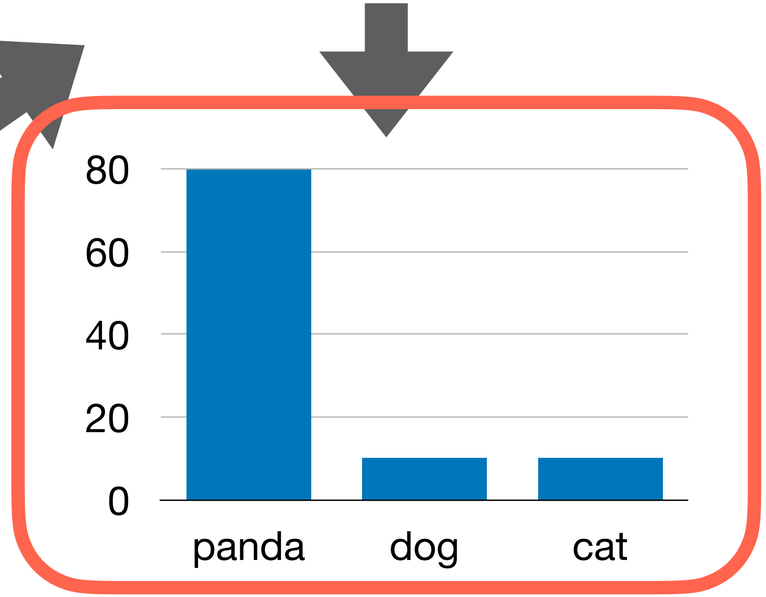


Membership Inference against Machine Learning

CISPA
HELMHOLTZ CENTER FOR INFORMATION SECURITY

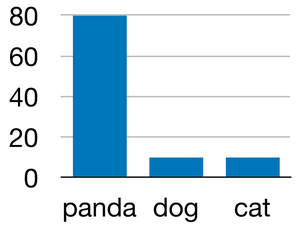


ML Model



Attack by Shokri et al.

CISPA HELMHOLTZ CENTER FOR INFORMATION SECURITY



Attack Models

member or non-member?

Multiple Attack Models

Overfitting leads to membership inference

Original Dataset

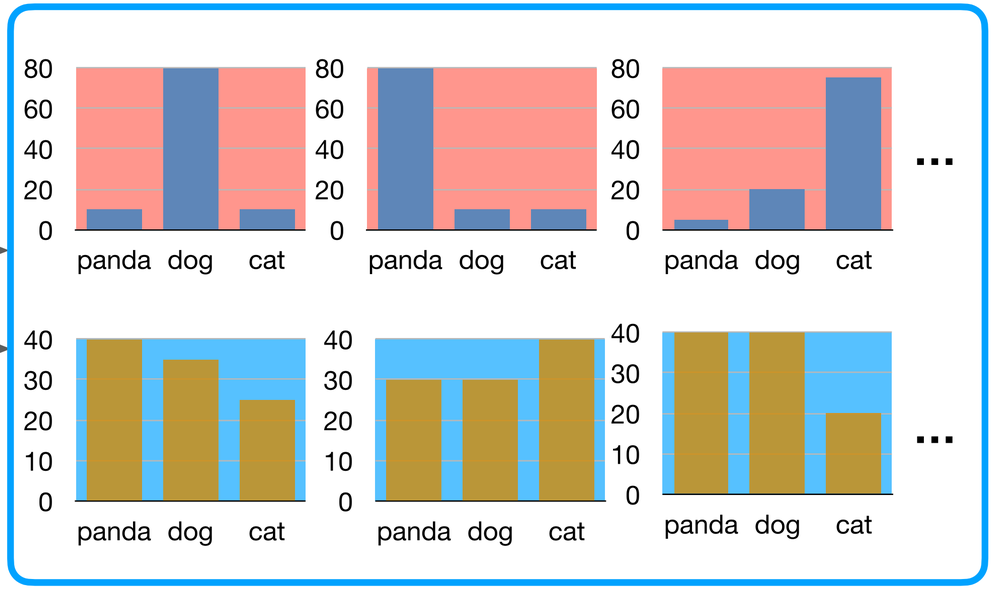
Dataset

Target Model

Train

Test

Shadow Models

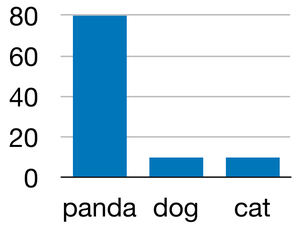


Same Distribution

Multiple Shadow Models

Attack 1

CISPA HELMHOLTZ CENTER FOR INFORMATION SECURITY



Original Dataset

Dataset

Train

Test

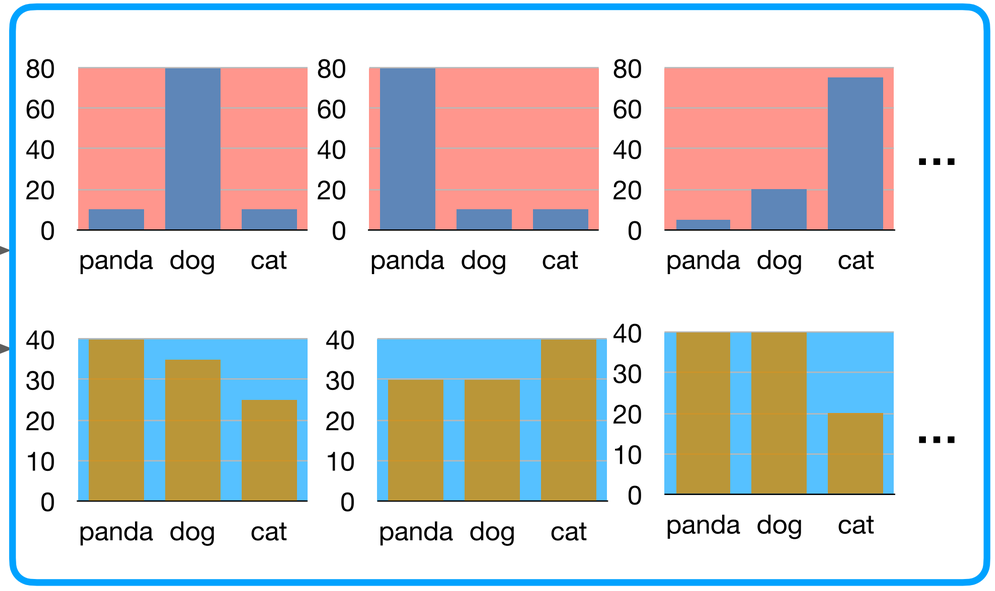
Target Model

Multiple Attack Models

member or non-member?

