CISPA
HELMHOLTZ CENTER FOR INFORMATION SECURITY

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

Ahmed Salem, Yang Zhang, Mathias Humbert, Pascal Berrang,
Mario Fritz, Michael Backes

CISPA Helmholtz Center for Information Security, Swiss Data Science Center

The more a model learns about you, the better it gets

Cars

The more a model learns about you, the better it gets

Cars

The more a model learns about you, the better it gets

Activity Tracker

Cars

Social Media

The more a model learns about you, the better it gets

Activity Tracker

Cars

Social Media

The more a model learns about you, the better it gets

Activity Tracker

Personal Assistant

amazon alexa

How Sensitive The Data Can Be?

CISPA
HELMHOLTZ CENTER FOR INFORMATION SECURITY

How Sensitive The Data Can Be?

CISPA
HELMHOLTZ CENTER FOR INFORMATION SECURITY

- Financial data

How Sensitive The Data Can Be?

CISPA
HELMHOLTZ CENTER FOR INFORMATION SECURITY

- Financial data
- Location and activity data

How Sensitive The Data Can Be?

CISPA
HELMHOLTZ CENTER FOR INFORMATION SECURITY

- Financial data
- Location and activity data
- Biomedical data

How Sensitive The Data Can Be?

CISPA
HELMHOLTZ CENTER FOR INFORMATION SECURITY

- Financial data
- Location and activity data
- Biomedical data
- etc.

Privacy in Machine Learning

CISPA
HELMHOLTZ CENTER FOR INFORMATION SECURITY

Privacy in Machine Learning

CISPA
HELMHOLTZ CENTER FOR INFORMATION SECURITY

- ML models are trained on sensitive data

Privacy in Machine Learning

CISPA
HELMHOLTZ CENTER FOR INFORMATION SECURITY

- ML models are trained on sensitive data
- Main focus is performance

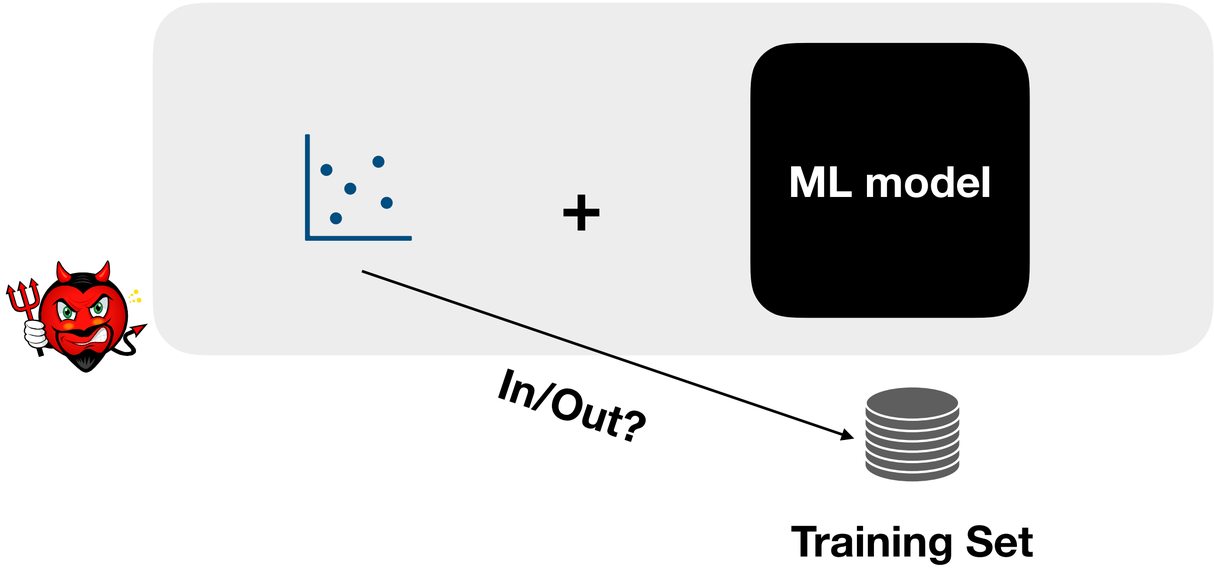
Privacy in Machine Learning

CISPA
HELMHOLTZ CENTER FOR INFORMATION SECURITY

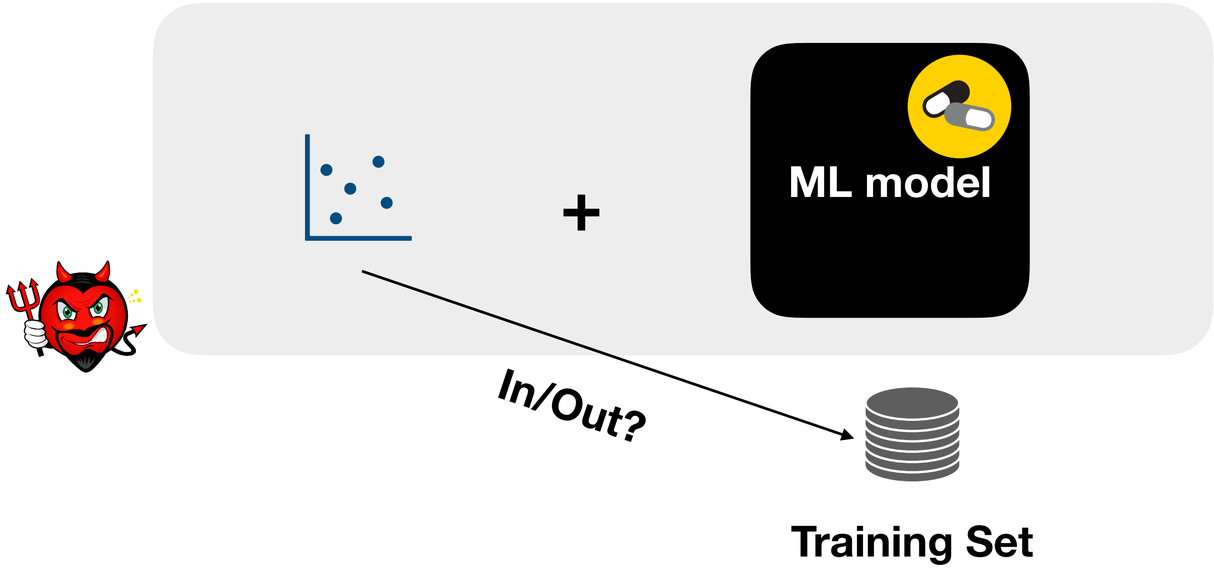
- ML models are trained on sensitive data
- Main focus is performance
  - Largely overlooked

Membership Inference

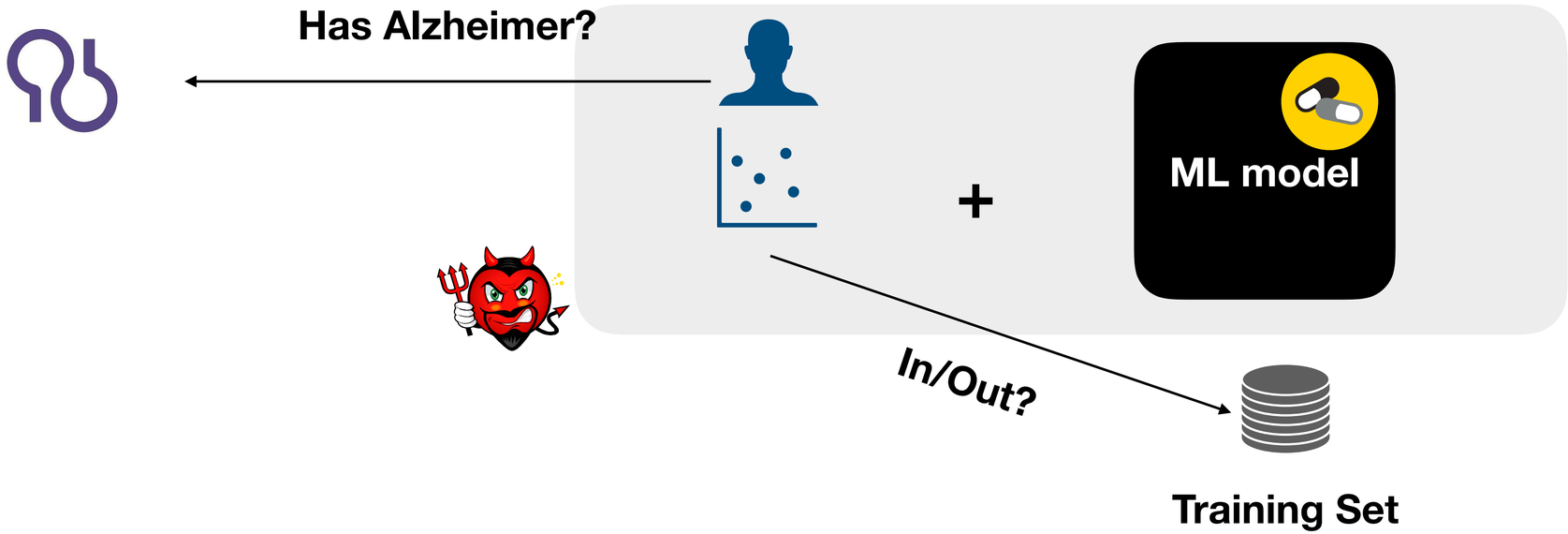
Membership Inference



Membership Inference



Membership Inference



Machine Learning Pipeline

Dataset

Model

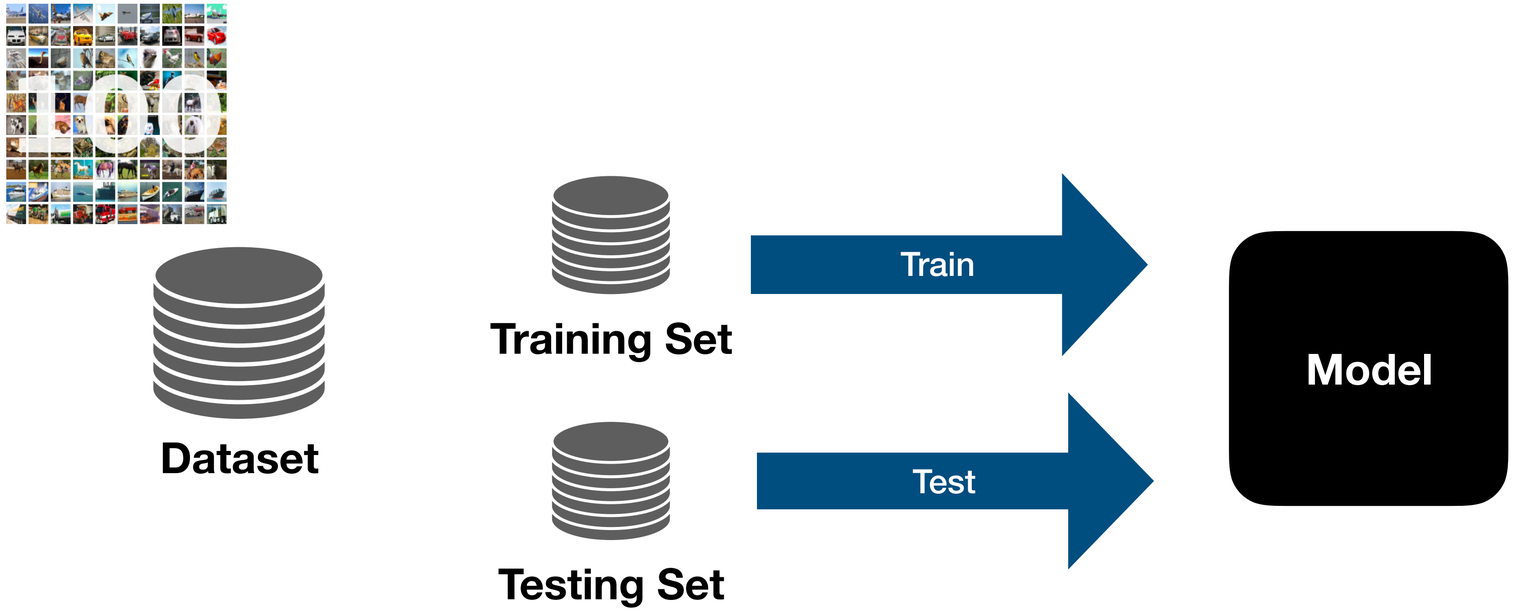
Machine Learning Pipeline

CISPA
HELMHOLTZ CENTER FOR INFORMATION SECURITY

Dataset

Model

Machine Learning Pipeline



Machine Learning Pipeline

Dataset

Training Set

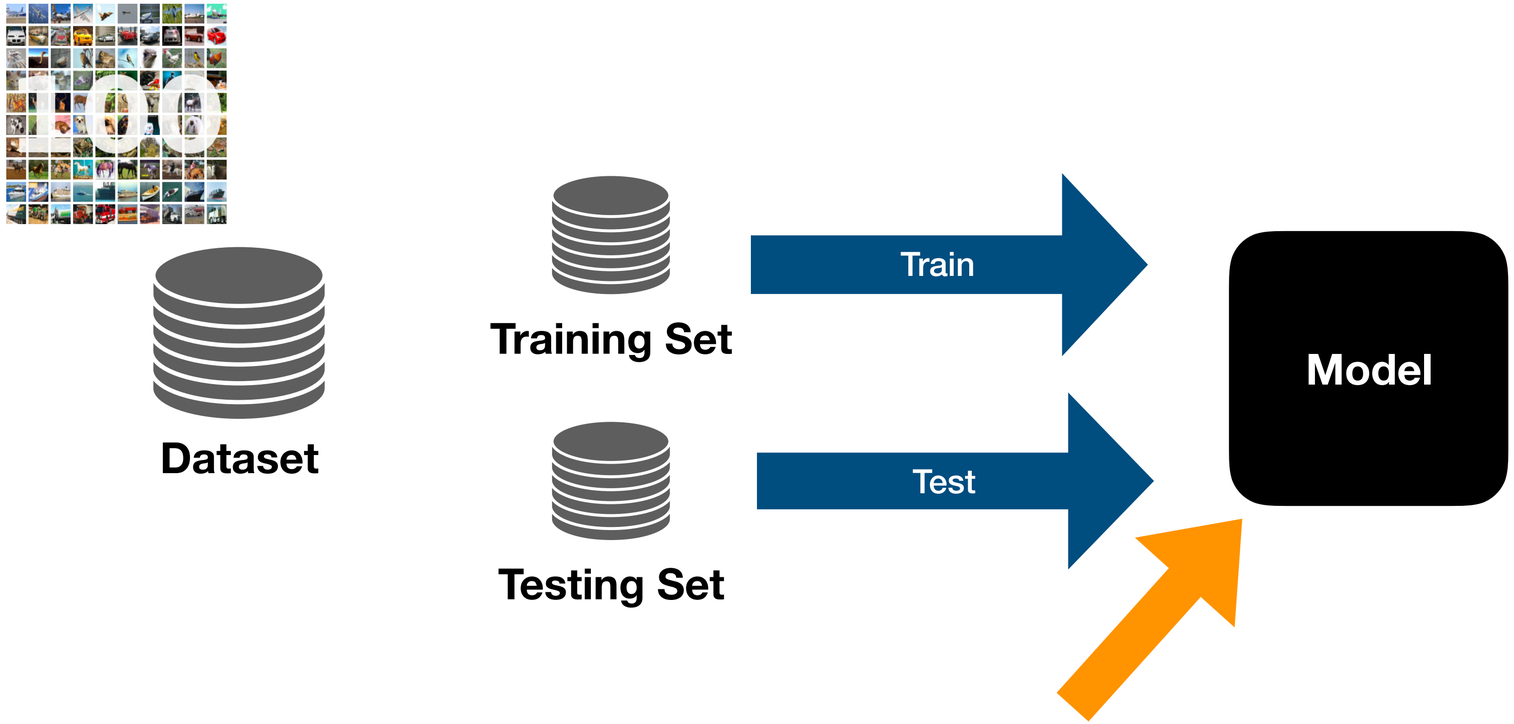
Testing Set

Train

Test

Model

Machine Learning Pipeline



Machine Learning Pipeline

Dataset

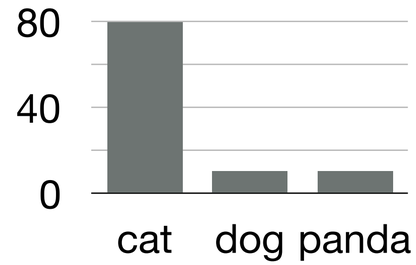
Training Set

Train

Testing Set

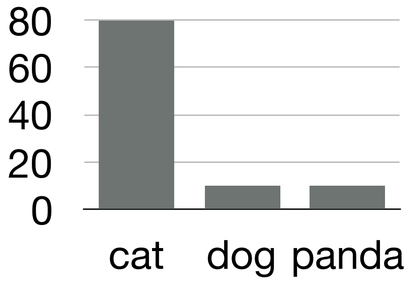
Test

Model

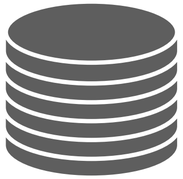


Threat Model

Target
ML Model



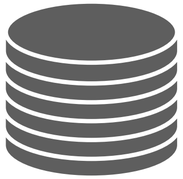
State Of The Art (Shokri et al.)



Target Dataset

Target Model

State Of The Art (Shokri et al.)



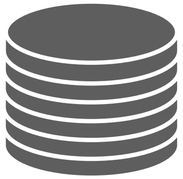
Target Dataset

Target Model

Attack Models

Attack Models

State Of The Art (Shokri et al.)



Target Dataset

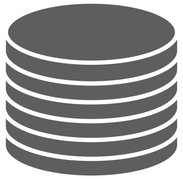
Target Model

Ground Truth?

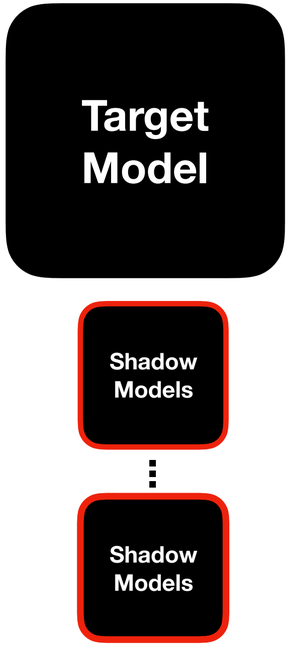
Attack Models

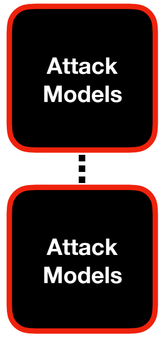
Attack Models

State Of The Art (Shokri et al.)

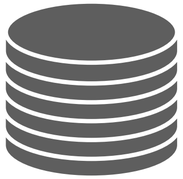


Target Dataset



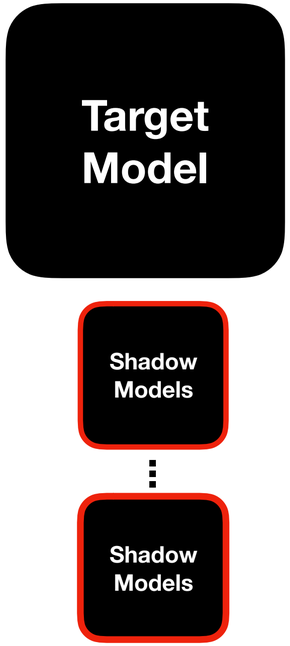


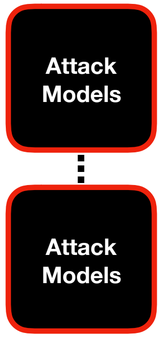
State Of The Art (Shokri et al.)



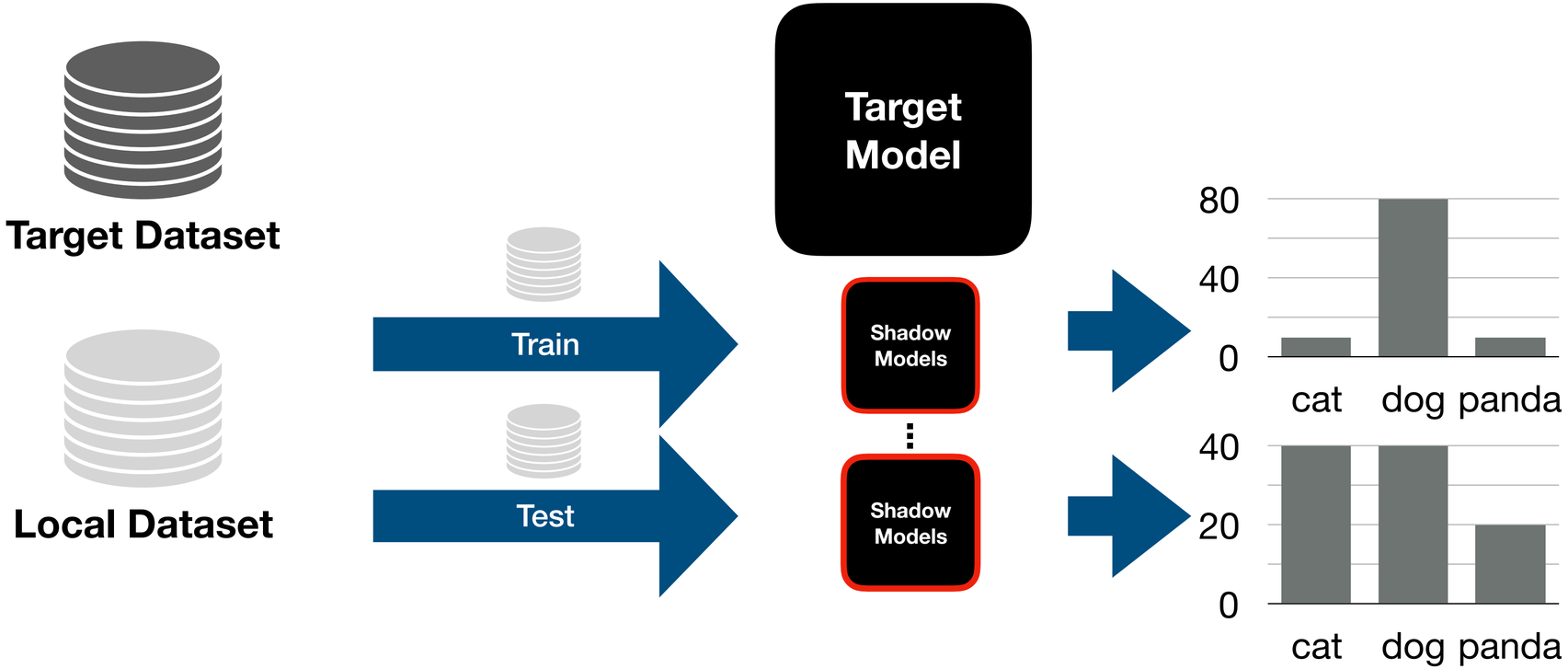
Target Dataset

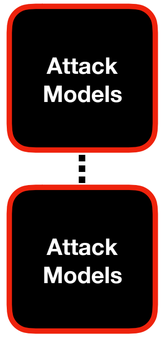
Local Dataset



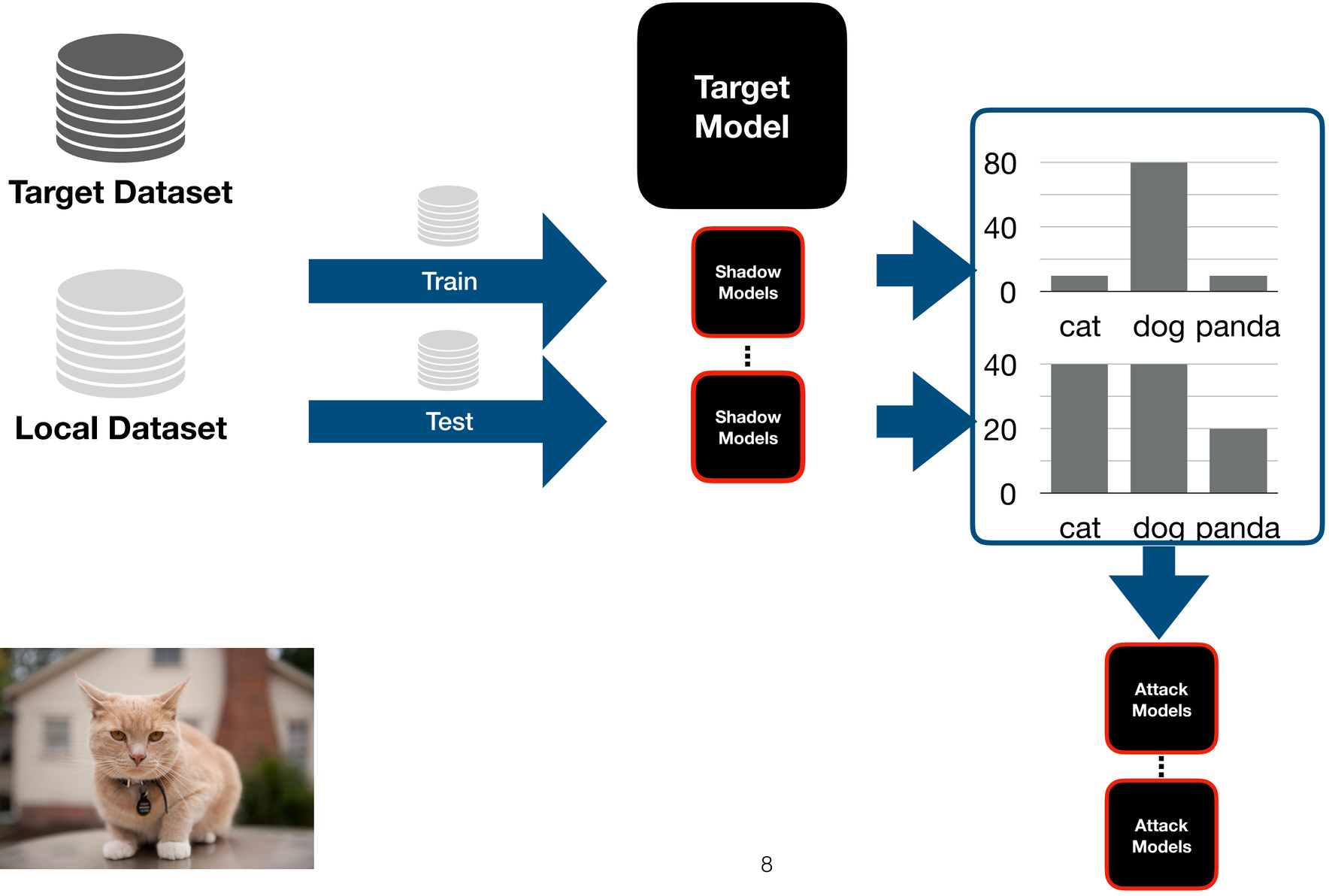


State Of The Art (Shokri et al.)

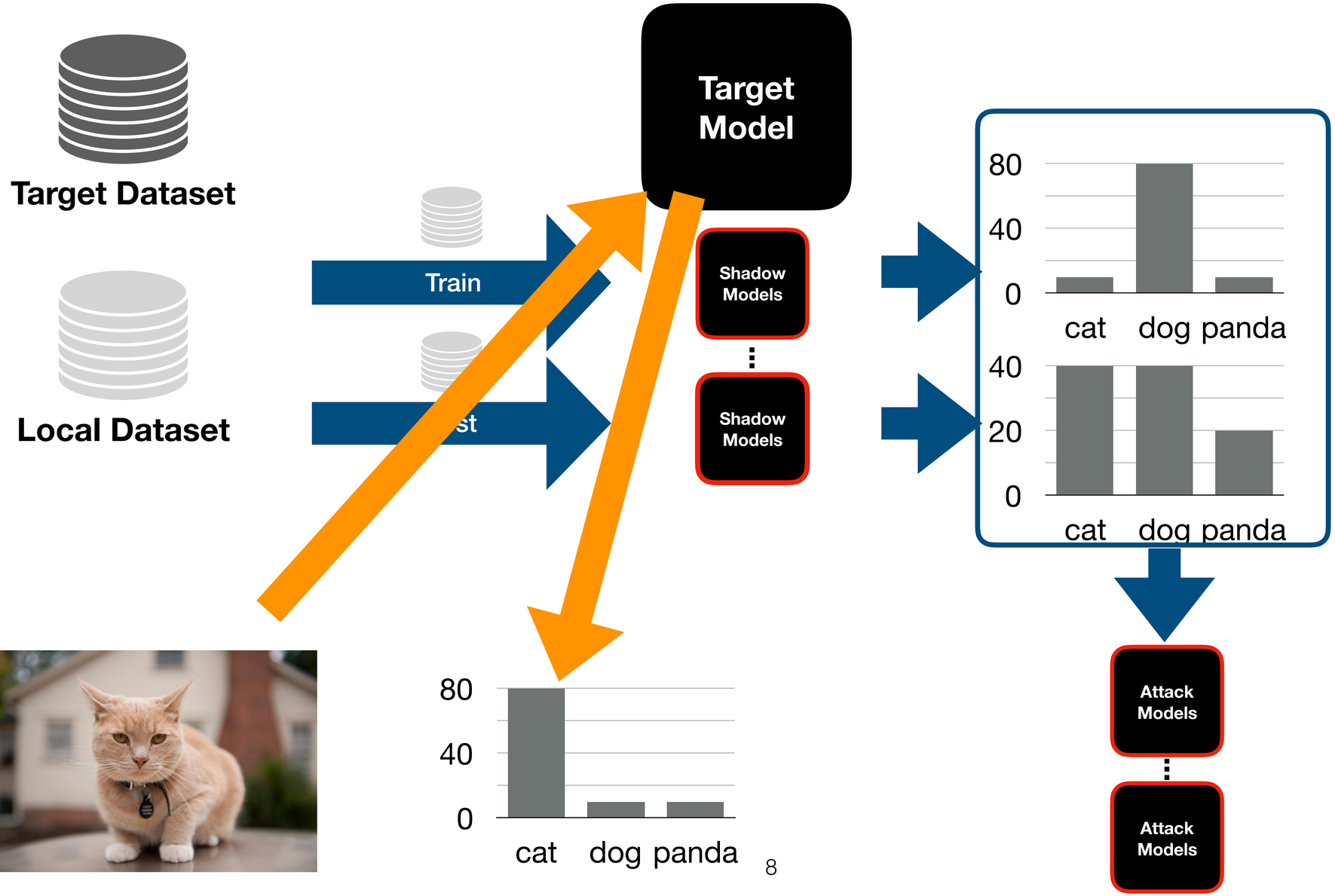




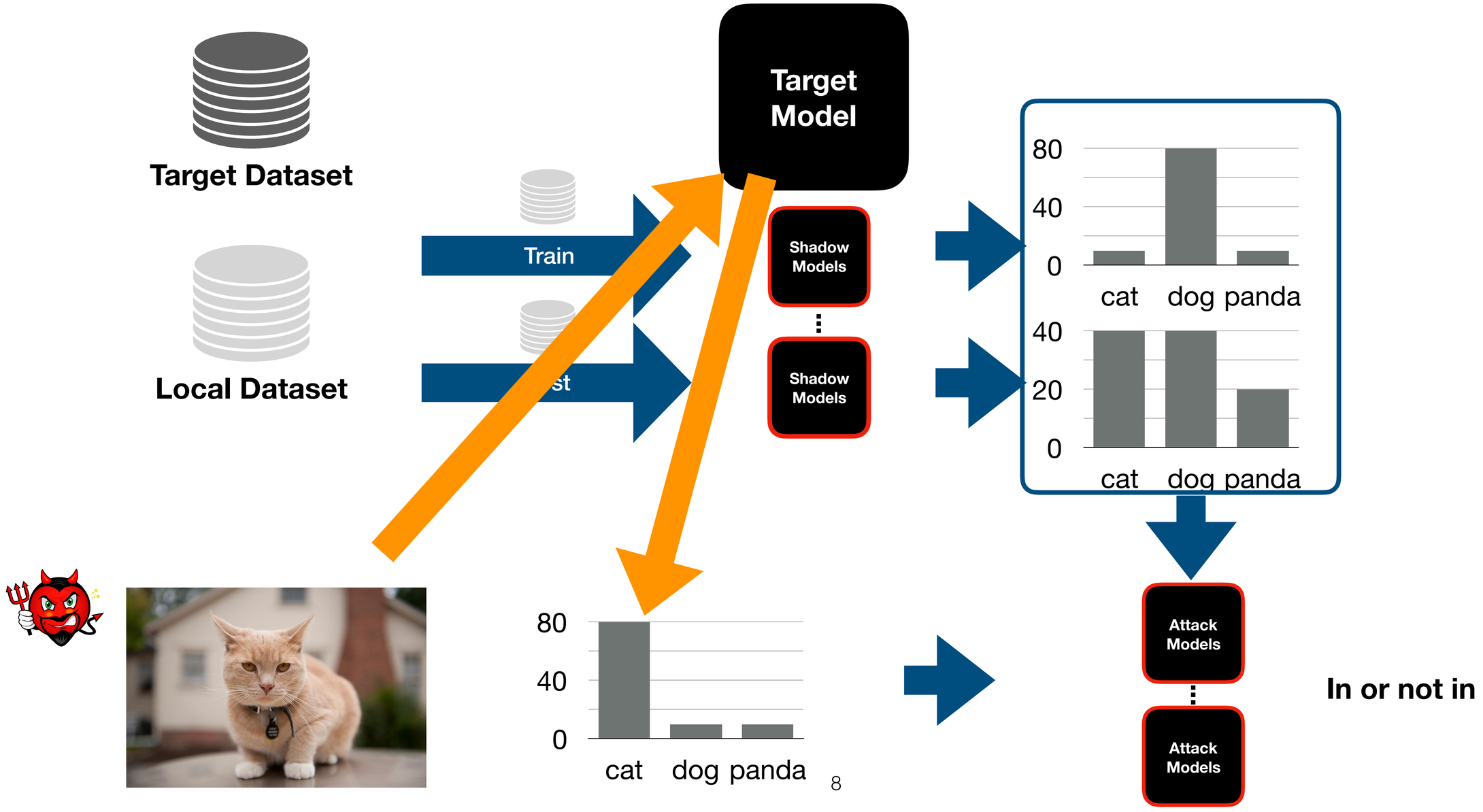
State Of The Art (Shokri et al.)



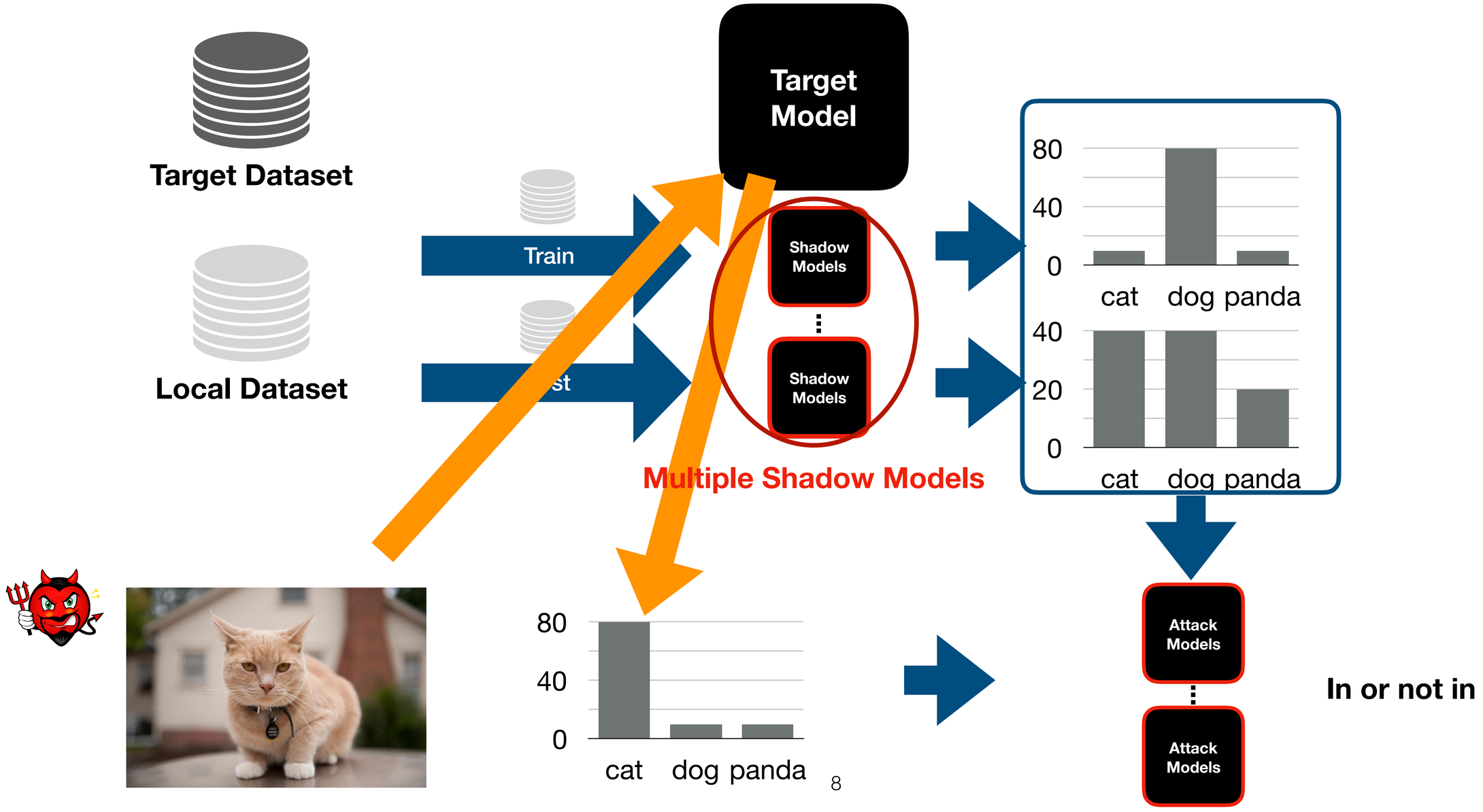
State Of The Art (Shokri et al.)



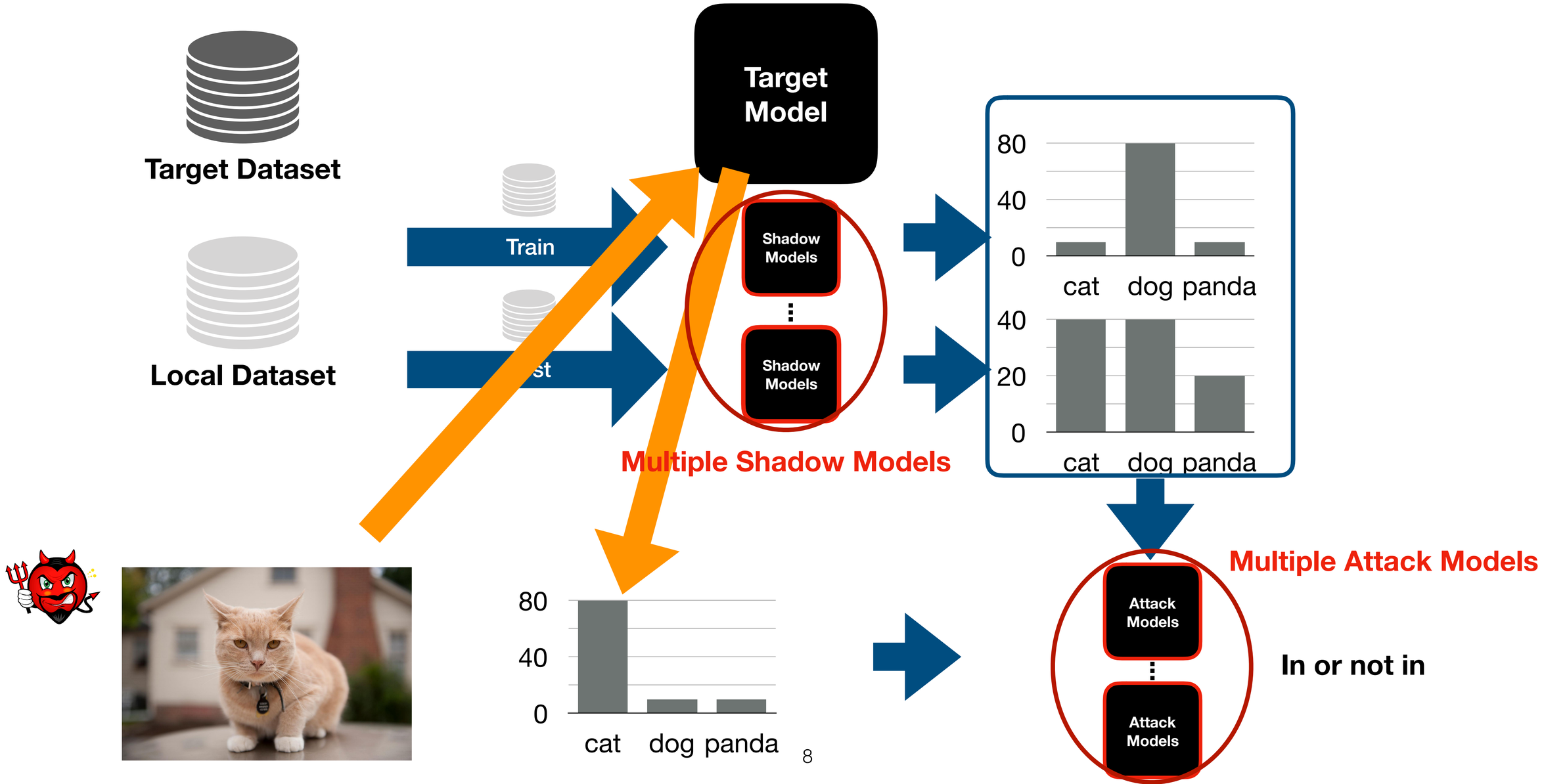
State Of The Art (Shokri et al.)