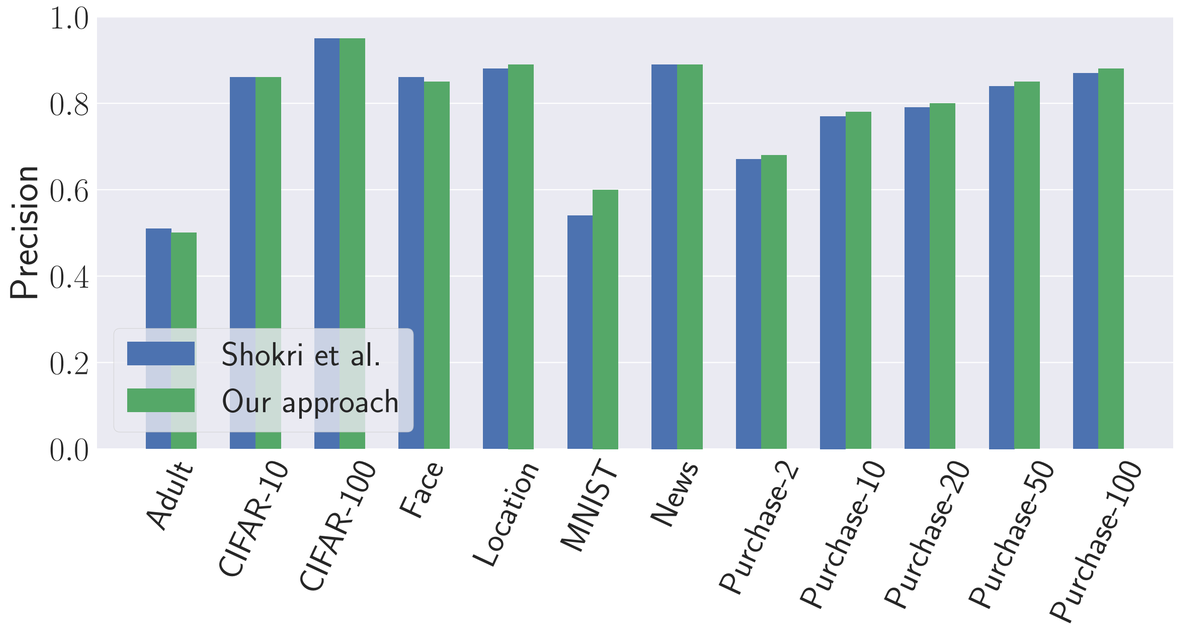
Same Distribution

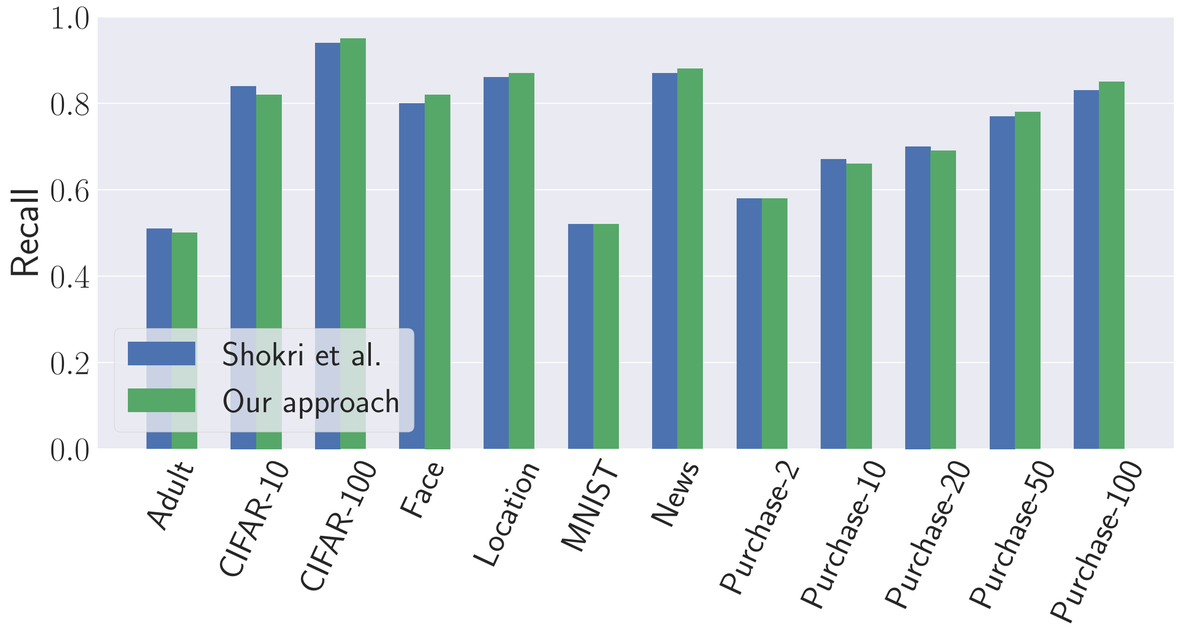
Multiple Shadow Models



Attack 1

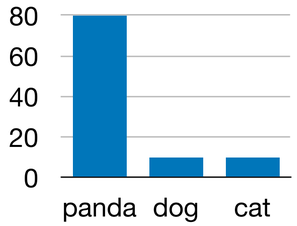
CISPA
HELMHOLTZ CENTER FOR INFORMATION SECURITY





Attack 1

CISPA HELMHOLTZ CENTER FOR INFORMATION SECURITY



One Attack Model

member or non-member?

Original Dataset

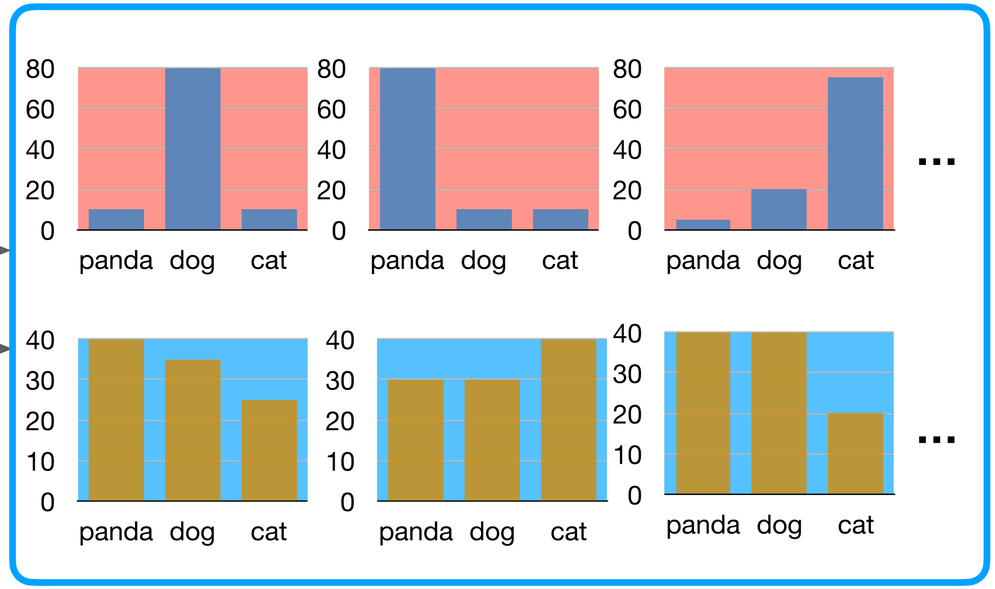
Dataset

Target Model

Train

Test

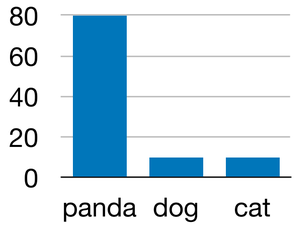
One Shadow Model



Same Distribution

Attack 1

CISPA HELMHOLTZ CENTER FOR INFORMATION SECURITY



One Attack Model

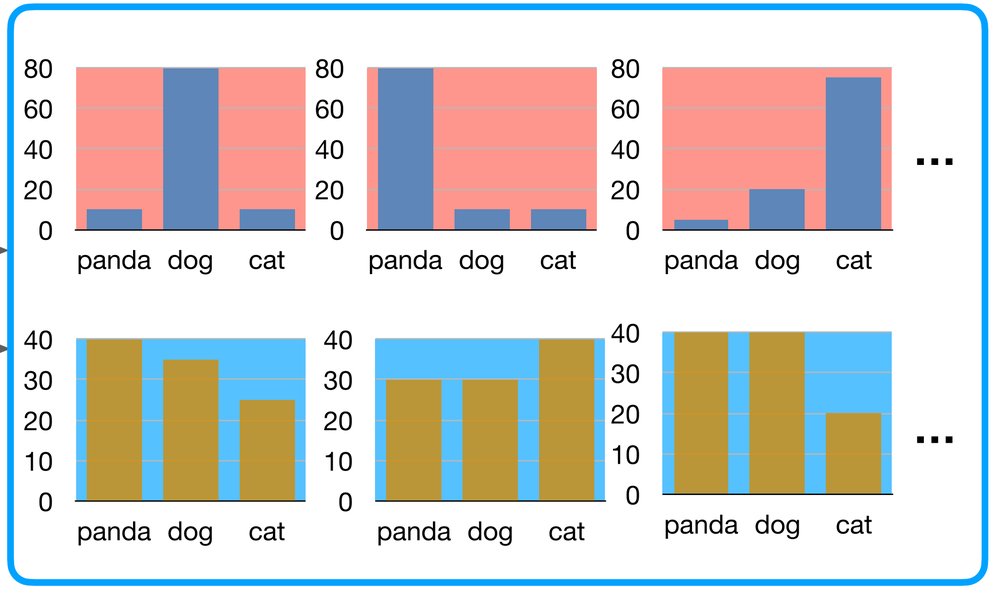
member or non-member?

Target Model

Train

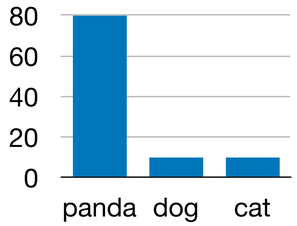
Test

One Shadow Model



Same Distribution

Attack 2



One Attack Model

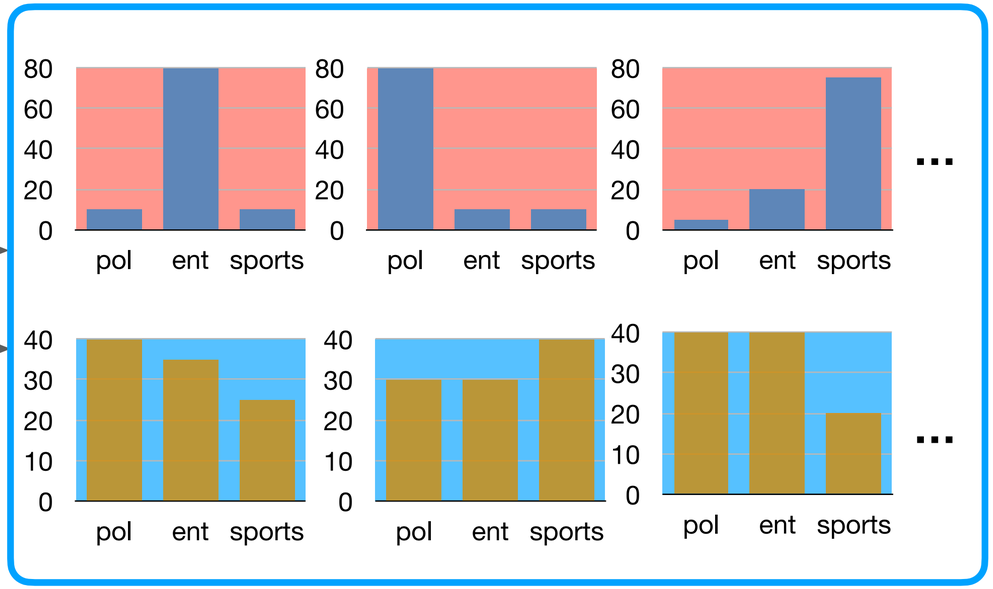
member or non-member?

Target Model

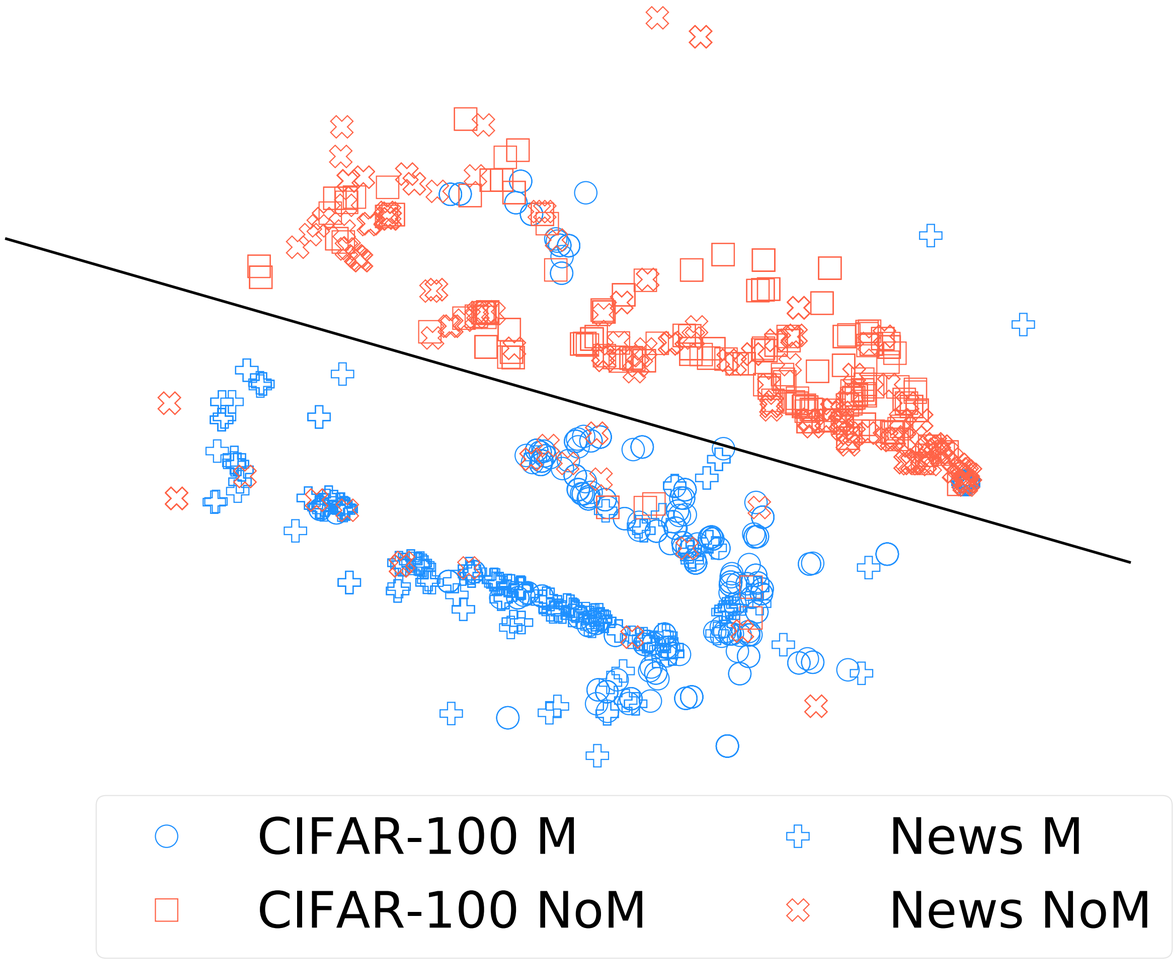
Train

Test

One Shadow Model



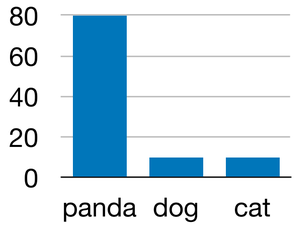
Attack 2



We didn’t transfer the dataset!

ML models got overfitted in a similar way!

Attack 2



One Attack Model

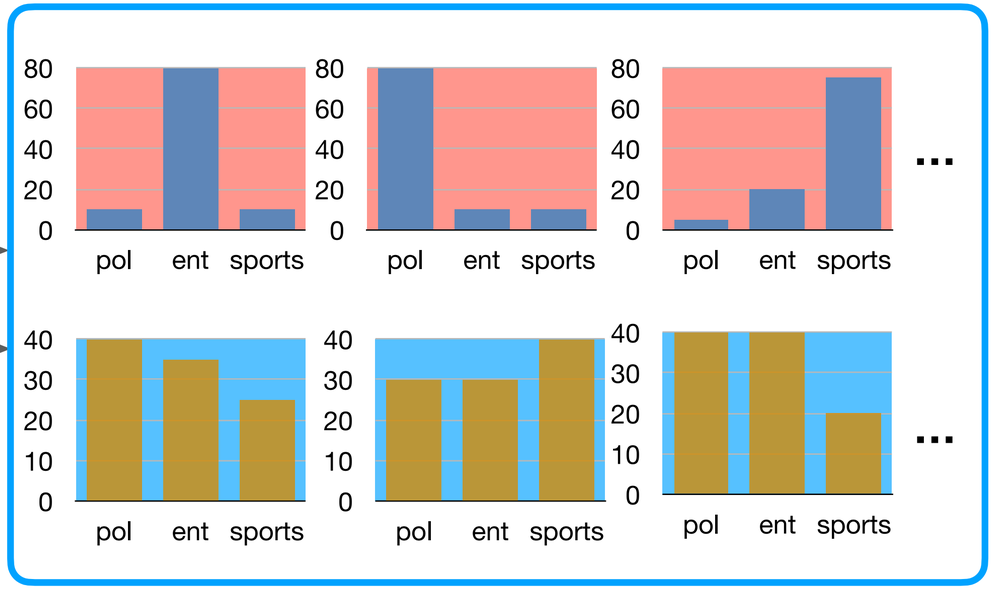
member or non-member?