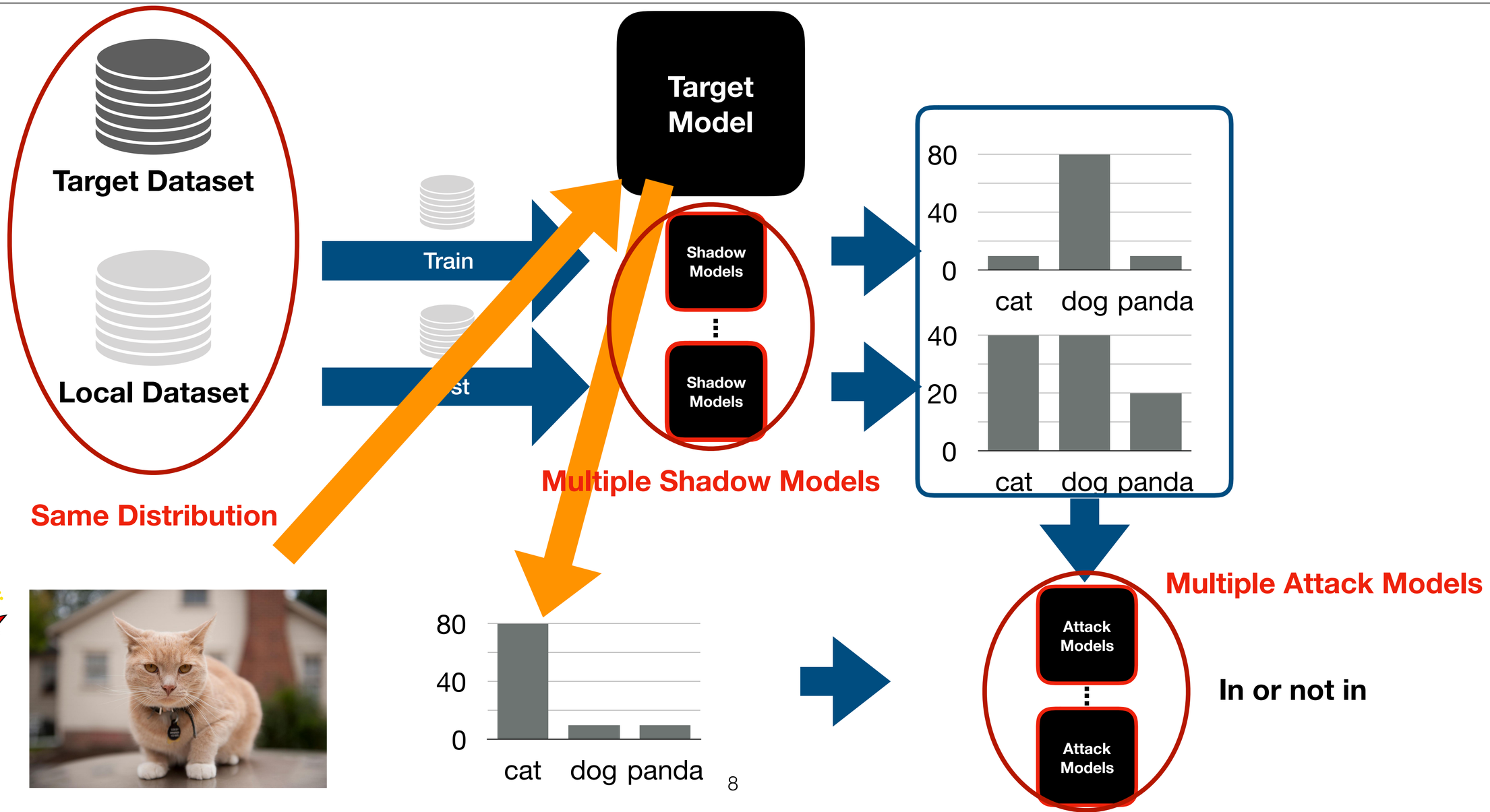
State Of The Art (Shokri et al.)



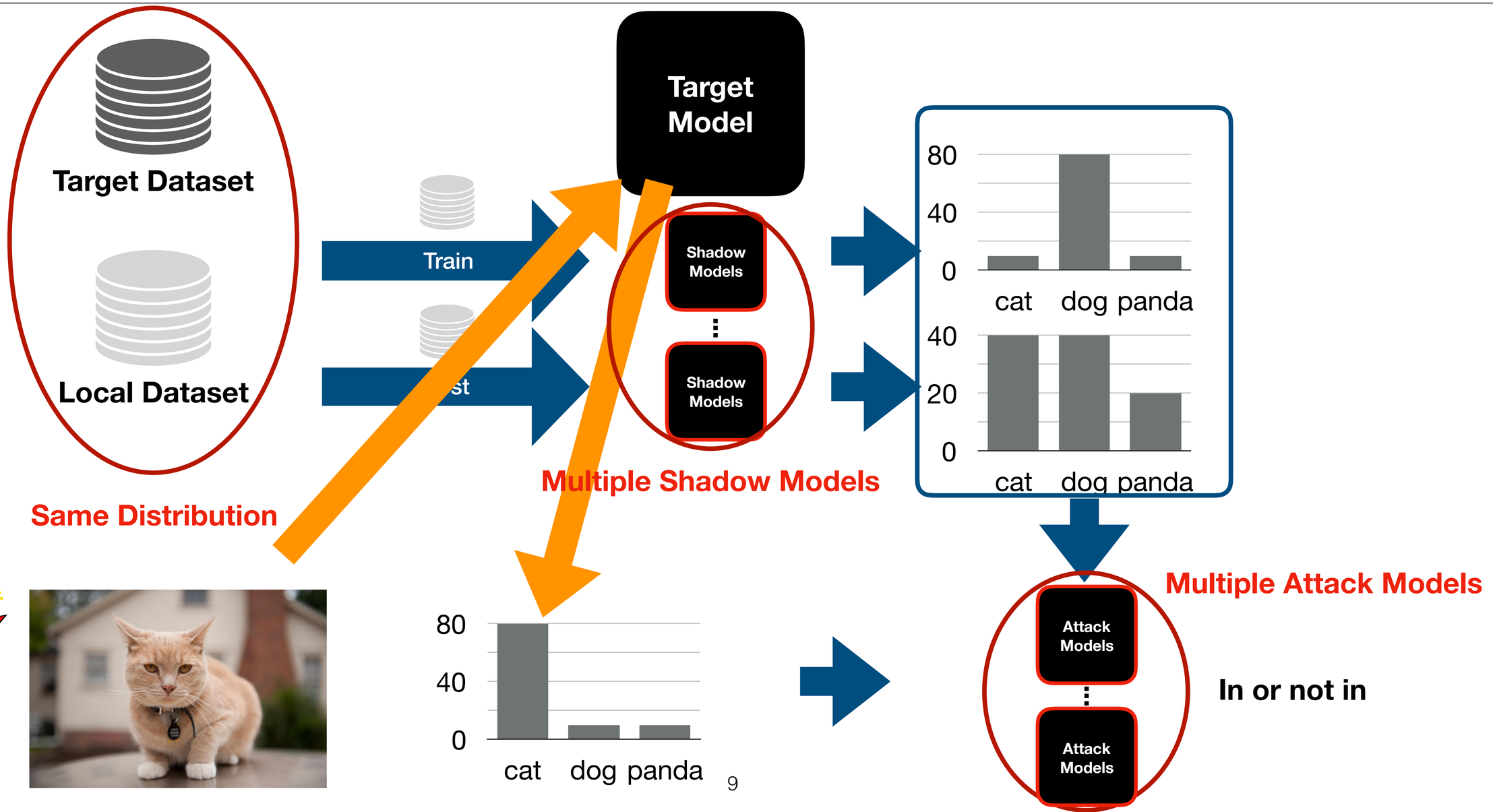
State Of The Art (Shokri et al.)



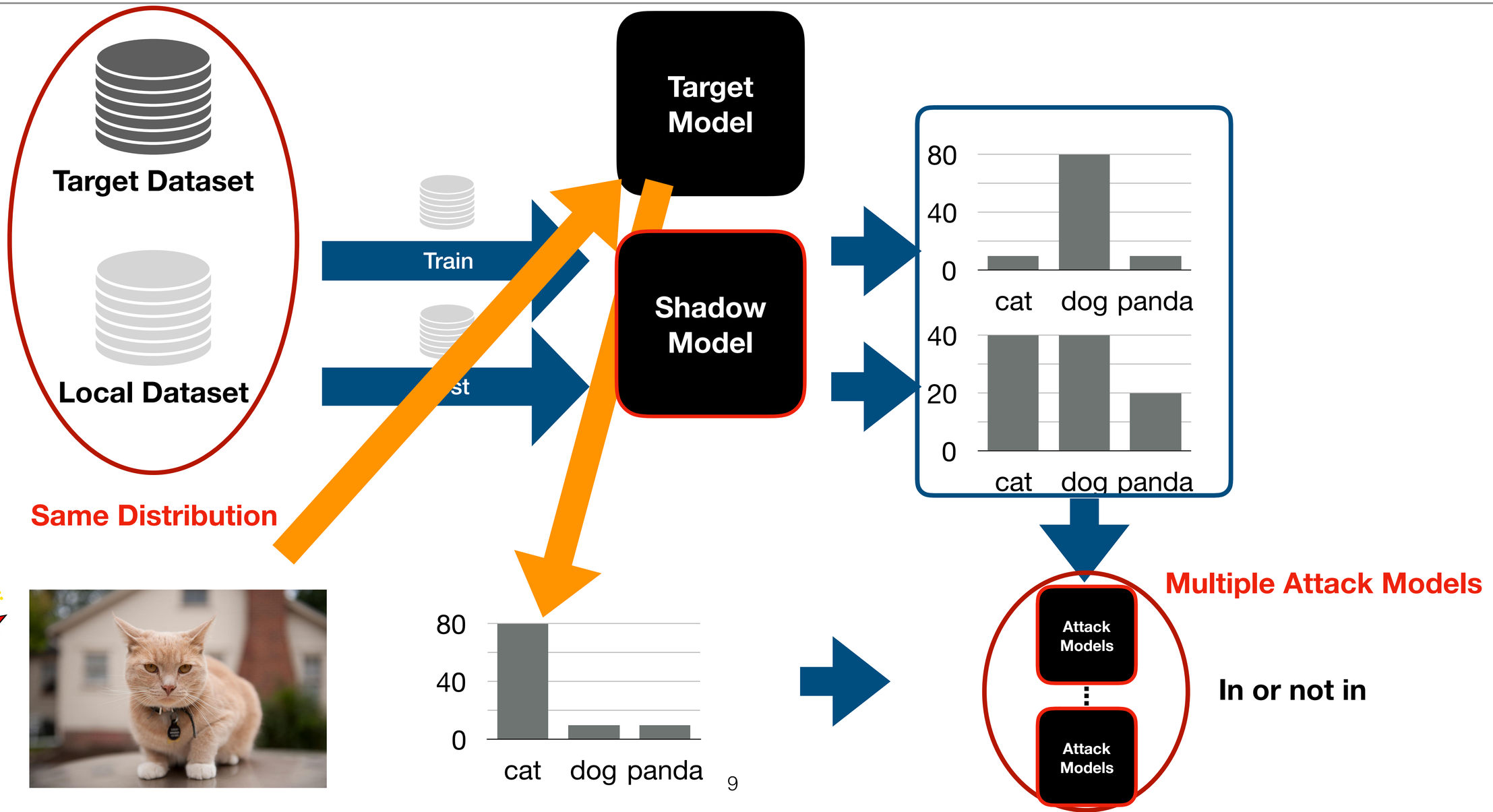
State Of The Art (Shokri et al.)



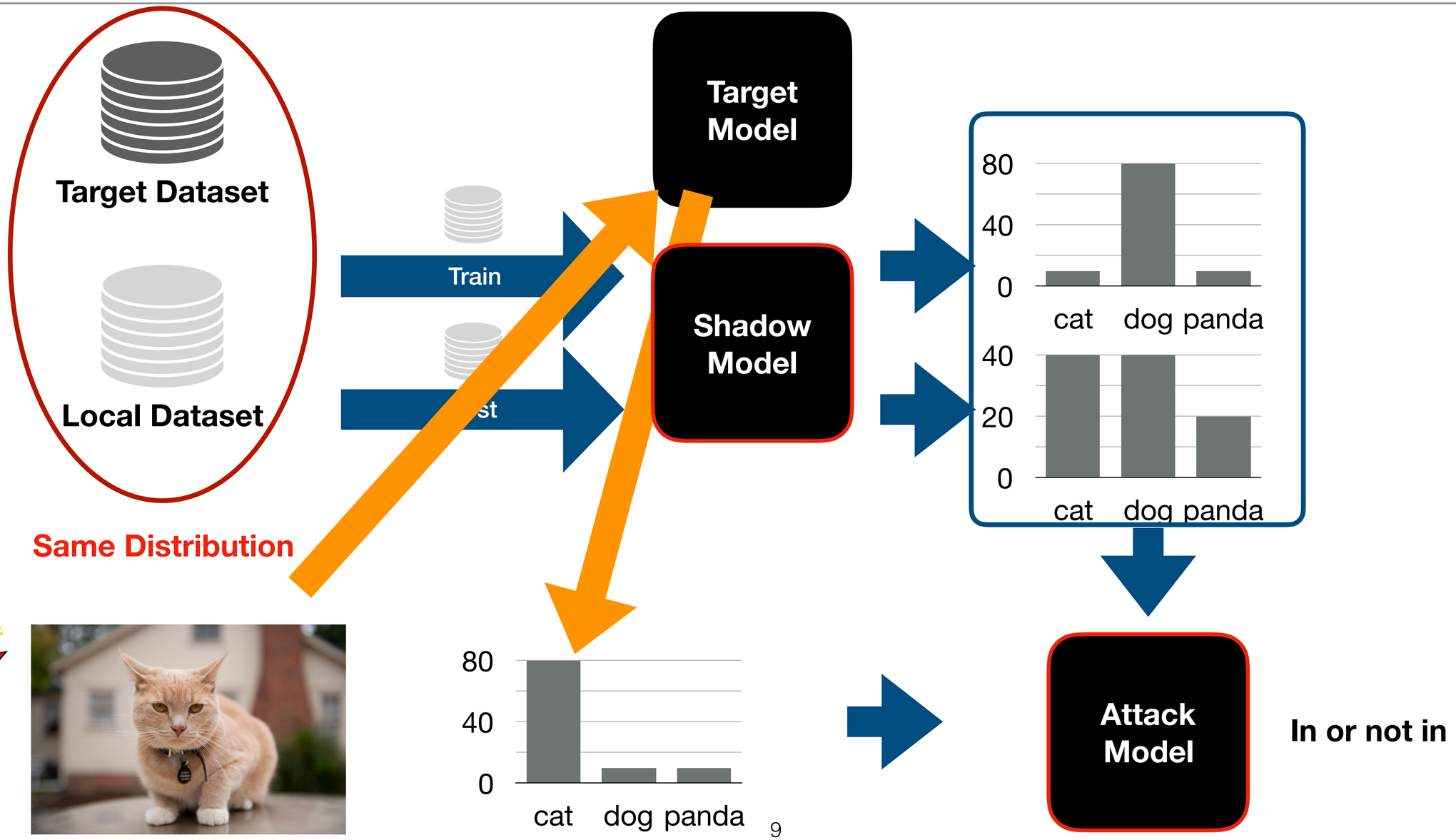
Our First Attack (Adversary 1)



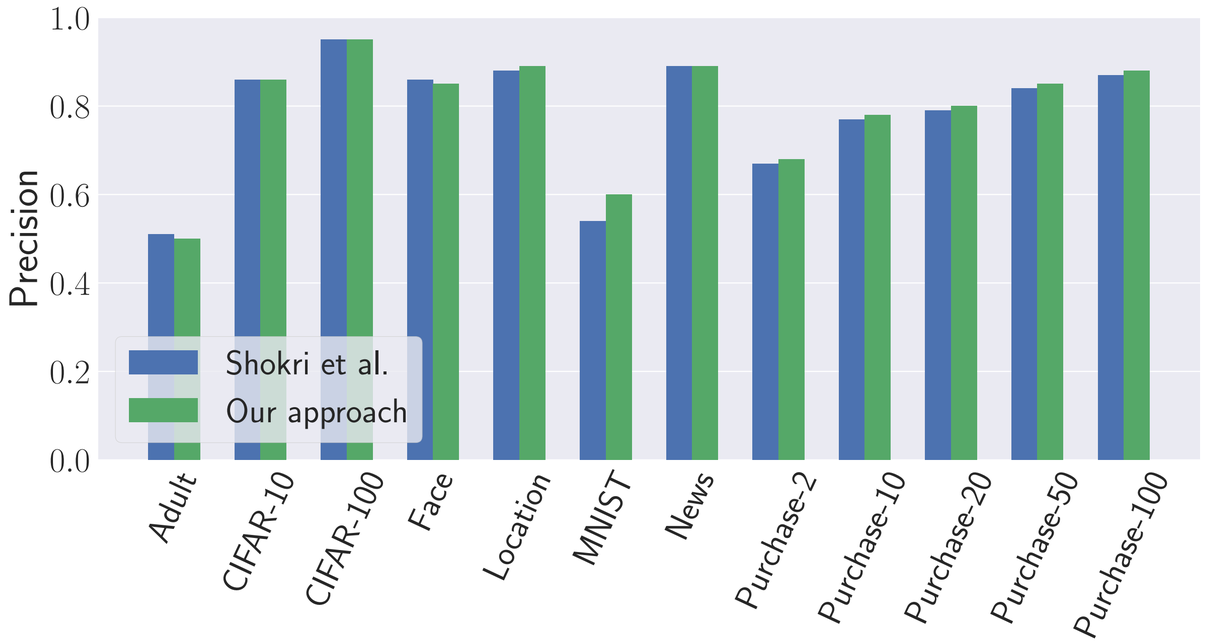
Our First Attack (Adversary 1)

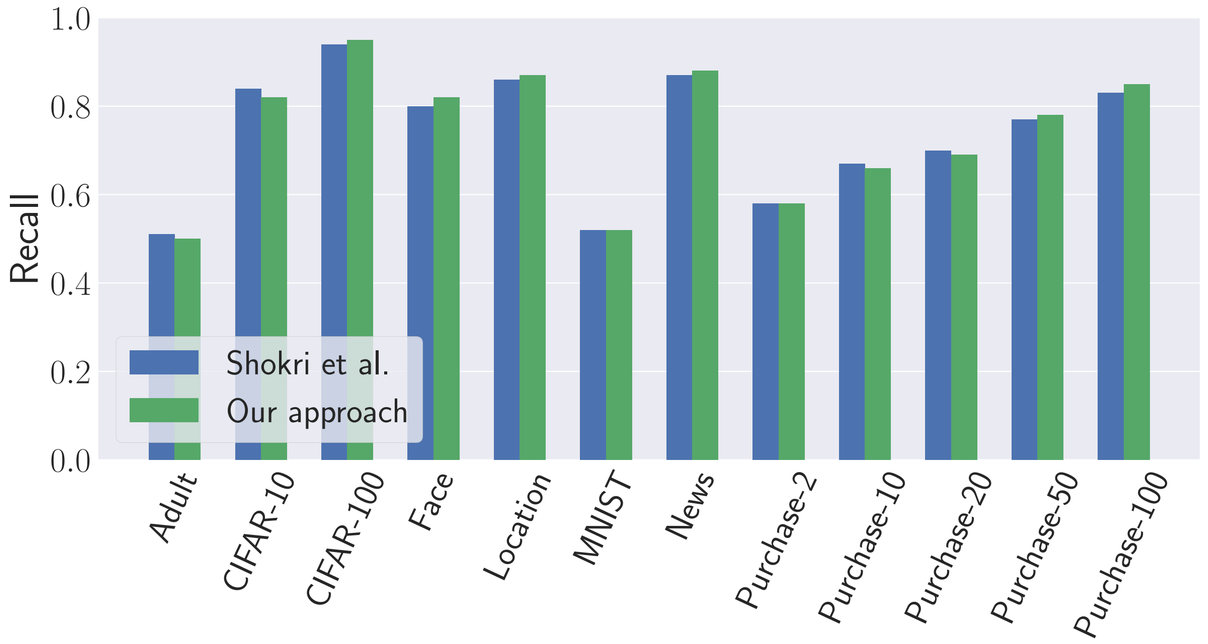


Our First Attack (Adversary 1)

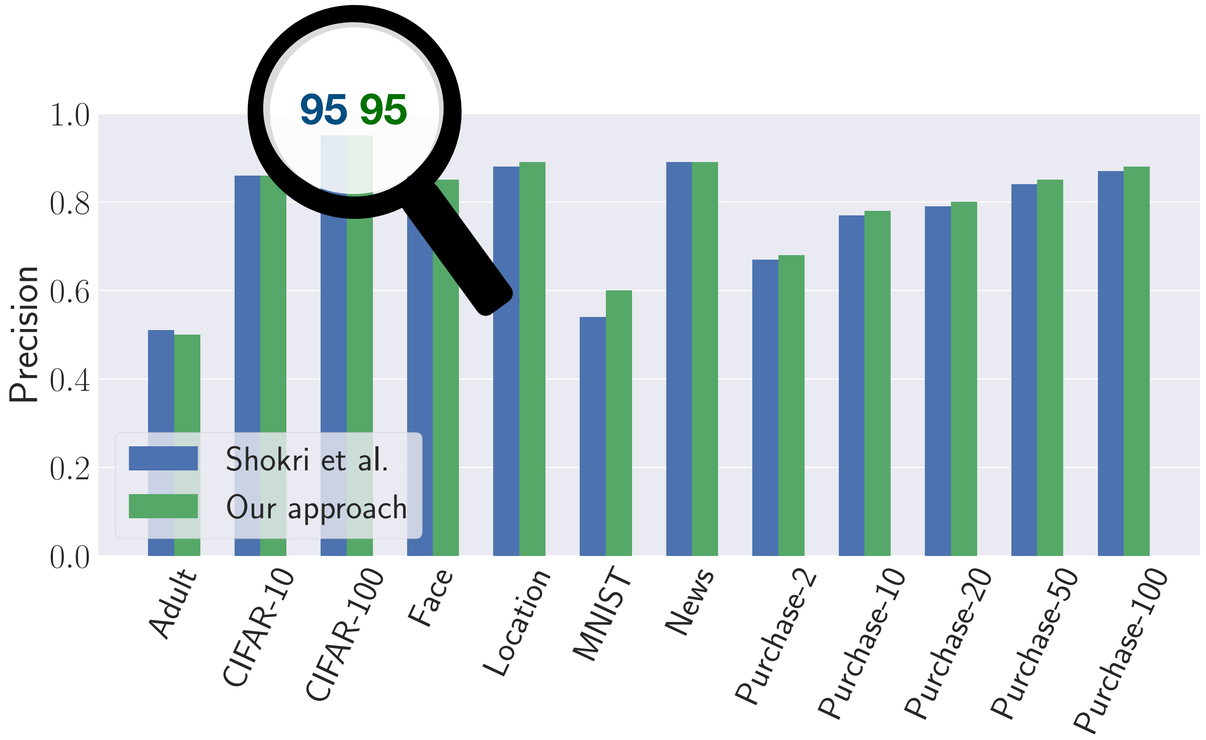


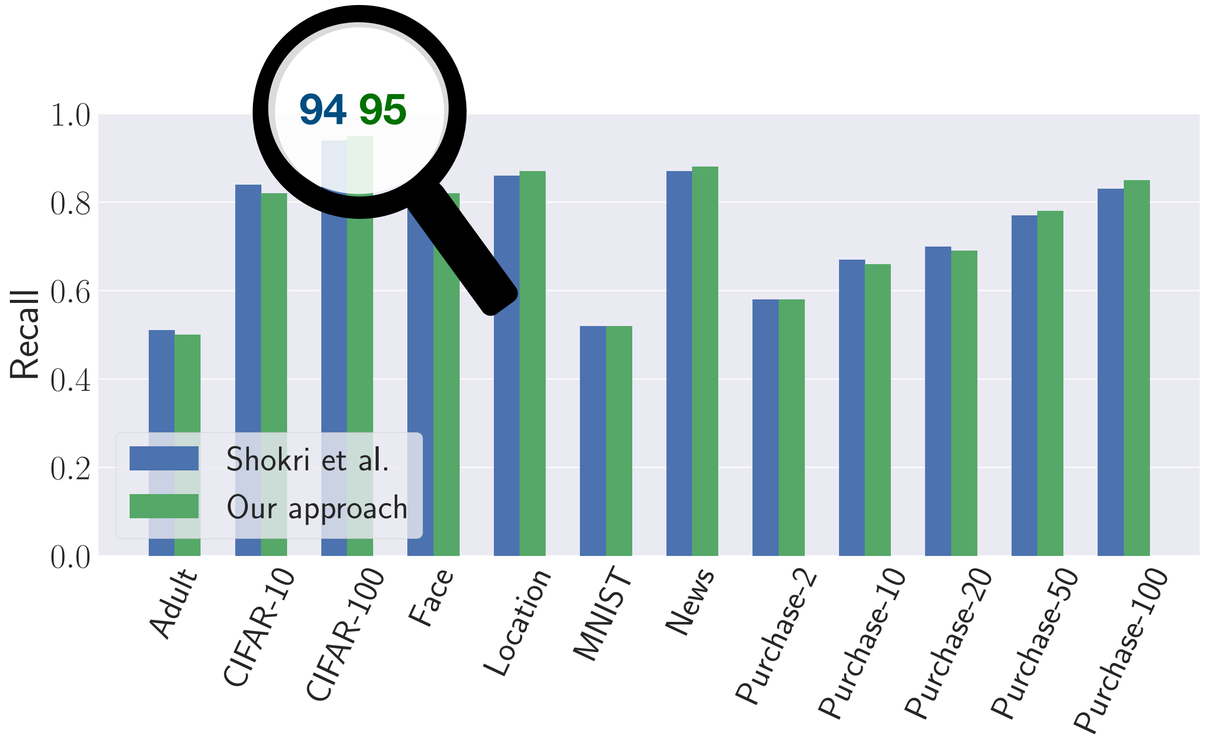
Performance Comparison



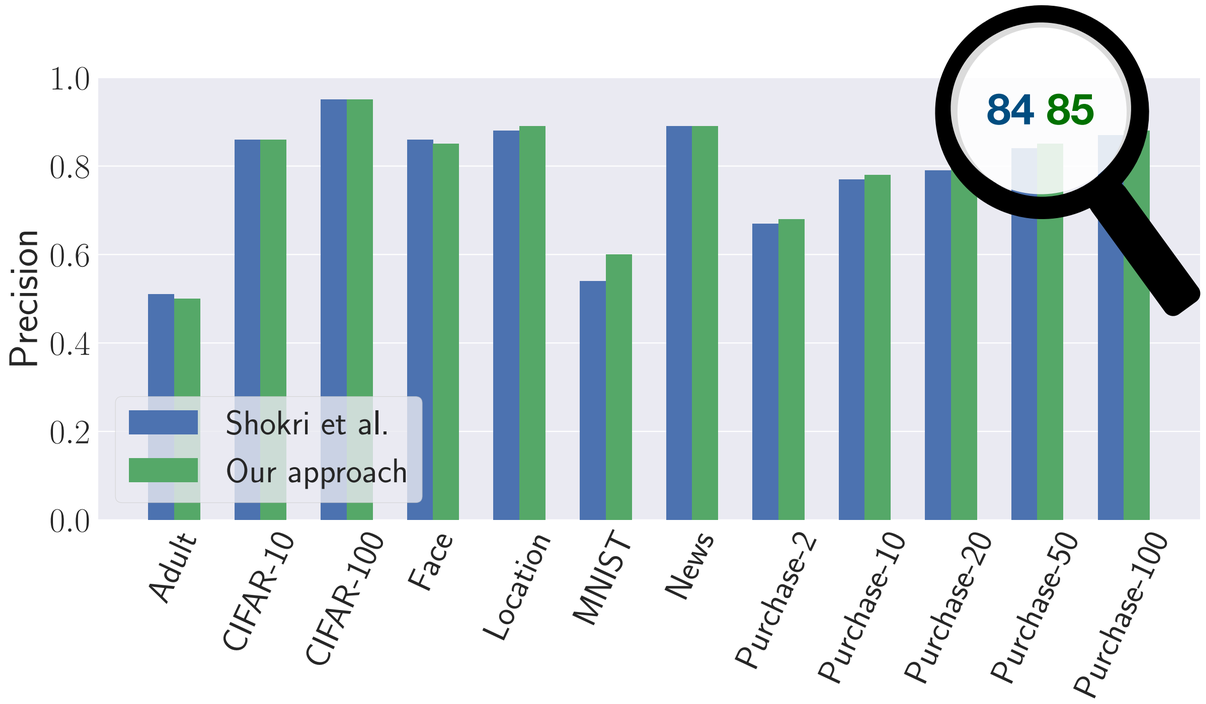


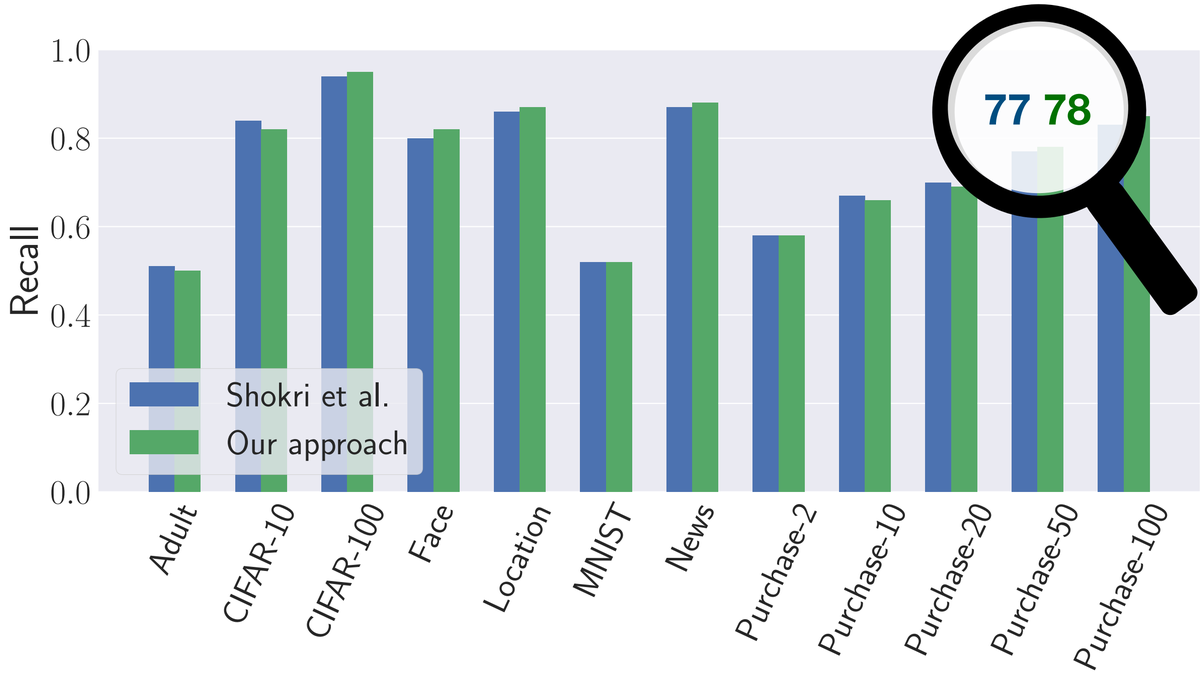
Performance Comparison



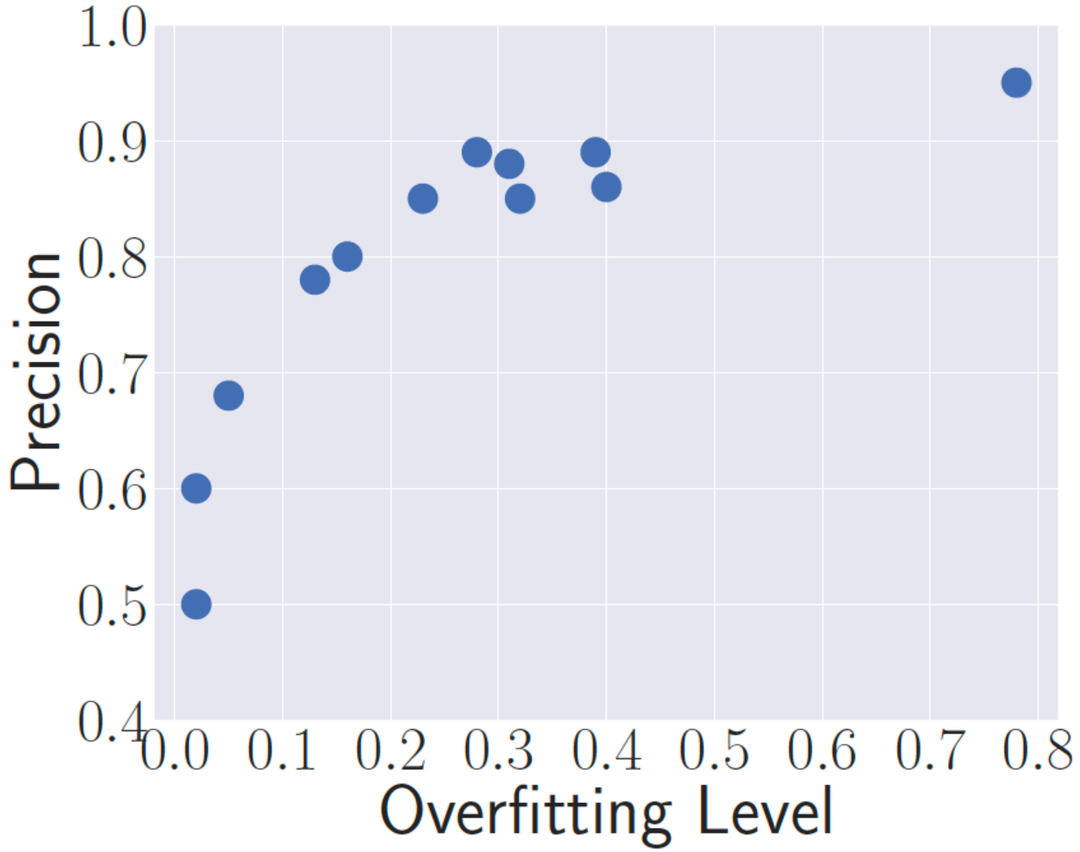


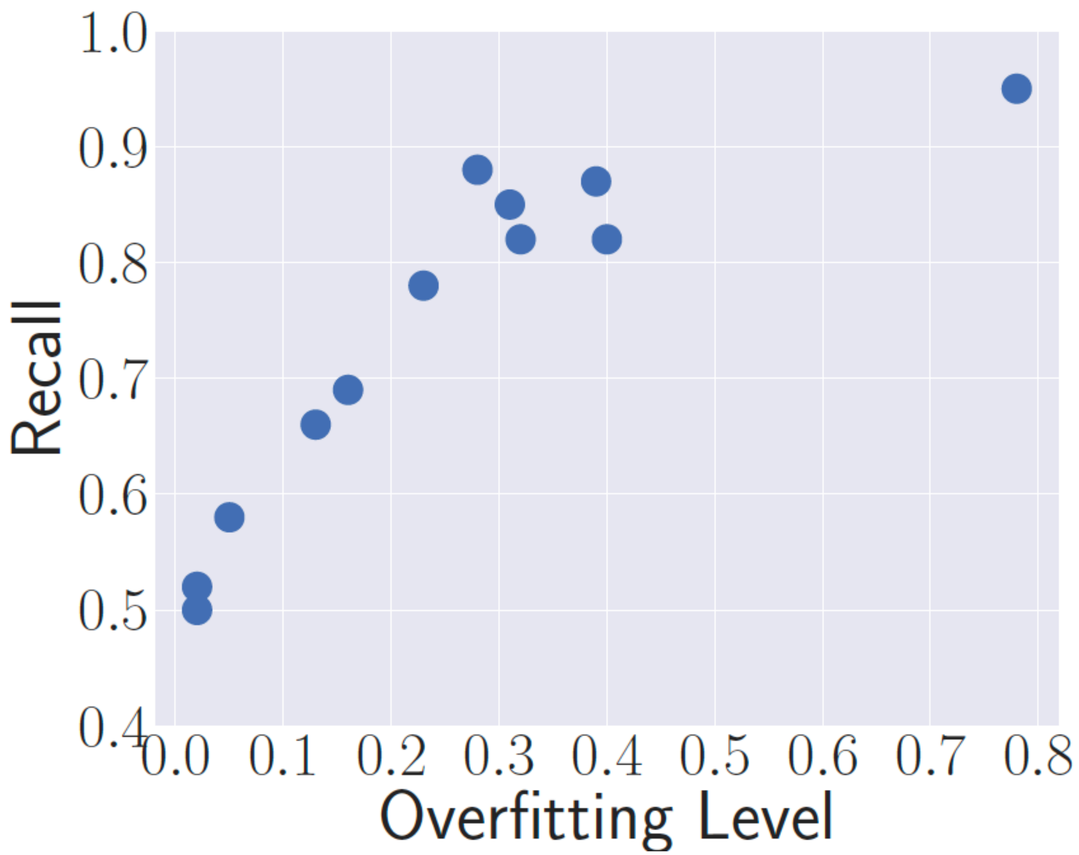
Performance Comparison





Overfitting vs Attack's Performance





Towards a Data Independent Attack

Towards a Data Independent Attack

CISPA
HELMHOLTZ CENTER FOR INFORMATION SECURITY

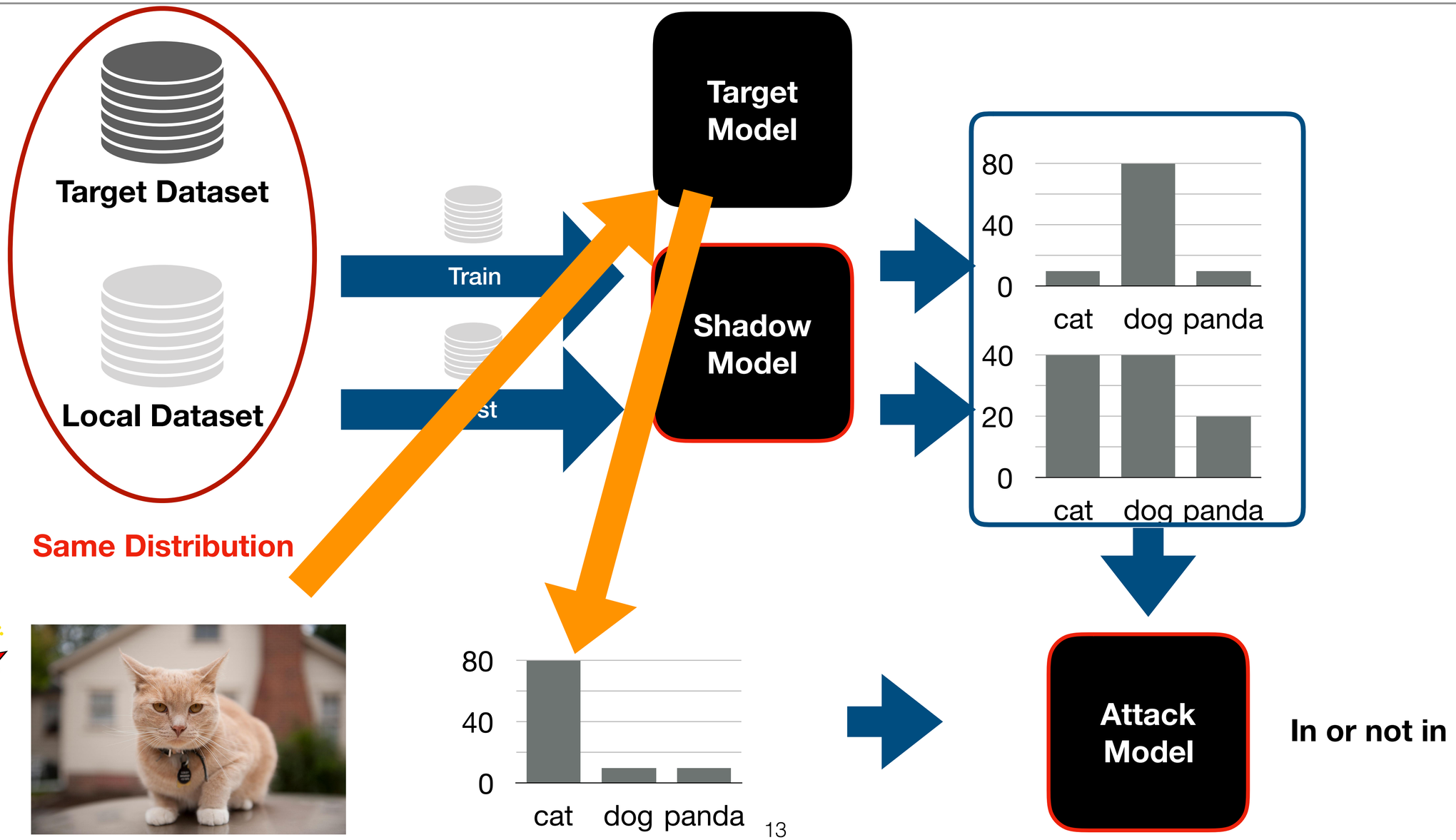
- Assumption on the datasets' distribution