Target Model

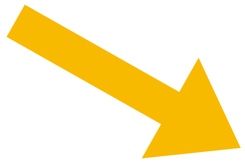
Train

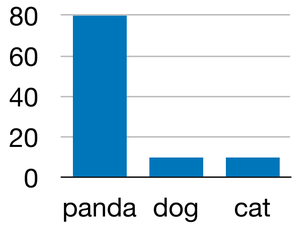
Test

One Shadow Model



Attack 3

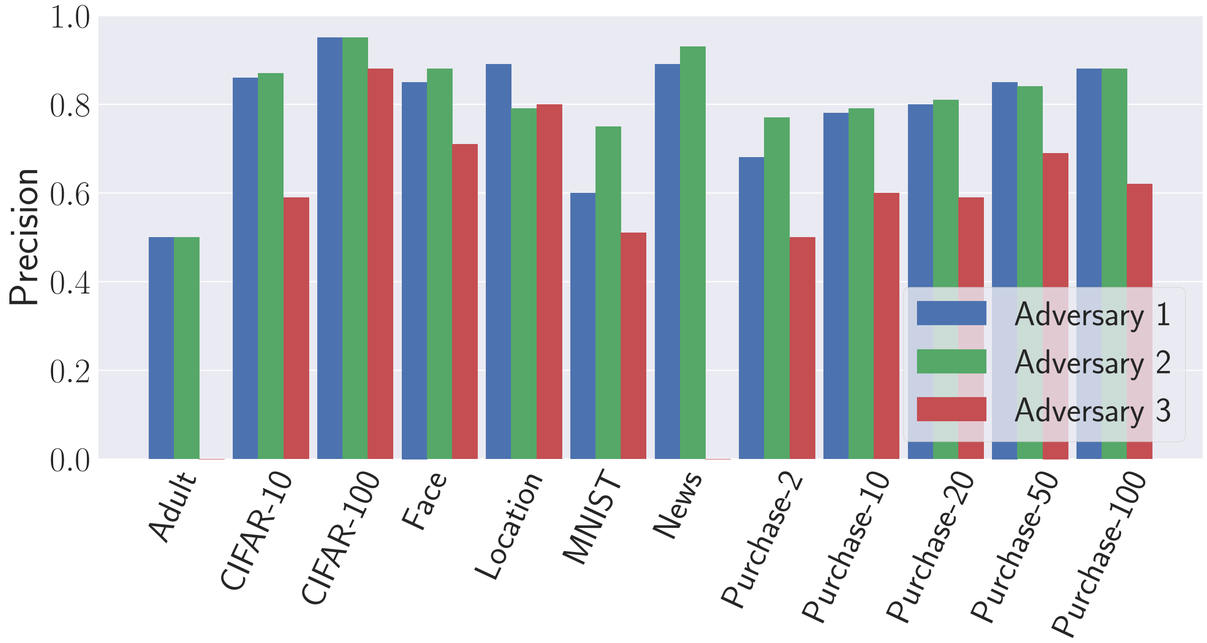


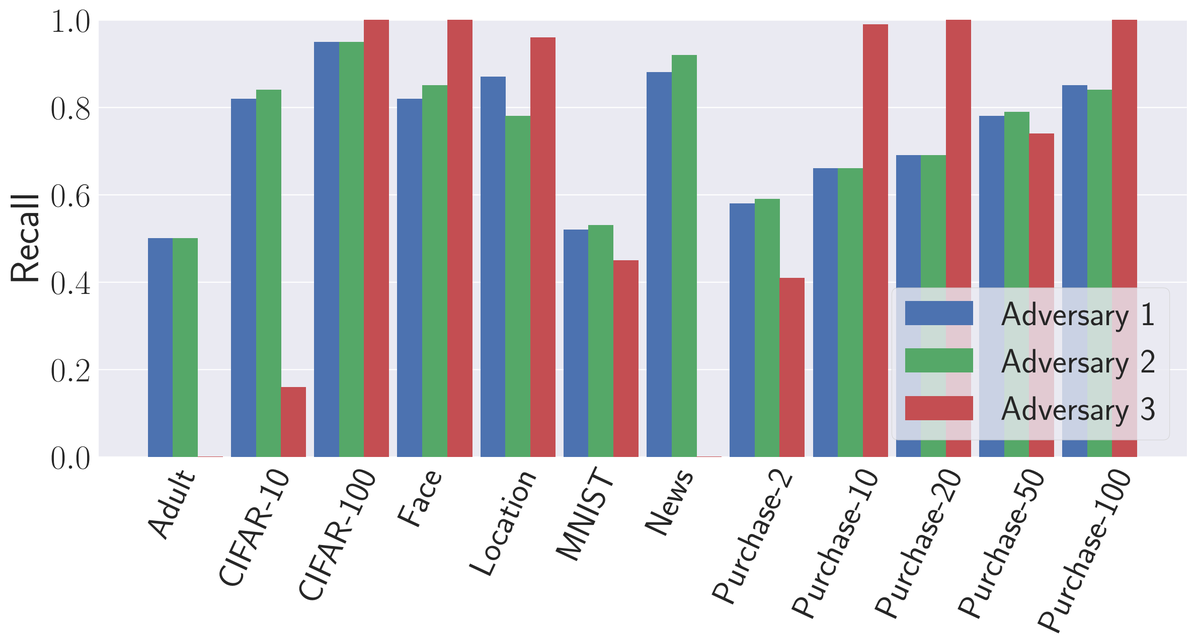


Threshold Picking

member or non-member?