Towards a Data Independent Attack

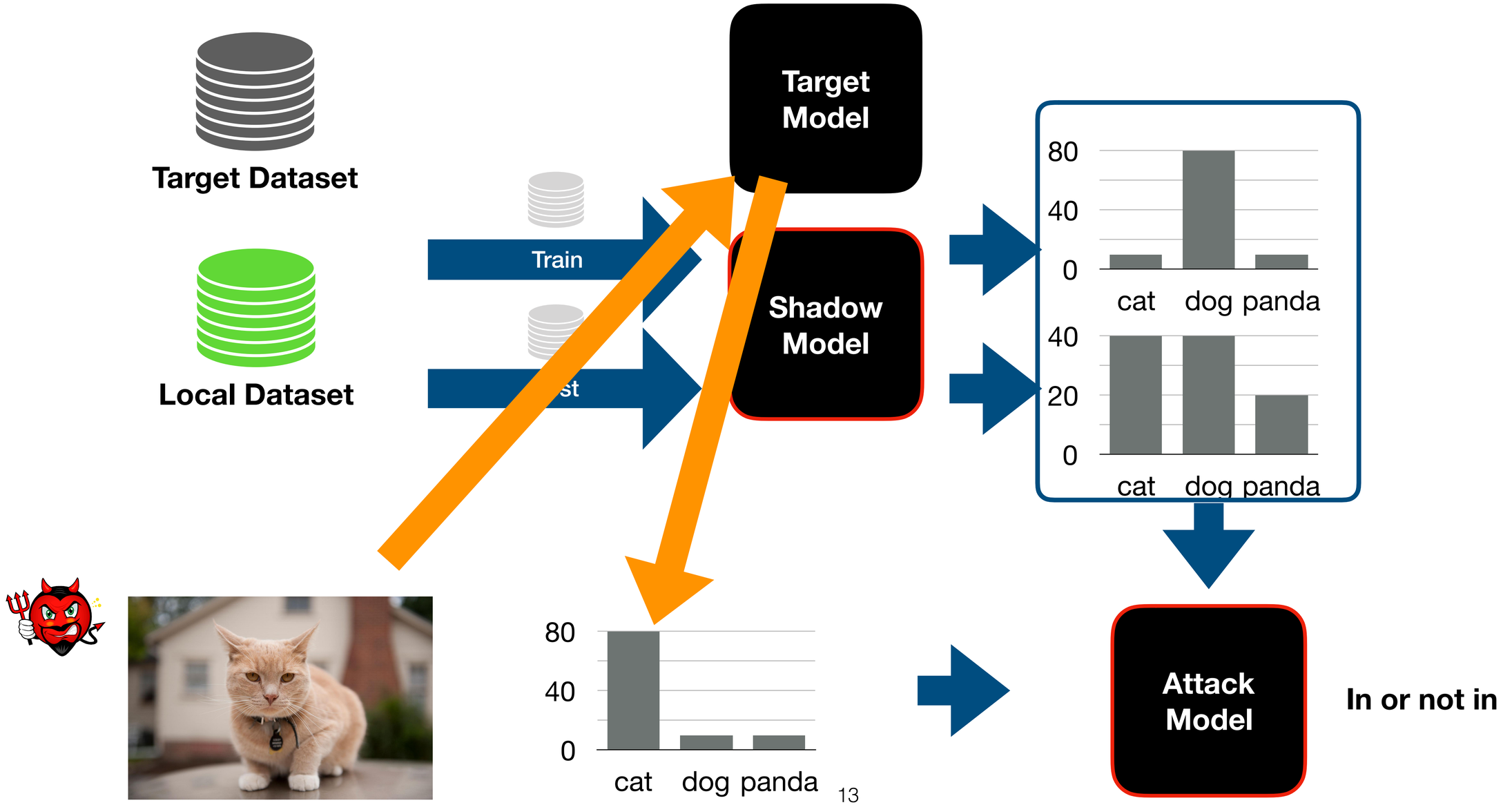
CISPA
HELMHOLTZ CENTER FOR INFORMATION SECURITY

- Assumption on the datasets' distribution
  - Data transferring attack

Data Transferring Attack (Adversary 2)



Data Transferring Attack (Adversary 2)



Data Transferring Attack (Adversary 2)