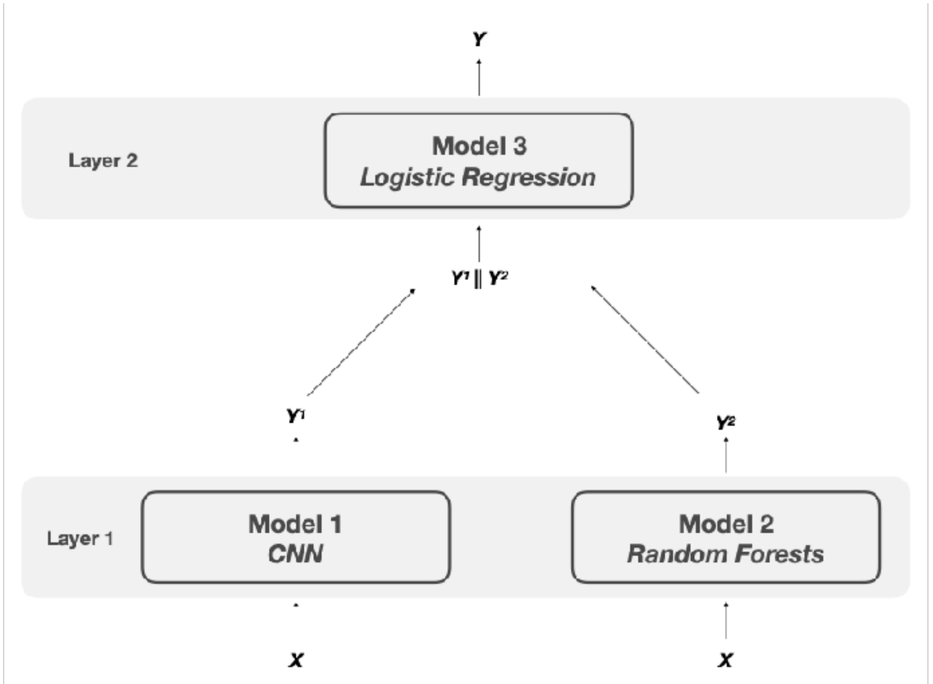
All Together

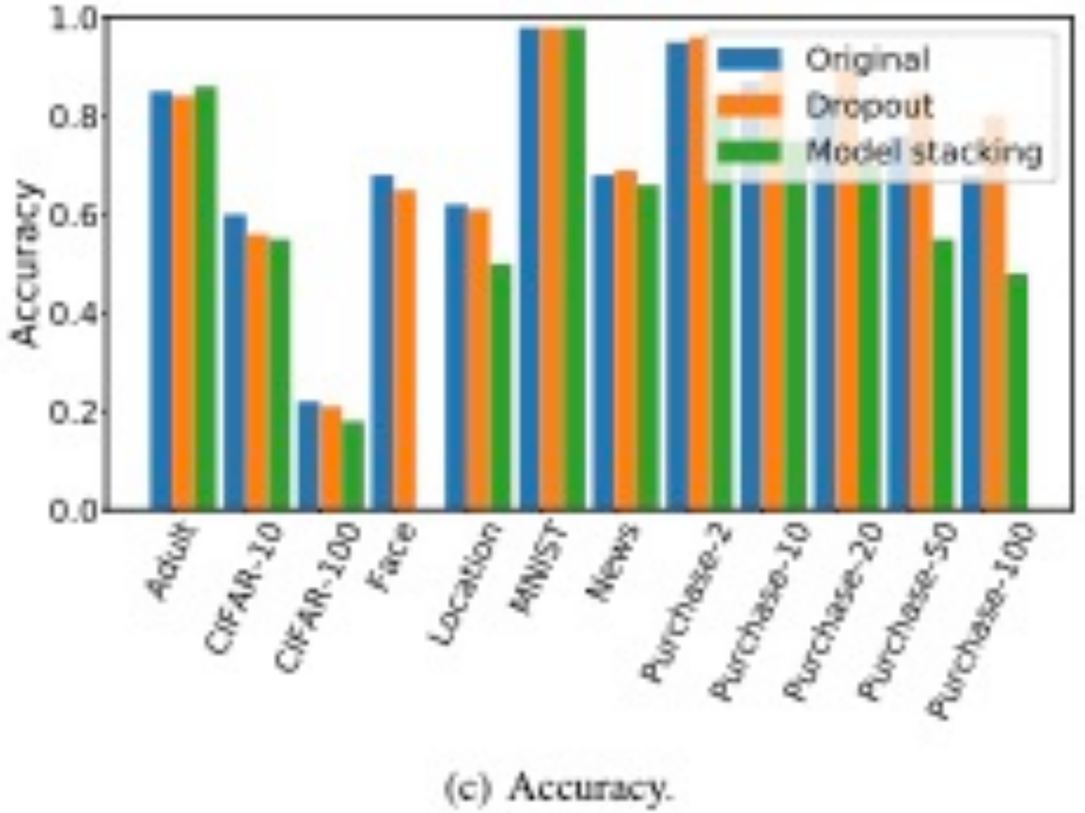




How To Defend the Attack?

- Reduce overfitting
  - E.g. regularize: Dropout
  - Combine models:Model Stacking
  - Differential Privacy (later)





(c) Accuracy.

- Membership Inference can tell if a data point was in training set
- Works for classification model

- How about other models?

Segmentations Leak: Membership Inference Attacks and Defenses in Semantic Image Segmentation

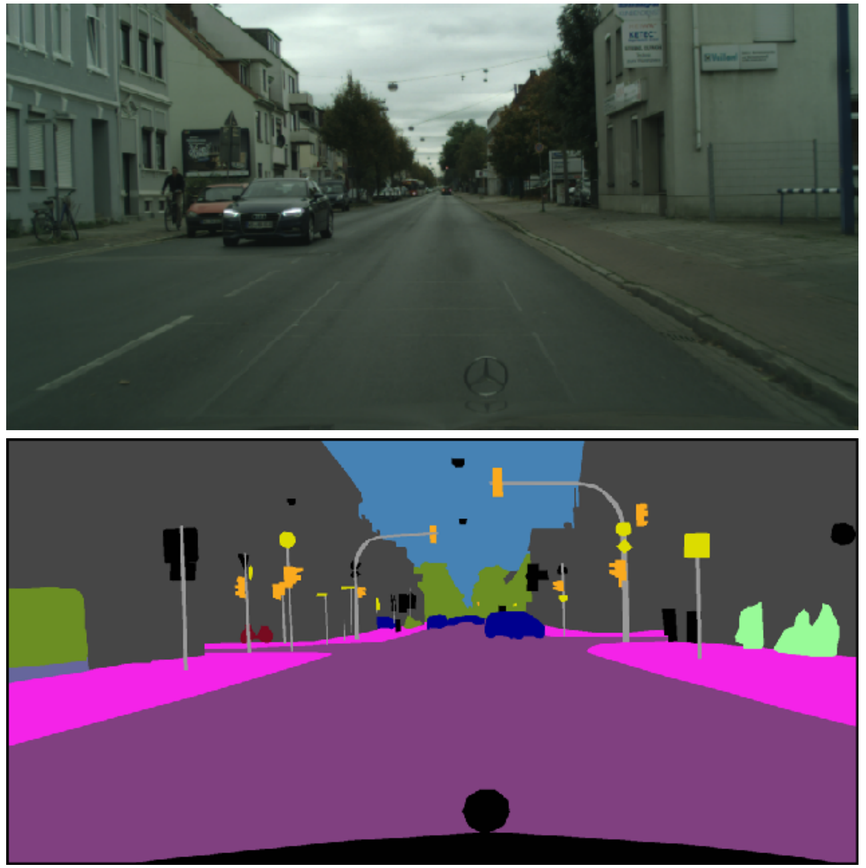
Yang He, Shadi Rahimian, Bernt Schiele and Mario Fritz

ECCV2020

Semantic Segmentation

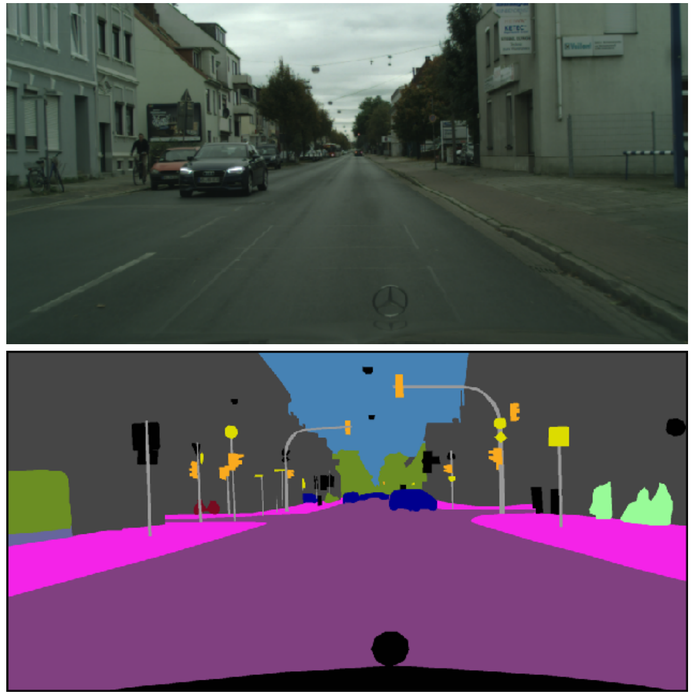
CISPA
HELMHOLTZ CENTER FOR INFORMATION SECURITY

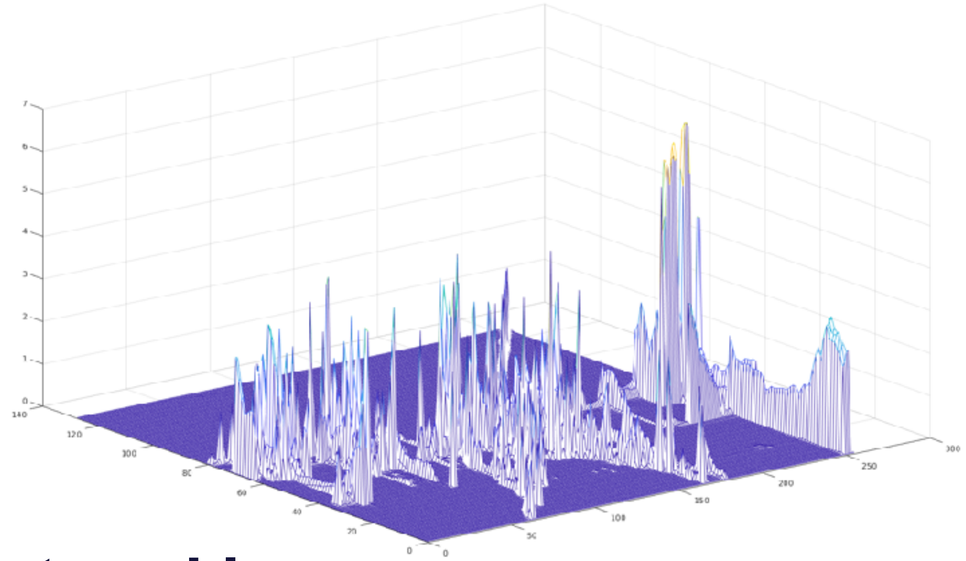
- Deep Fully Convolution Networks can predict highly accurate segmentation masks
- One ingredient for autonomous driving
- Large datasets play key role
- Membership inference:
  - Privacy violation
  - Model forensics



Test Loss Map as Membership “Signal”

CISPA
HELMHOLTZ CENTER FOR INFORMATION SECURITY





Structured loss map

Results

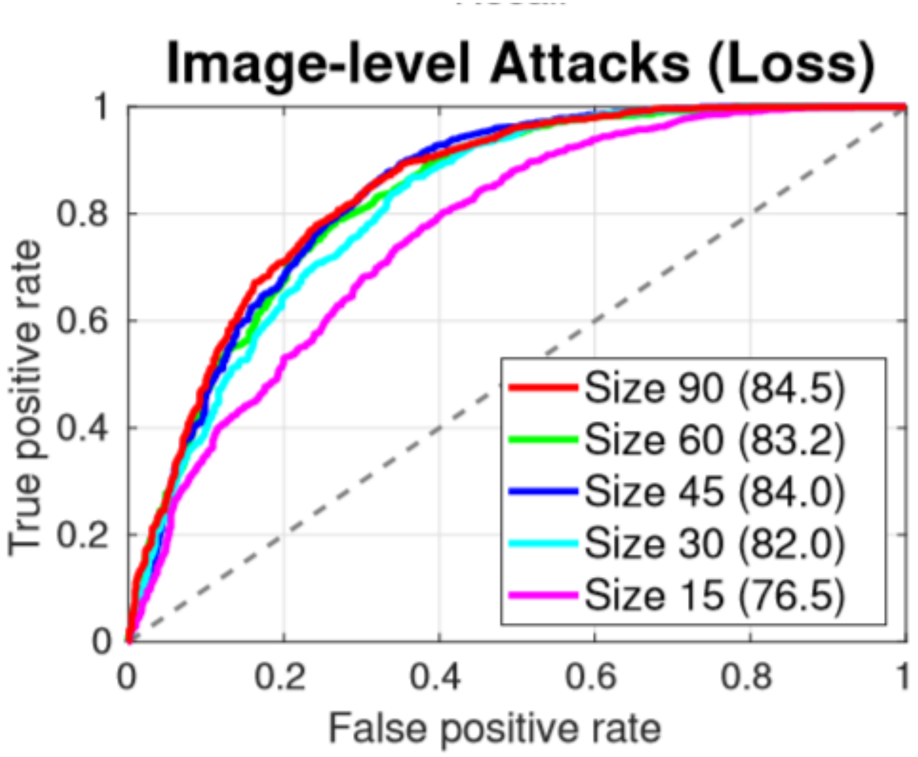
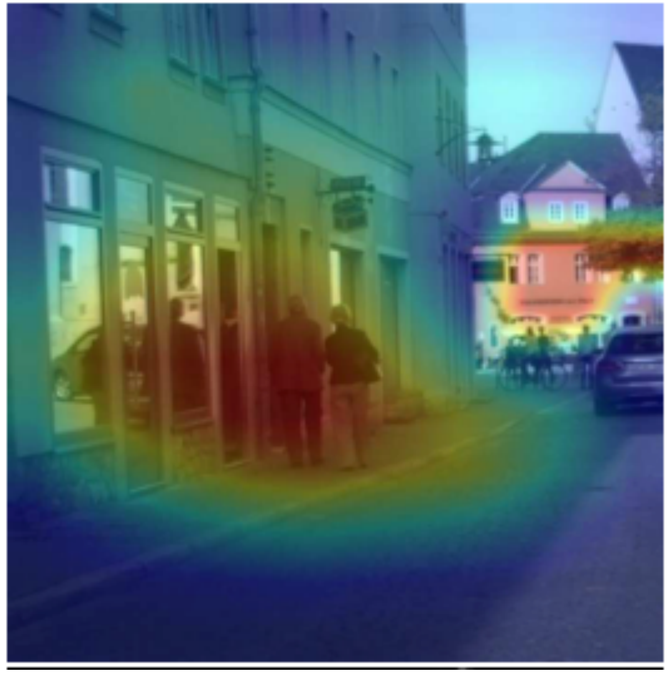


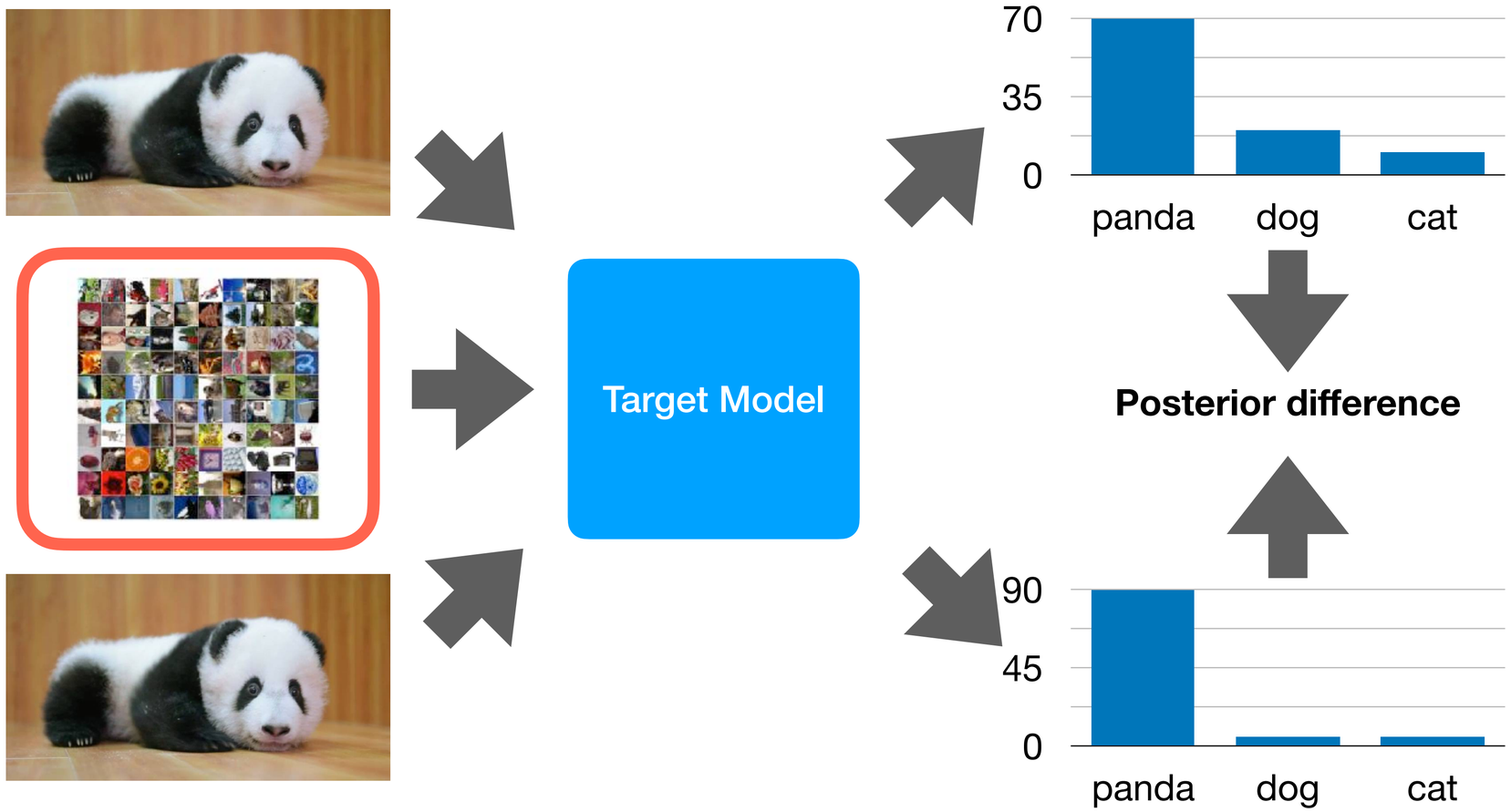
Image-level Attacks (Loss)
True positive rate
False positive rate
0
0.2
0.4
0.6
0.8
1
Size 90 (84.5)
Size 60 (83.2)
Size 45 (84.0)
Size 30 (82.0)
Size 15 (76.5)



- Membership Inference can tell if a data point was in training set
  - Classification, Segmentation, Generators
- … but maybe it’s not so bad. Or is it?