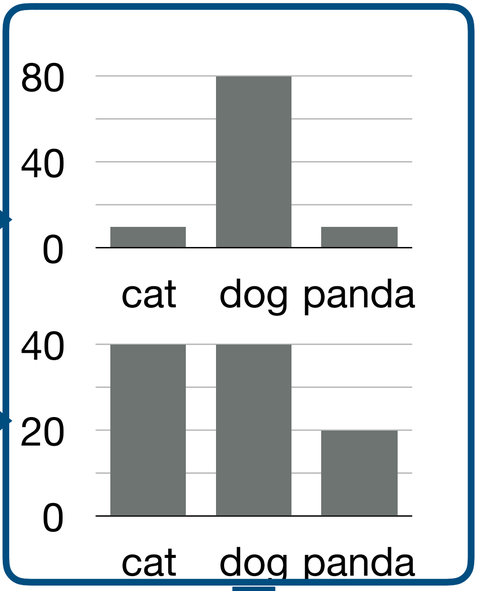
Target Dataset

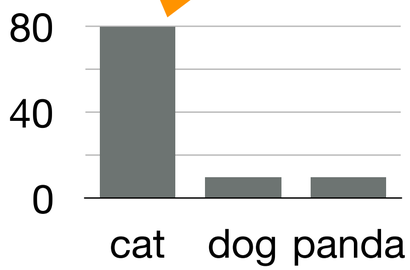
Local Dataset

Train

Target Model

Shadow Model

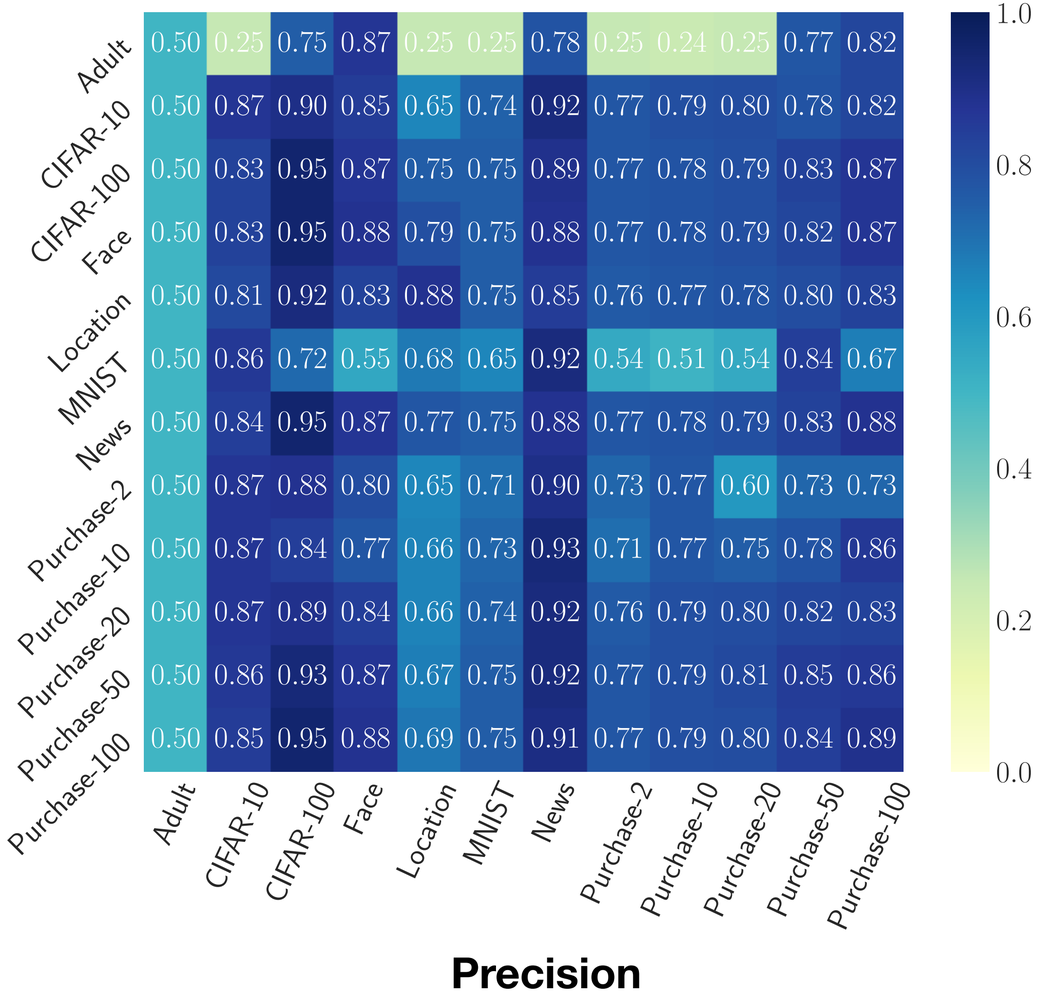




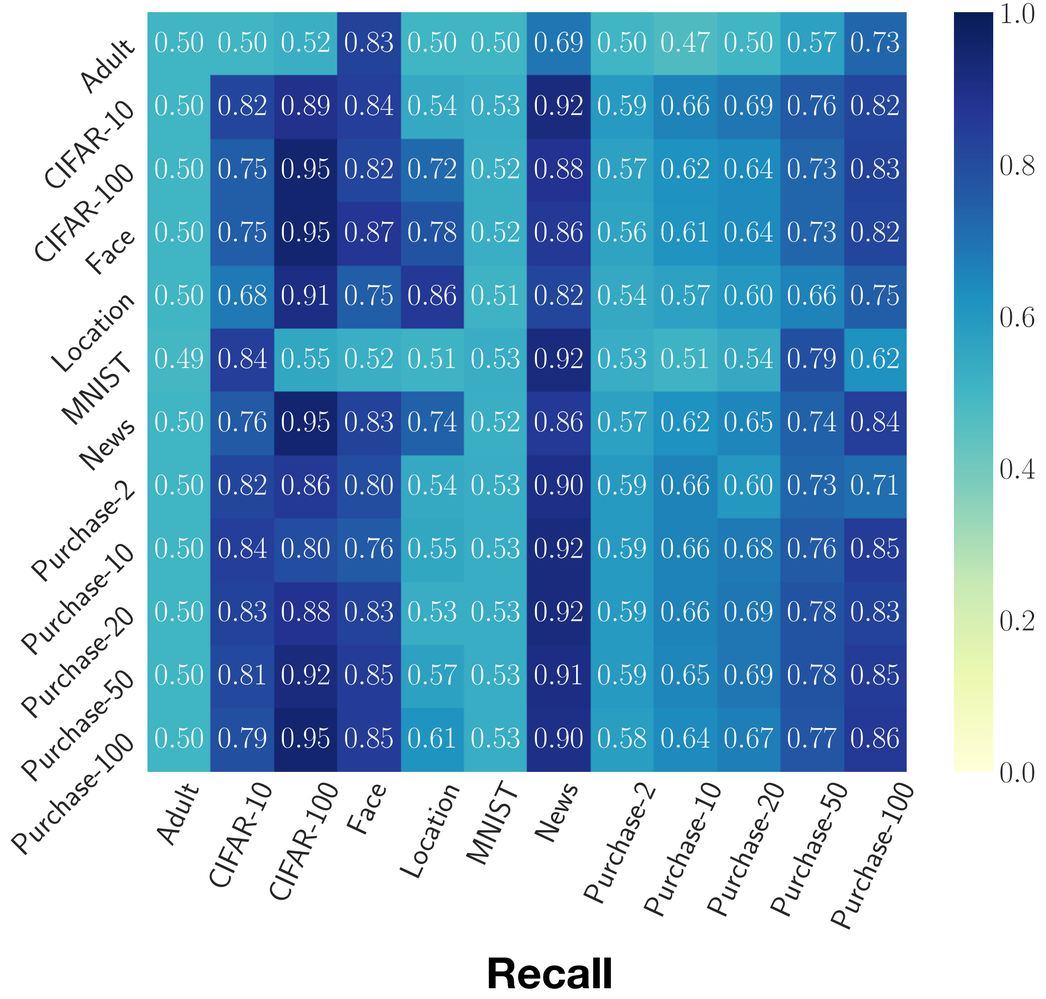
Attack Model

In or not in

Performance

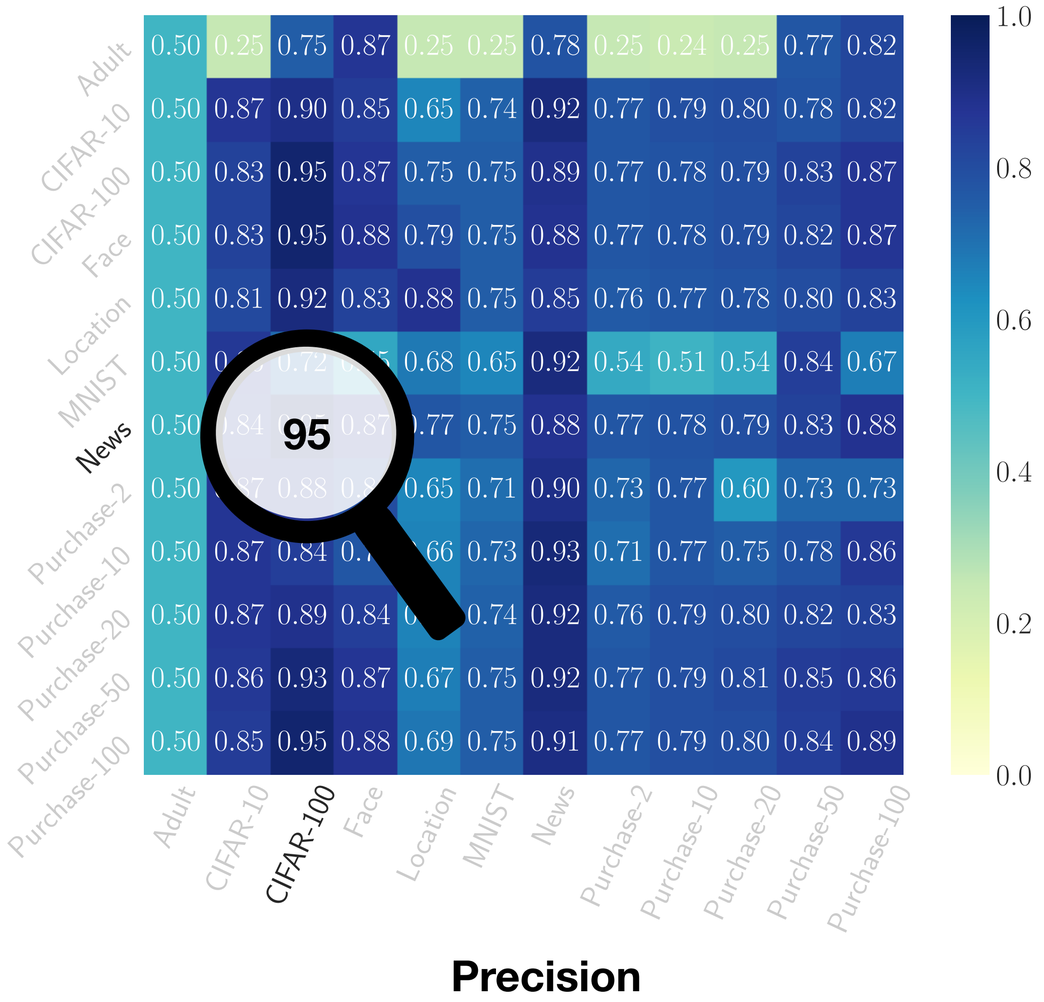


Precision

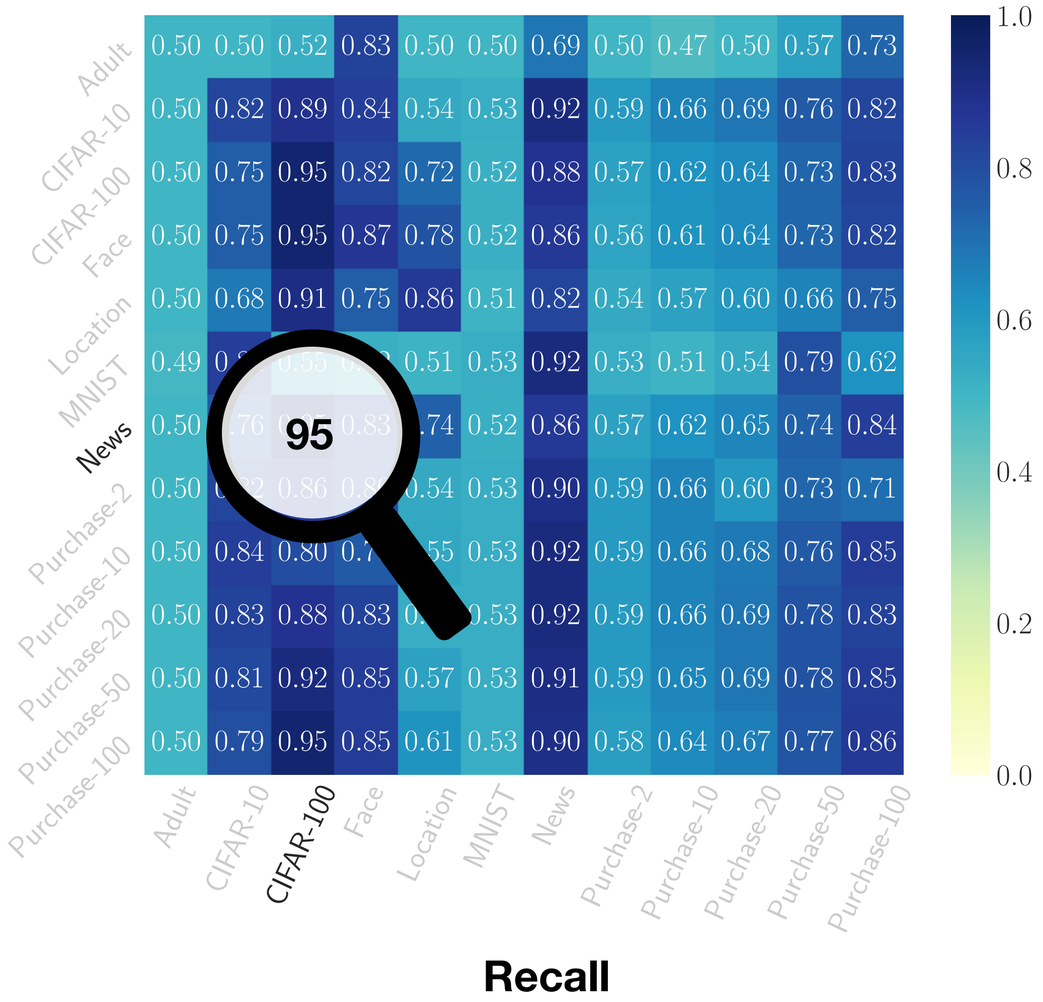


Recall

Performance

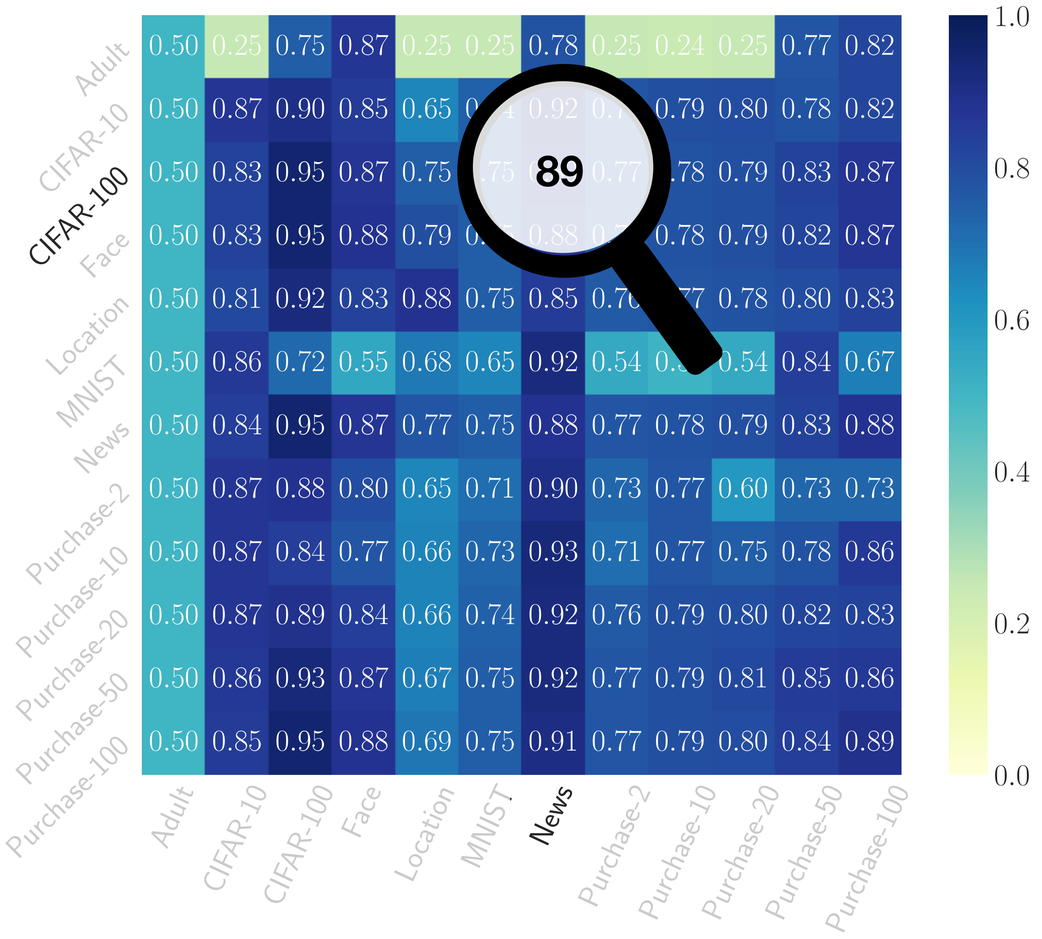


Precision

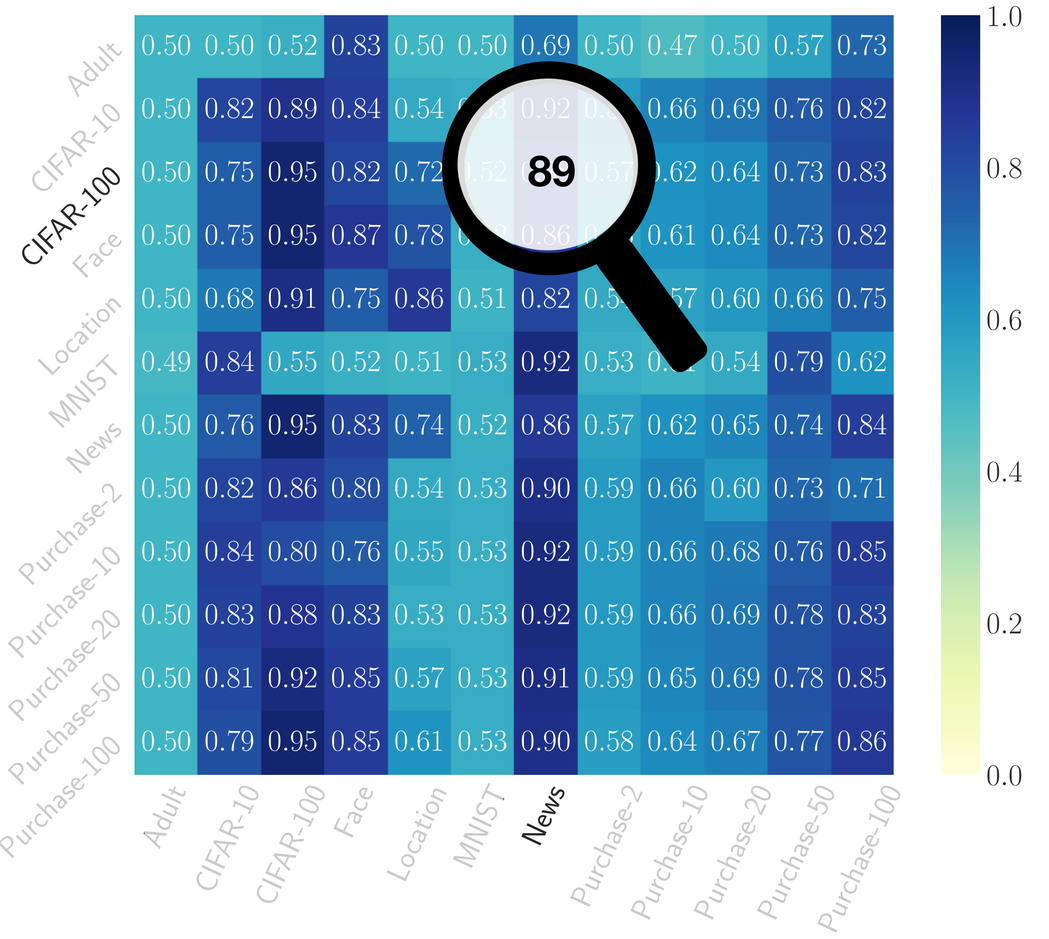


Recall

Performance



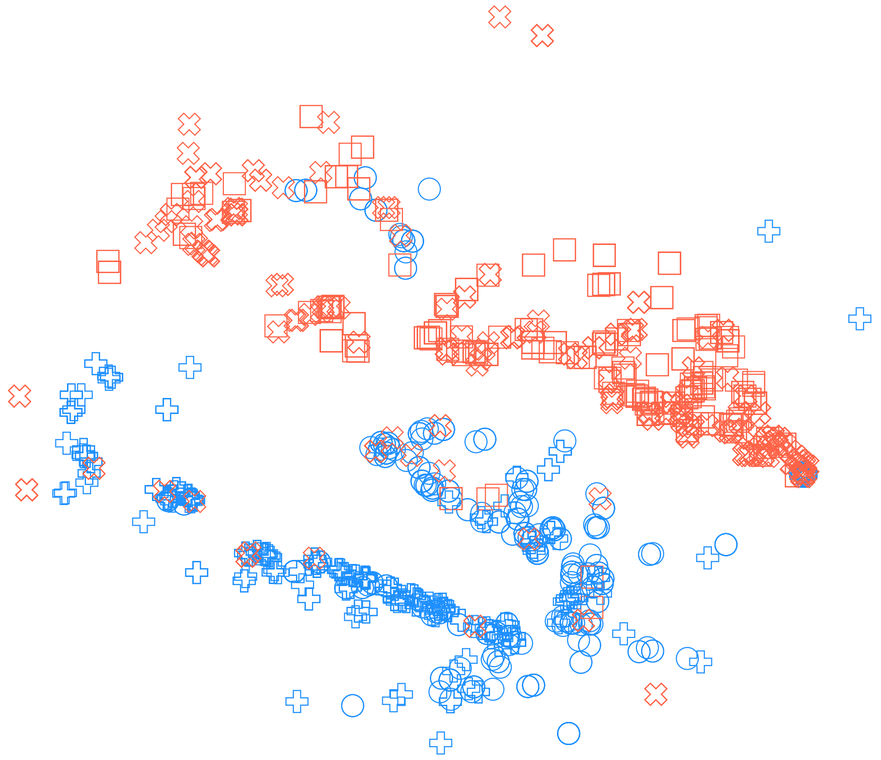
Precision



Recall

Sounds Magic, Why?

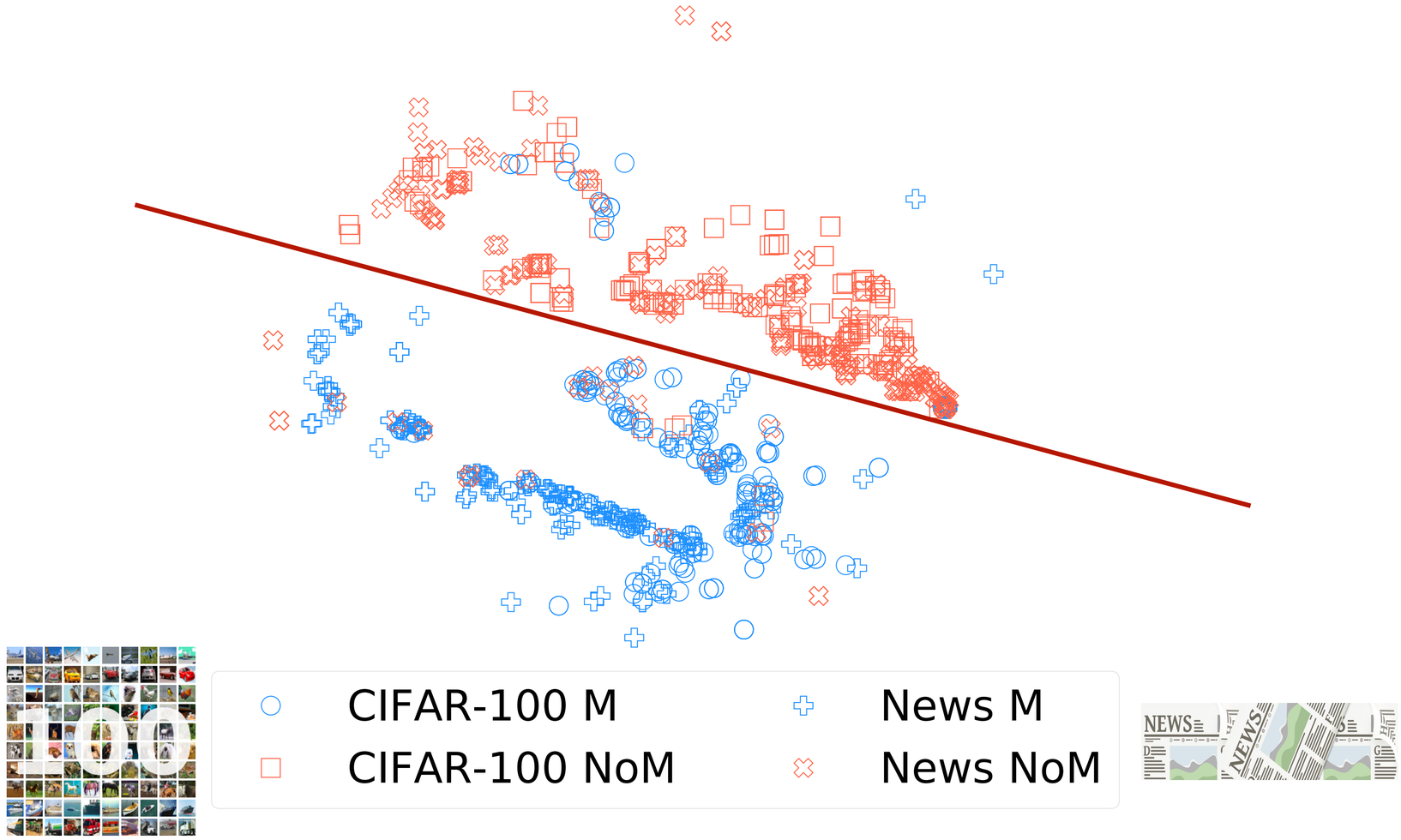
CISPA
HELMHOLTZ CENTER FOR INFORMATION SECURITY



CIFAR-100 M
CIFAR-100 NoM
News M
News NoM

Sounds Magic, Why?

CISPA
HELMHOLTZ CENTER FOR INFORMATION SECURITY



Towards an Unsupervised Attack

Towards an Unsupervised Attack

CISPA
HELMHOLTZ CENTER FOR INFORMATION SECURITY

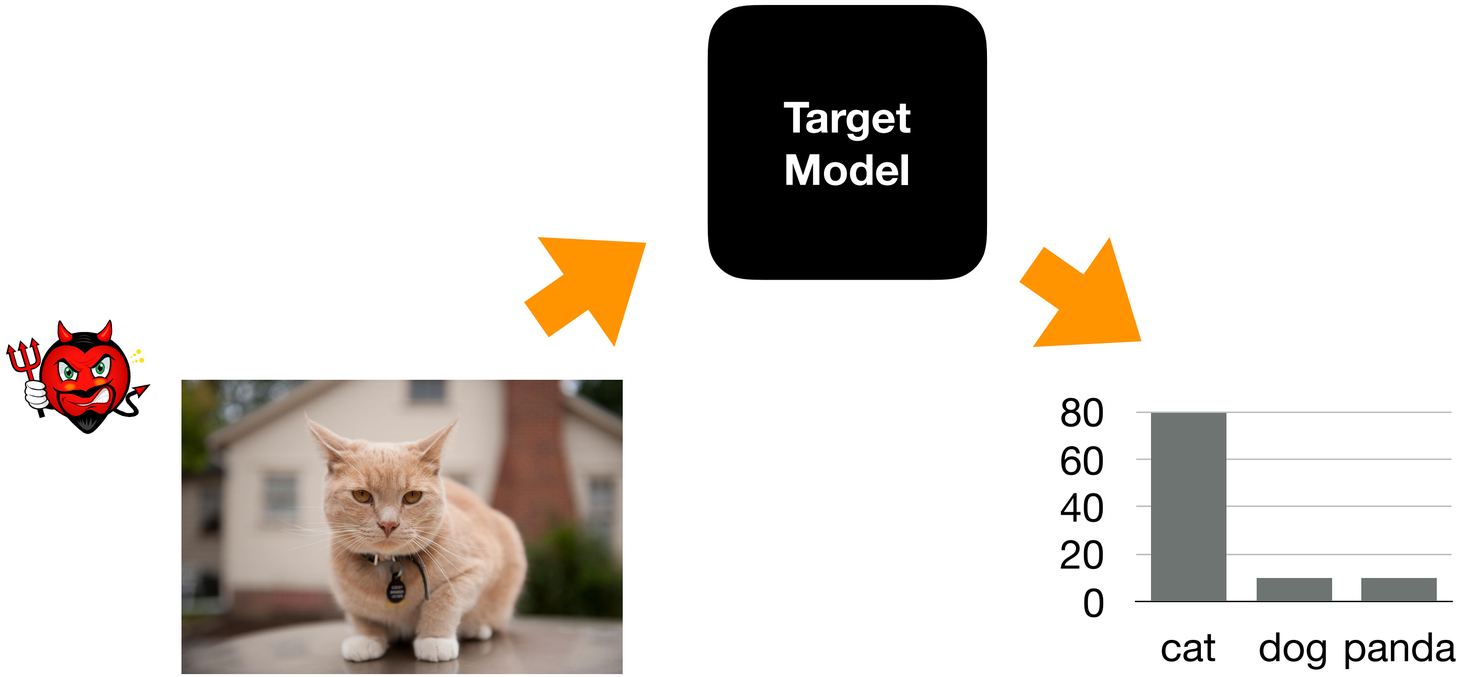
- Shadow model

Towards an Unsupervised Attack

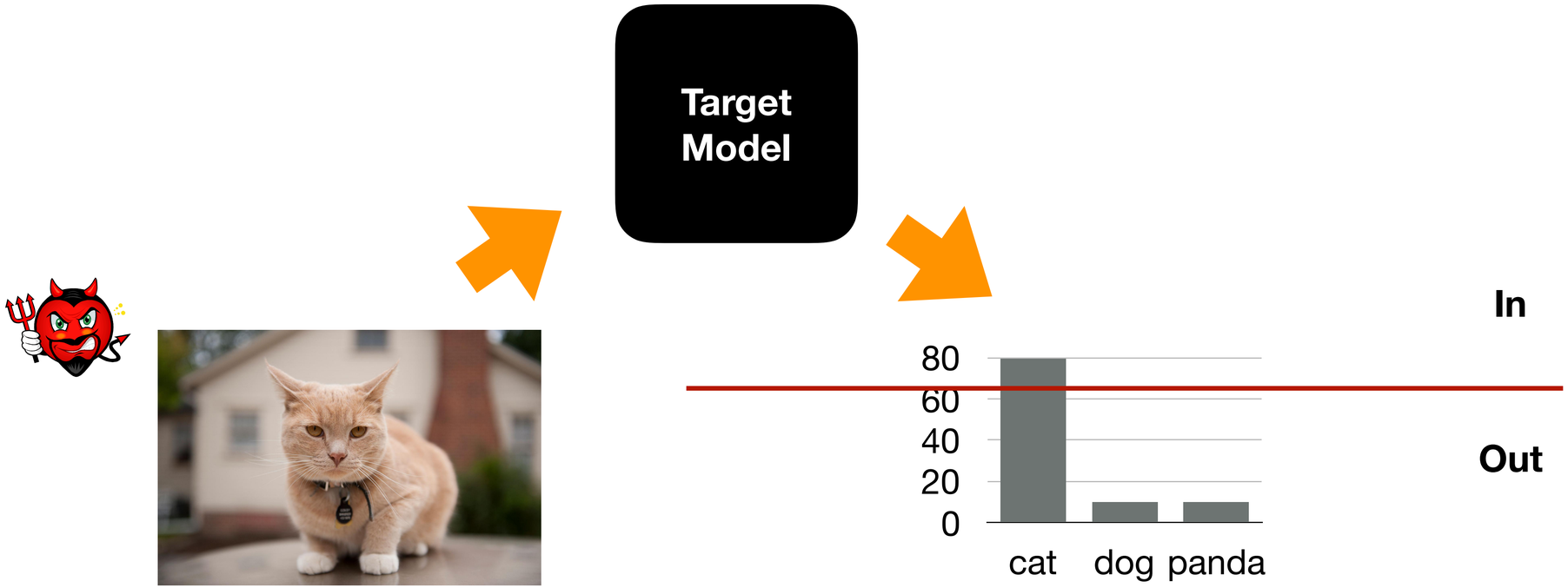
CISPA
HELMHOLTZ CENTER FOR INFORMATION SECURITY

- Shadow model
- Attack model

Our Third Attack (Adversary 3)



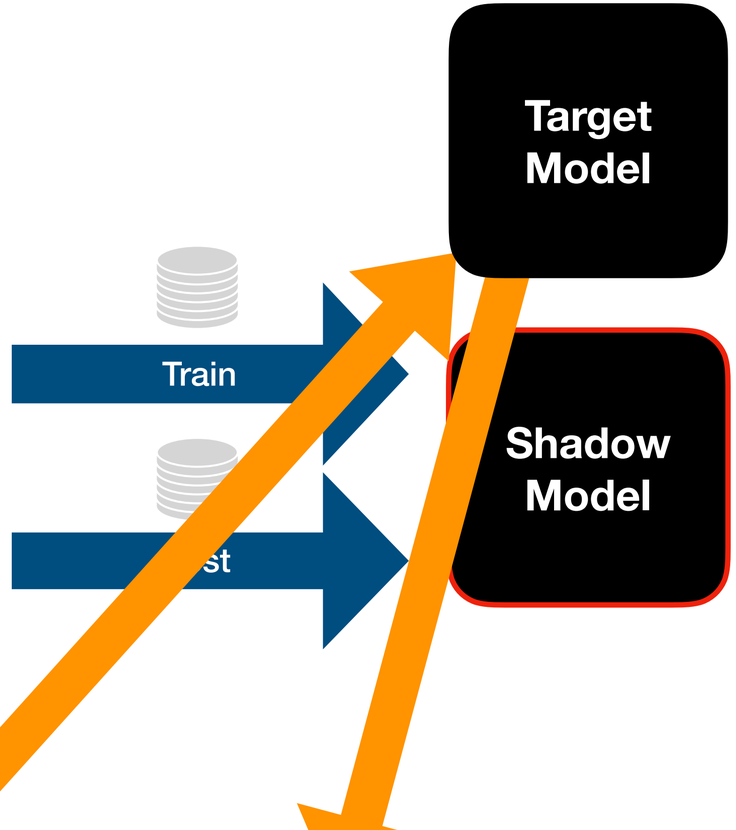
Our Third Attack (Adversary 3)

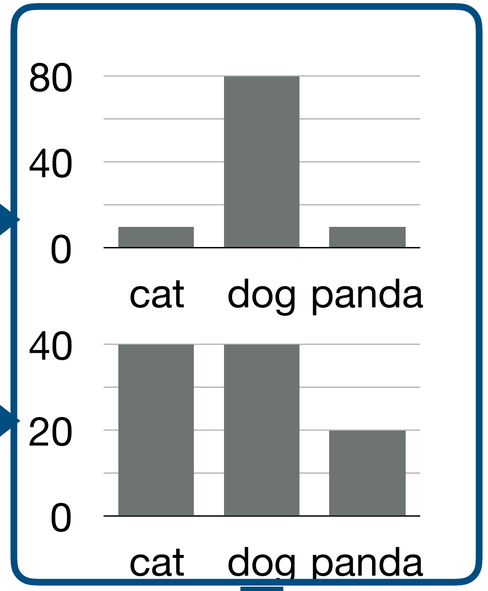


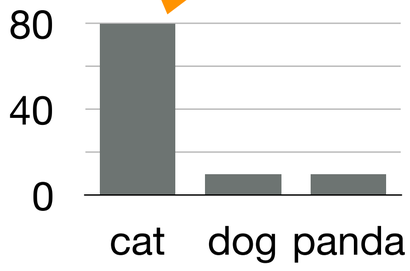
Our Third Attack (Adversary 3)

Target Dataset

Local Dataset







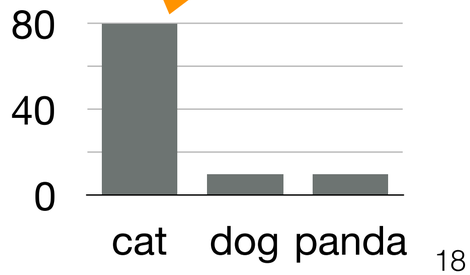
Attack Model

In or not in

Our Third Attack (Adversary 3)