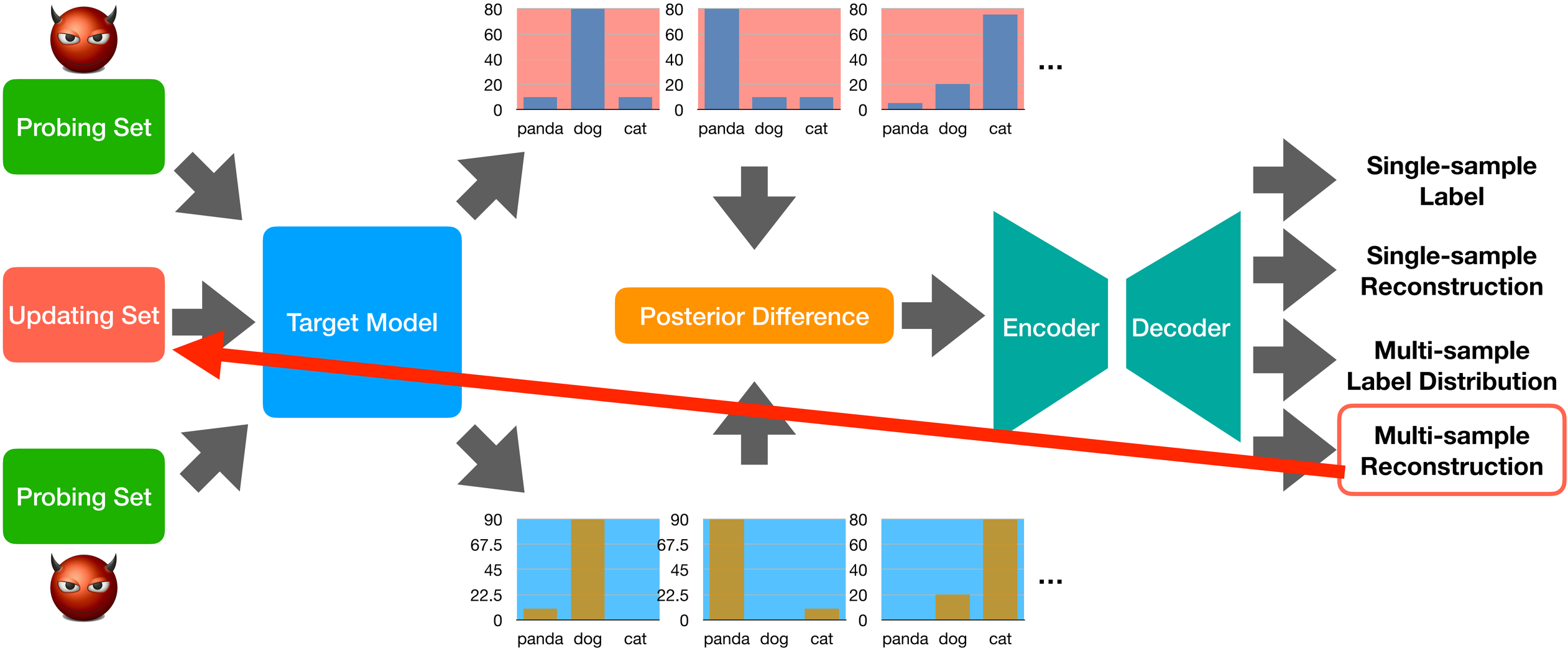
A New Attack Surface for Online Learning

CISPA
HELMHOLTZ CENTER FOR INFORMATION SECURITY

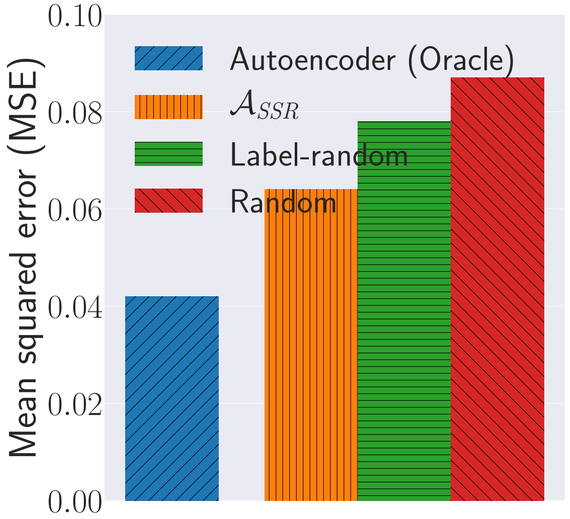


Research Question: Can this posterior difference be used to infer information of the updating set?