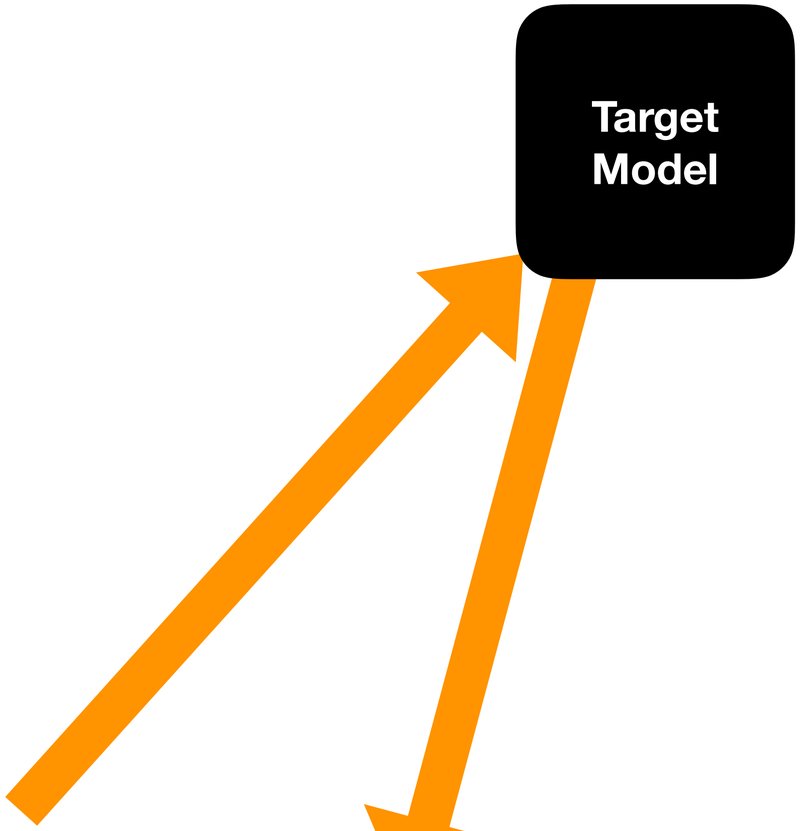
Target Dataset

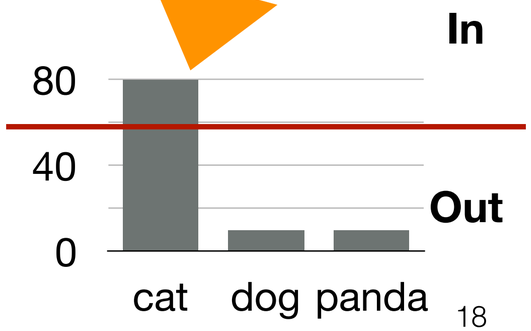
Target Model



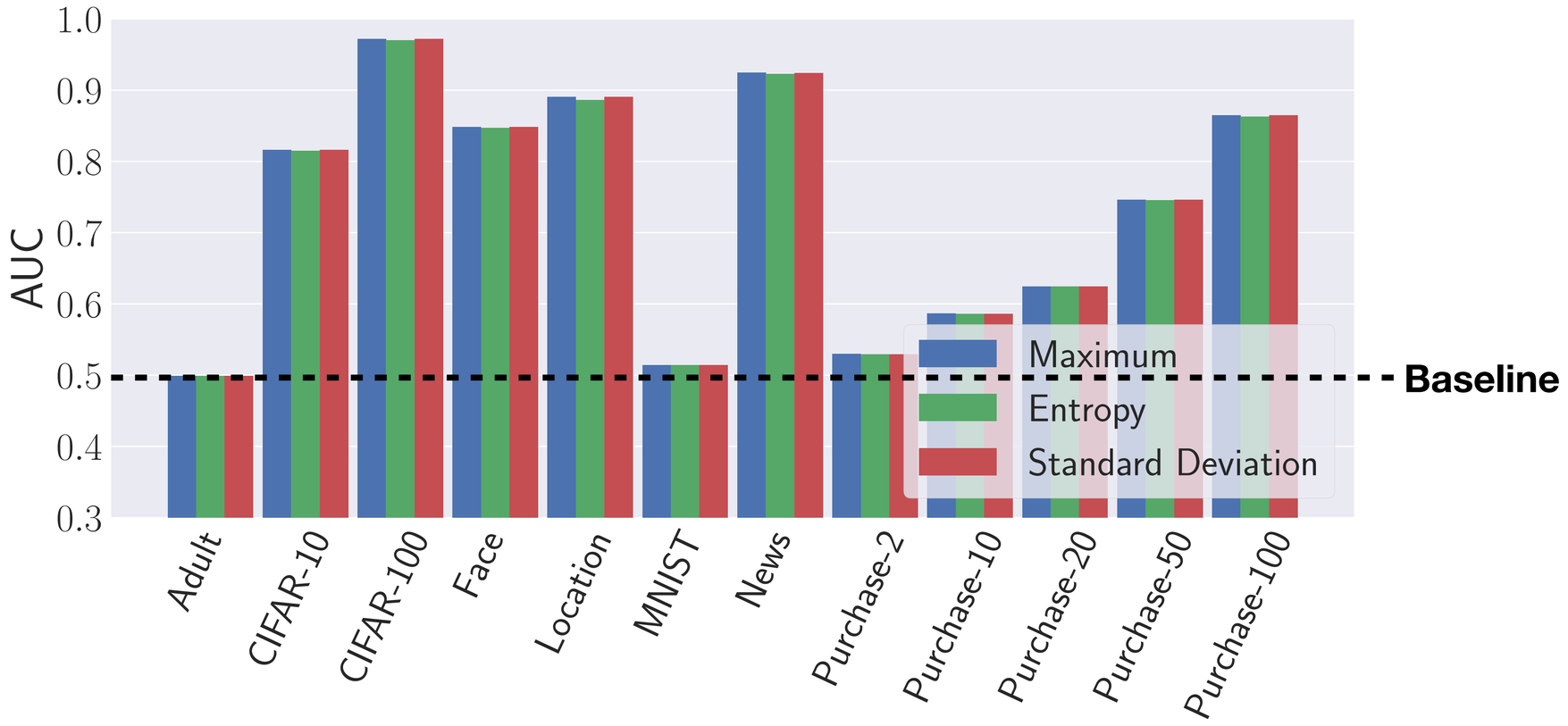
Our Third Attack (Adversary 3)