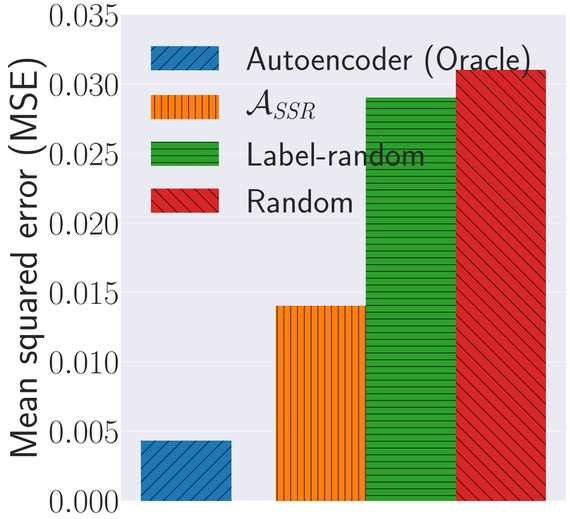
General Attack Pipeline



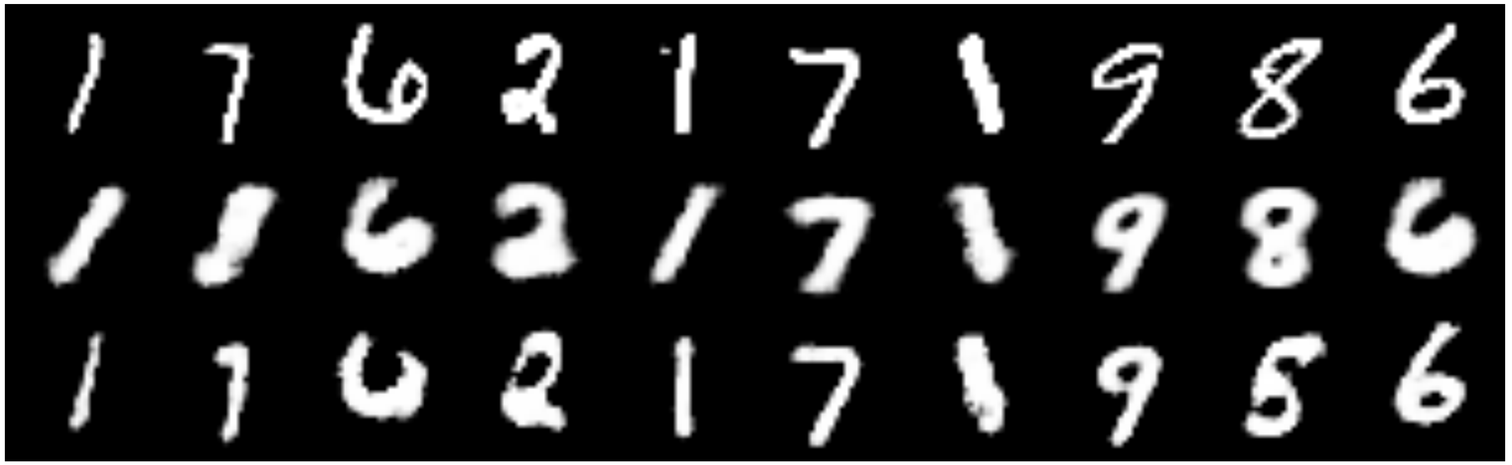
Single-Sample Reconstruction



CIFAR-10



MNIST



True update

Inferred update

Oracle reconstruction

Reconstruction of MNIST

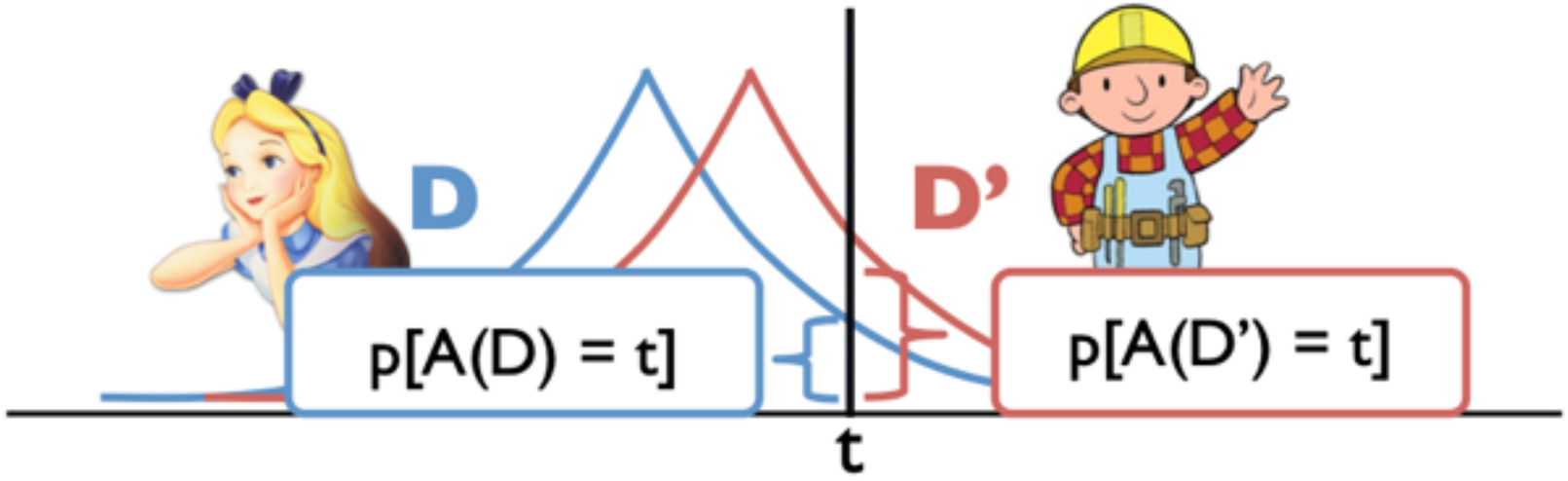
CISPA
HELMHOLTZ CENTER FOR INFORMATION SECURITY

- Under some assumptions - reconstructing the training is possible.
- How can we prevent privacy leakage?

Defenses - Privacy Preserving Machine Learning

CISPA HELMHOLTZ CENTER FOR INFORMATION SECURITY

- Membership Inference relies on overfitting
- Defenses by drop-out, ensembles, proper model training, return only label
  - No guarantees
- Principled defenses by differential privacy (DP)
- Algorithm can be transformed to satisfy DP, e.g. DP-SGD



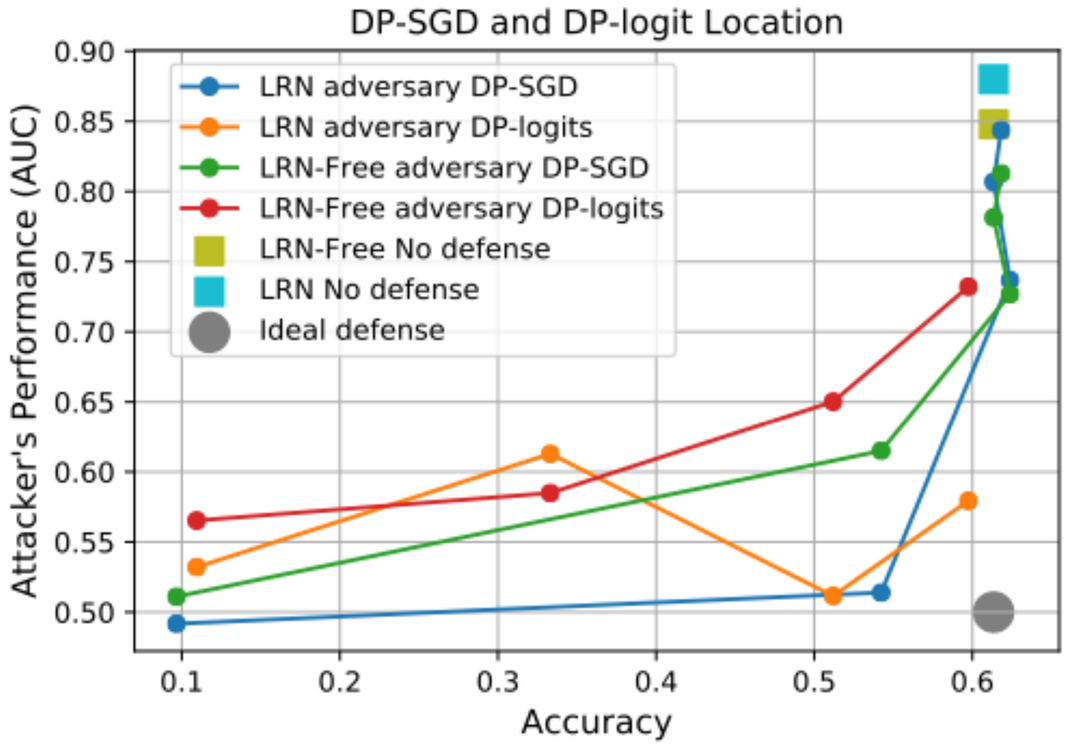
For all D, D' that differ in one person's value,

If A = ϵ-differentially private randomized algorithm, then:

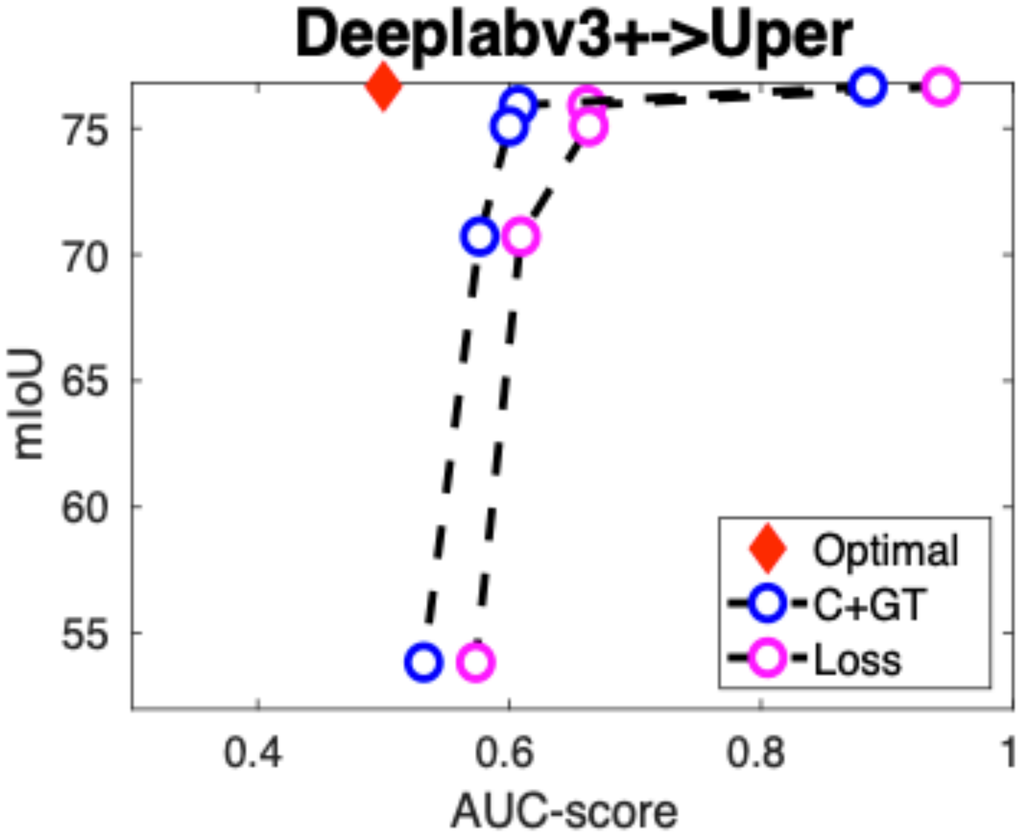
$$\sup_t \left| \log \frac{p(A(D) = t)}{p(A(D') = t)} \right| \leq \epsilon$$

Differential Privacy defends against Membership Inference

CISPA HELMHOLTZ CENTER FOR INFORMATION SECURITY



Classification



Segmentation

Machine Learning in Cybersecurity

- Adversarial Perturbations & Game Theory
- Membership Inference Attacks

Prof. Dr. Mario Fritz | 28.11.2019