Target Dataset



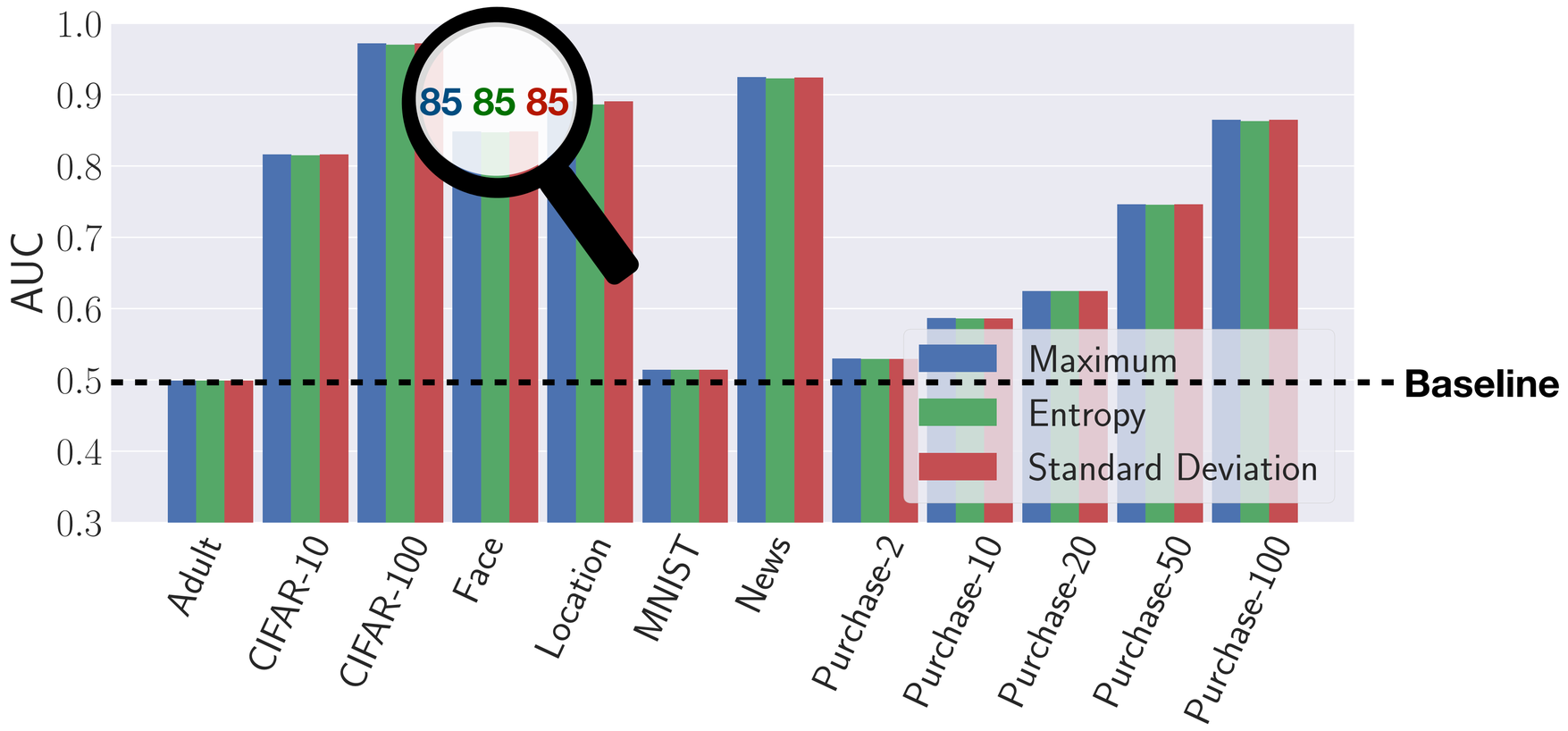


Performance



Performance

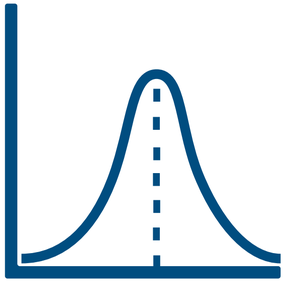
Performance



Threshold Picking

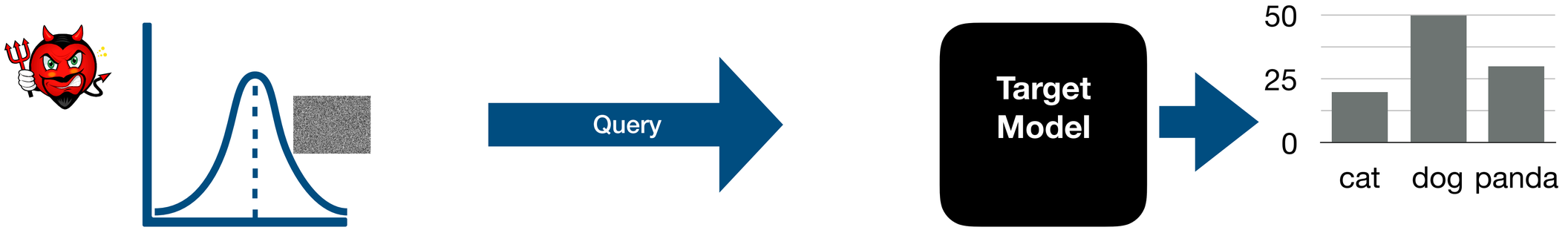
Target Model

Threshold Picking

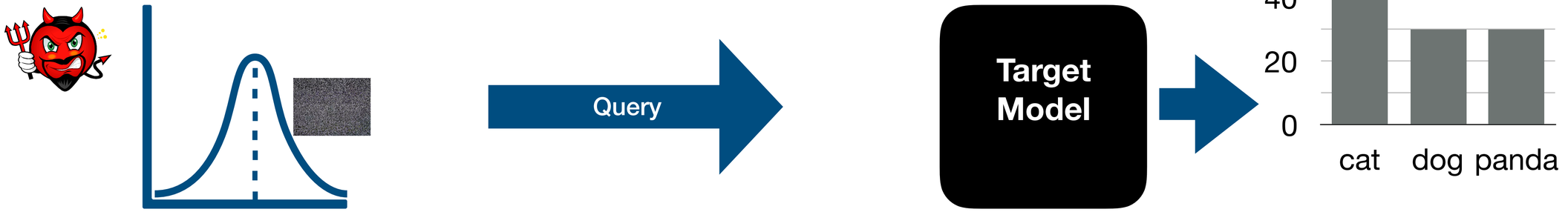


Target Model

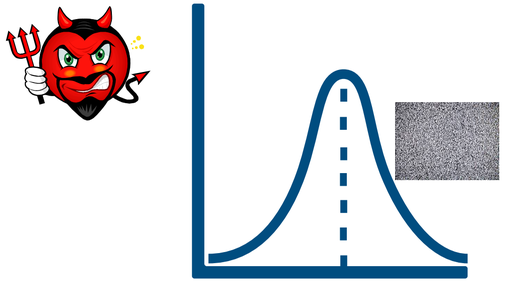
Threshold Picking



Threshold Picking

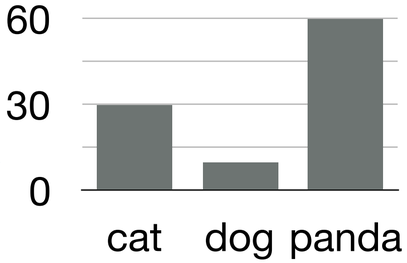


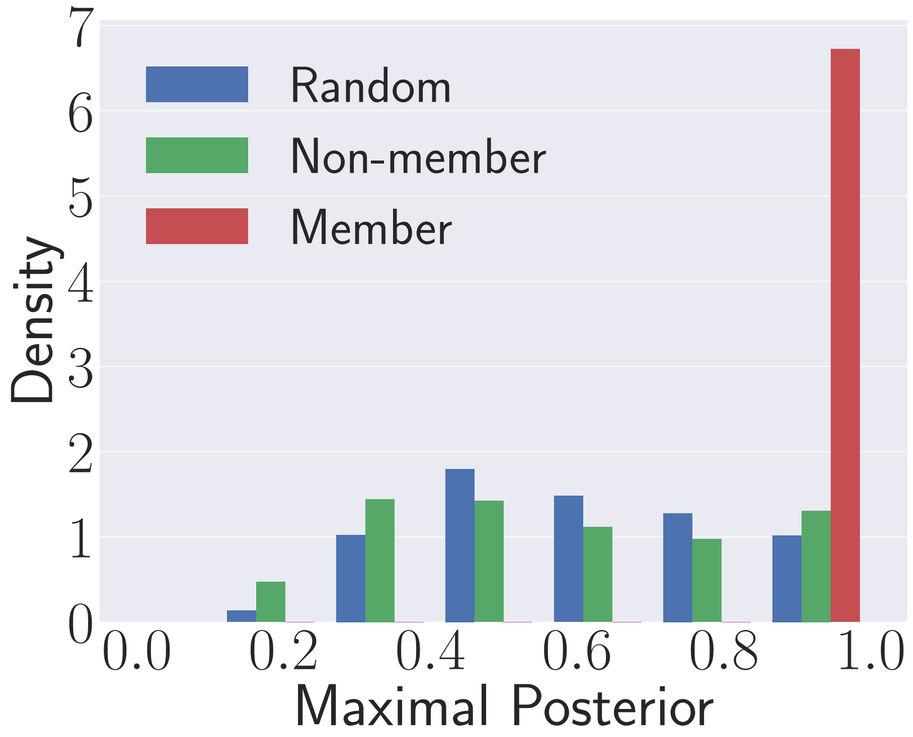
Threshold Picking



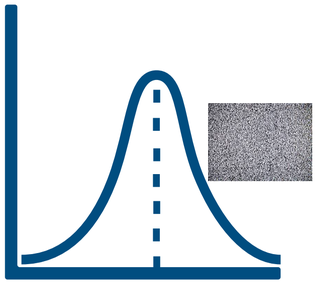
Query

Target Model



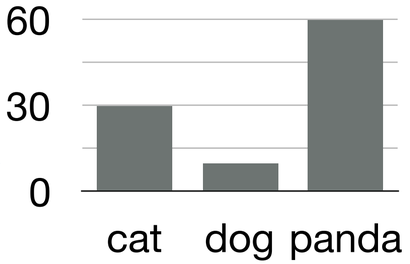


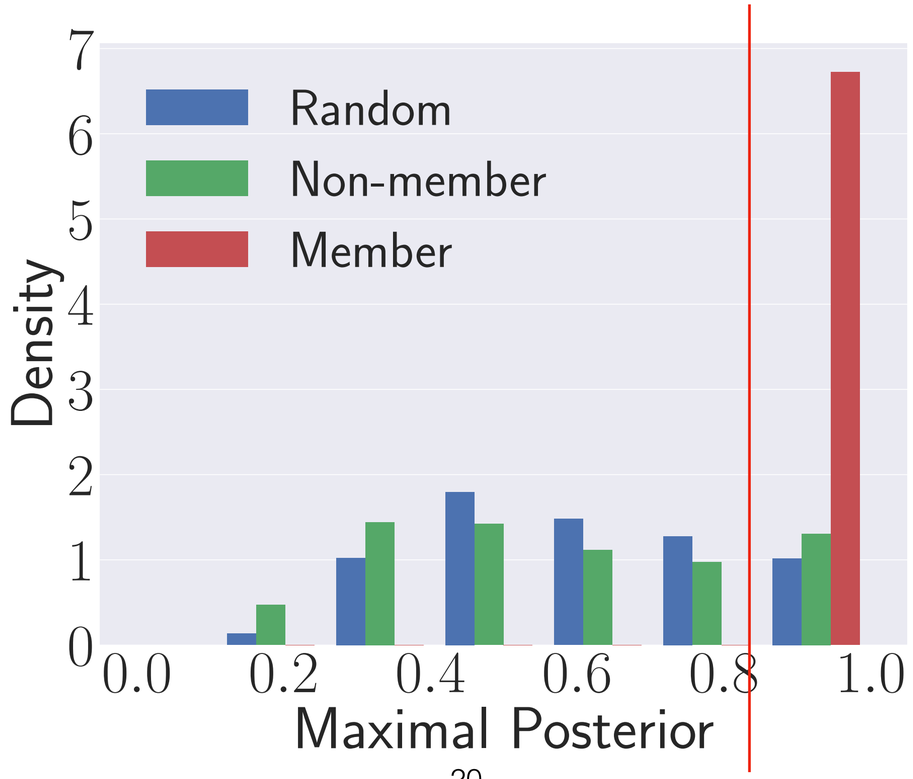
Threshold Picking



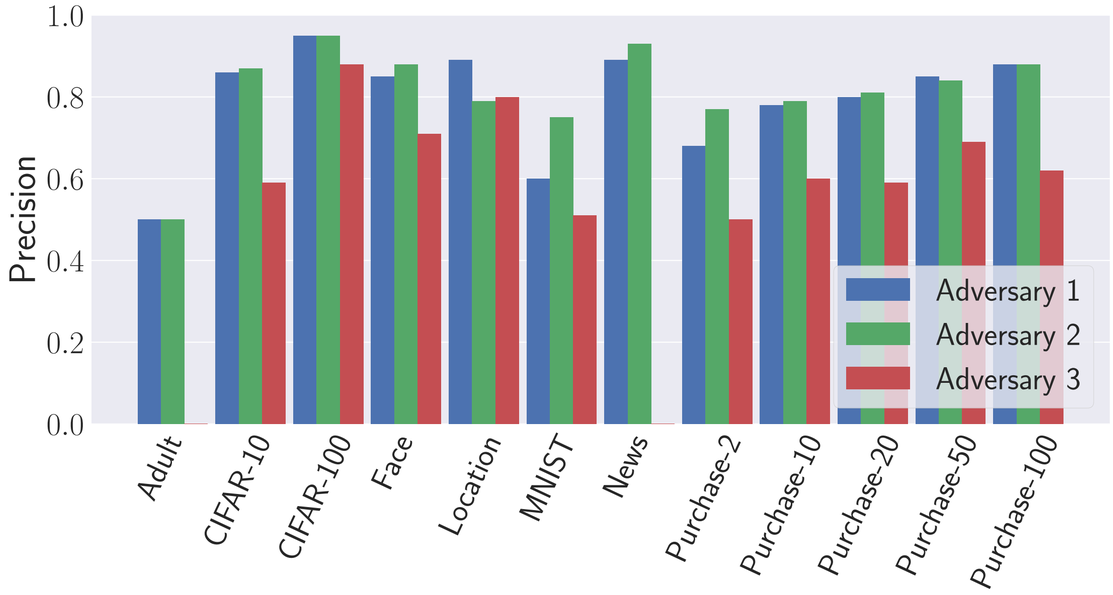
Query

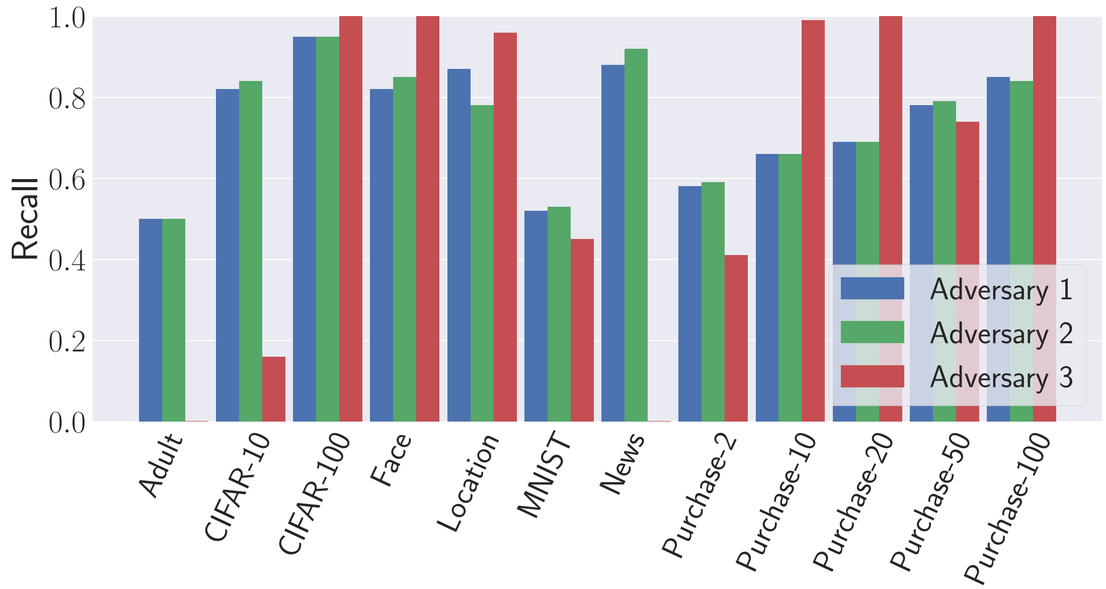
Target Model



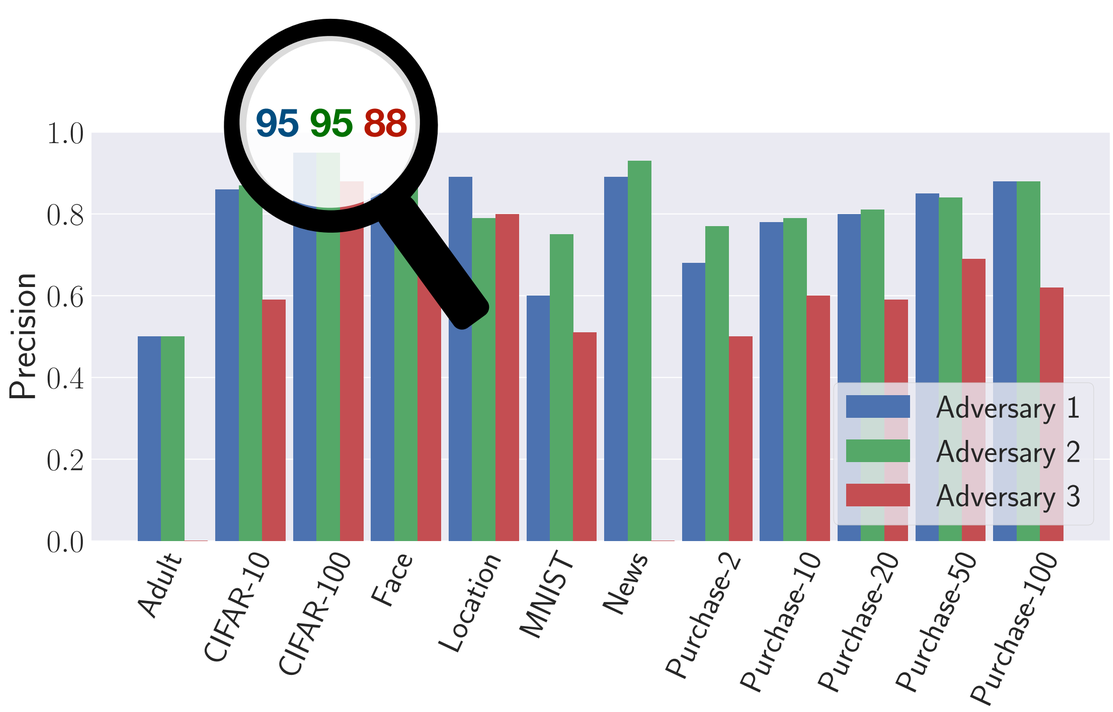


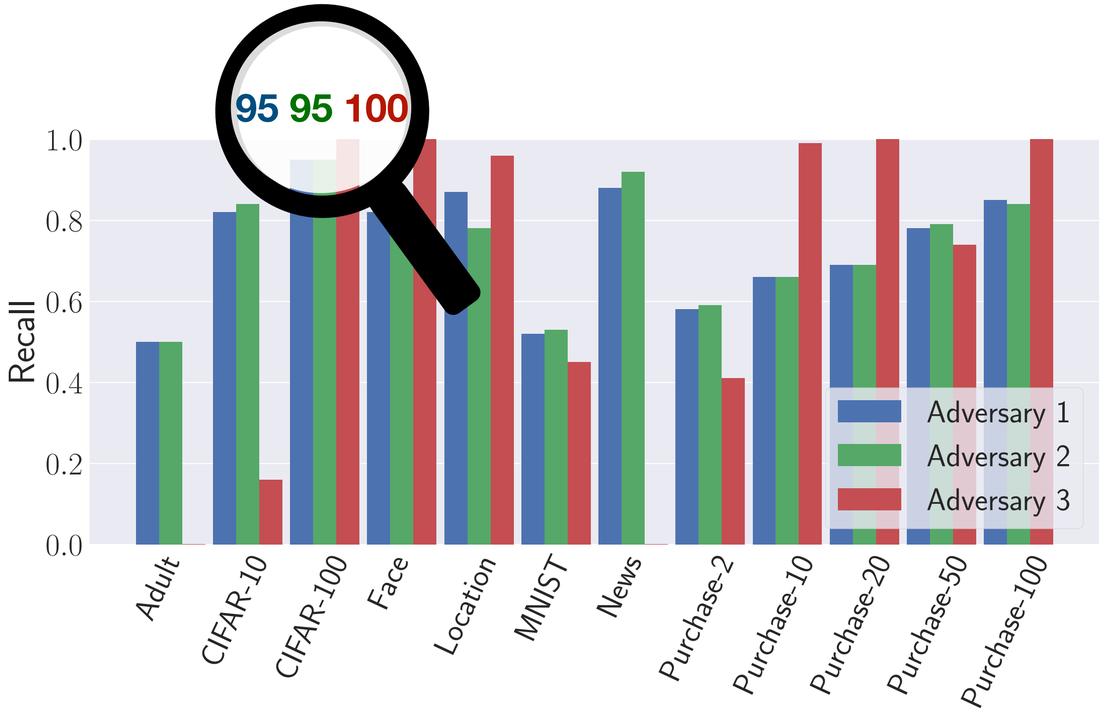
Comparing All Attacks





Comparing All Attacks

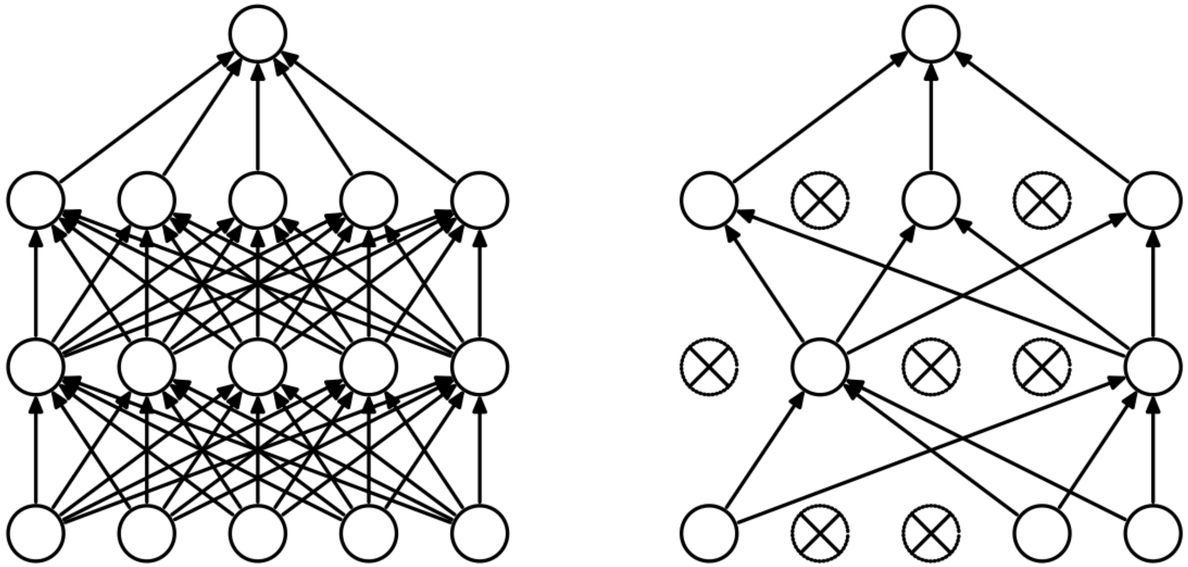




Defense

Defense

- Dropout

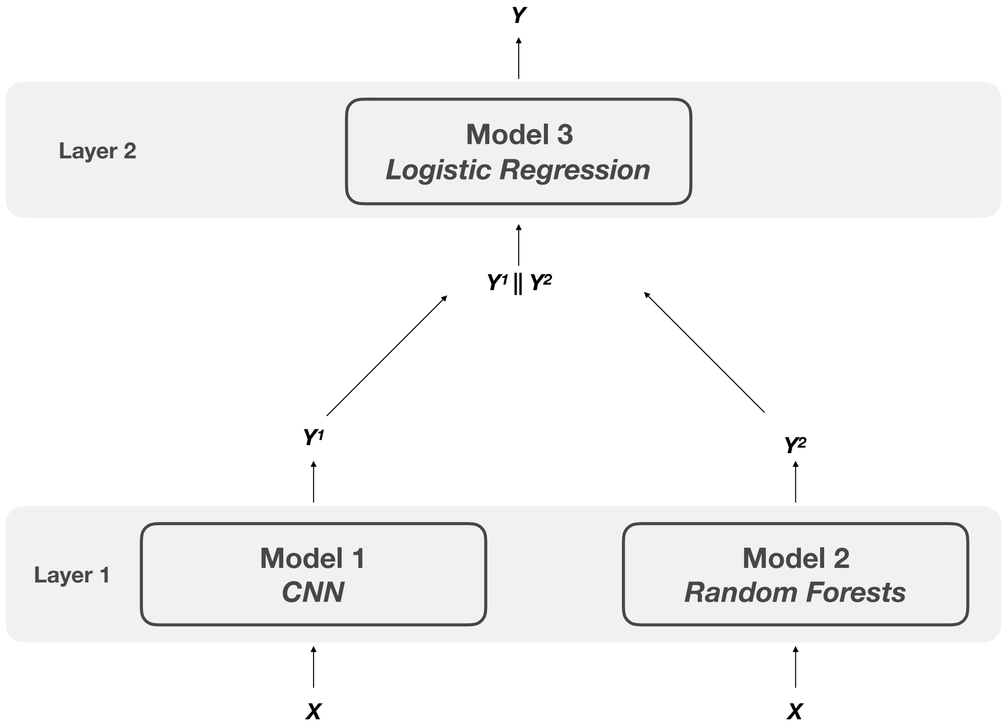


(a) Standard Neural Net (b) After applying dropout.

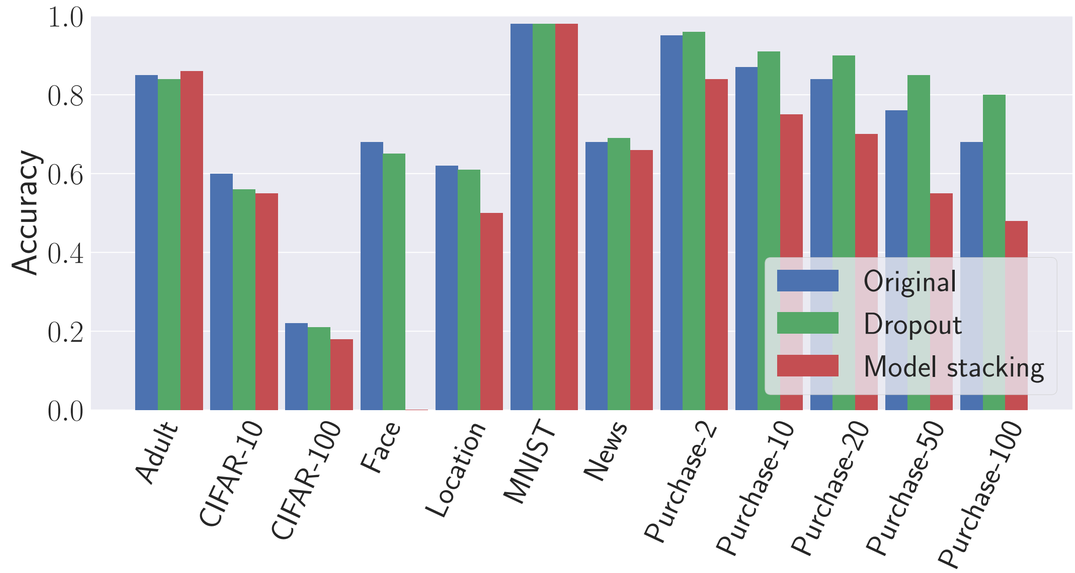
Srivastava, Nitish, et al. "Dropout: a simple way to prevent neural networks from overfitting", JMLR 2014

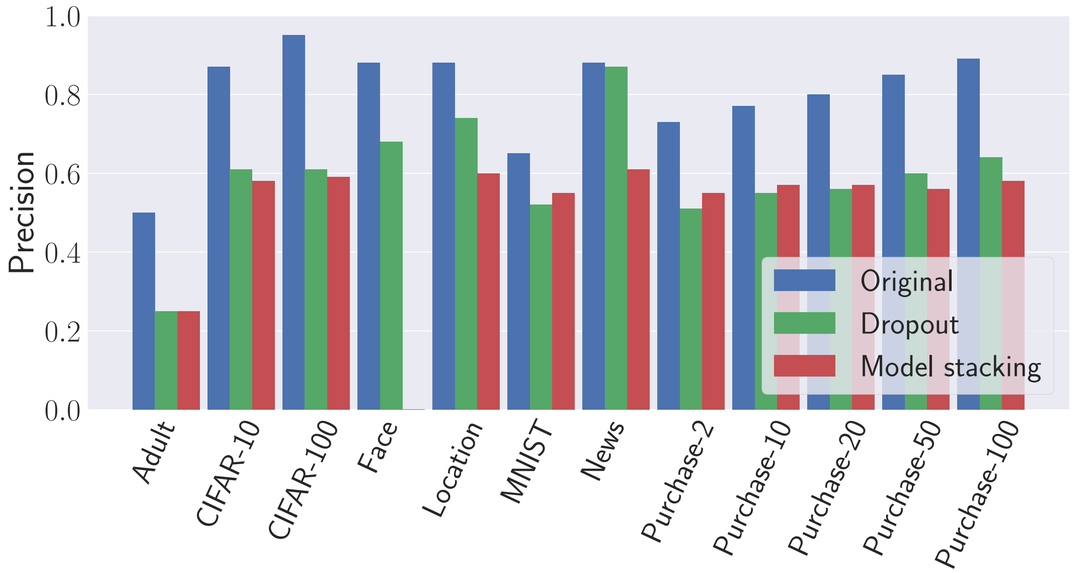
Defense

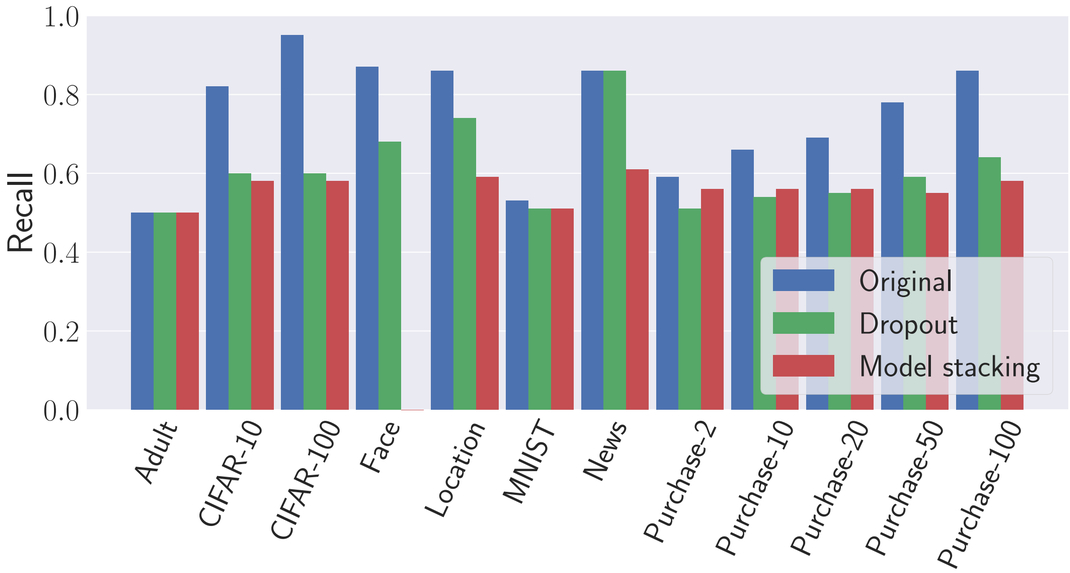
- Dropout
- Model Stacking



Performance

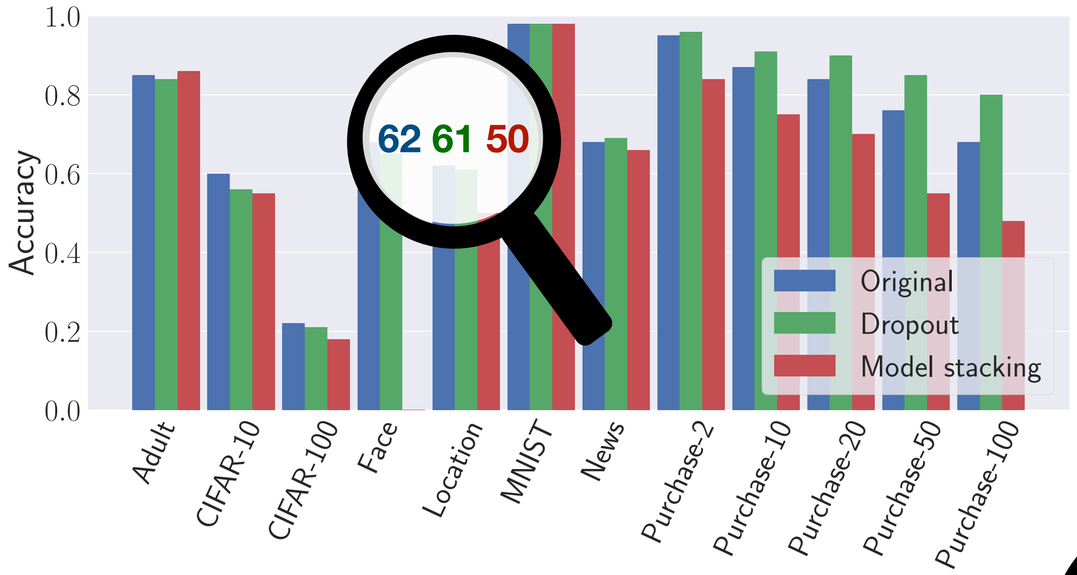


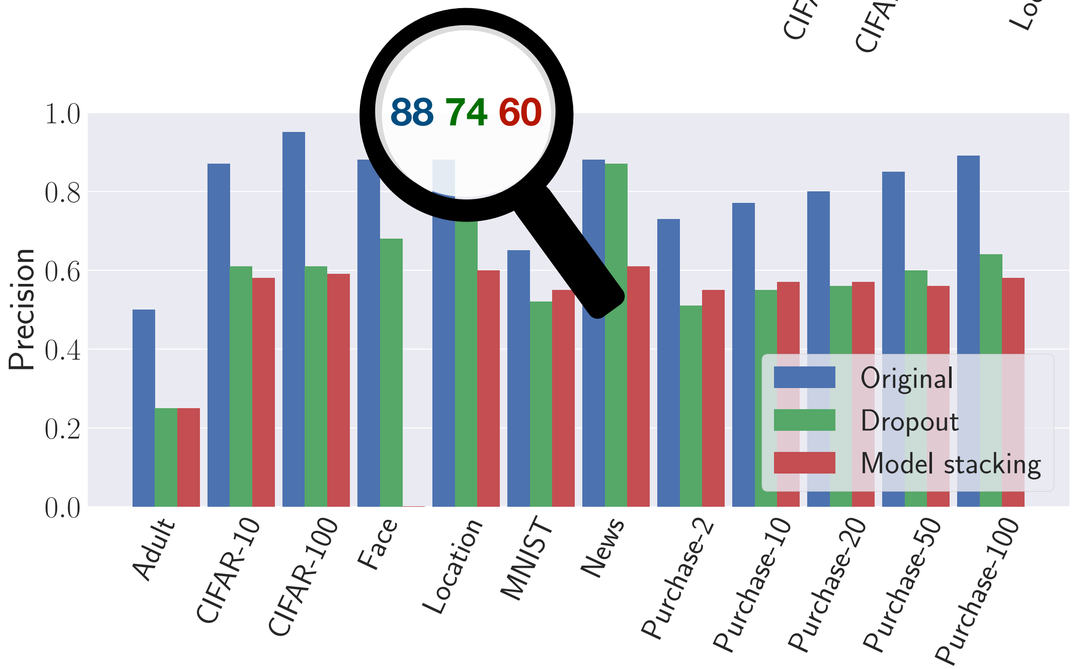


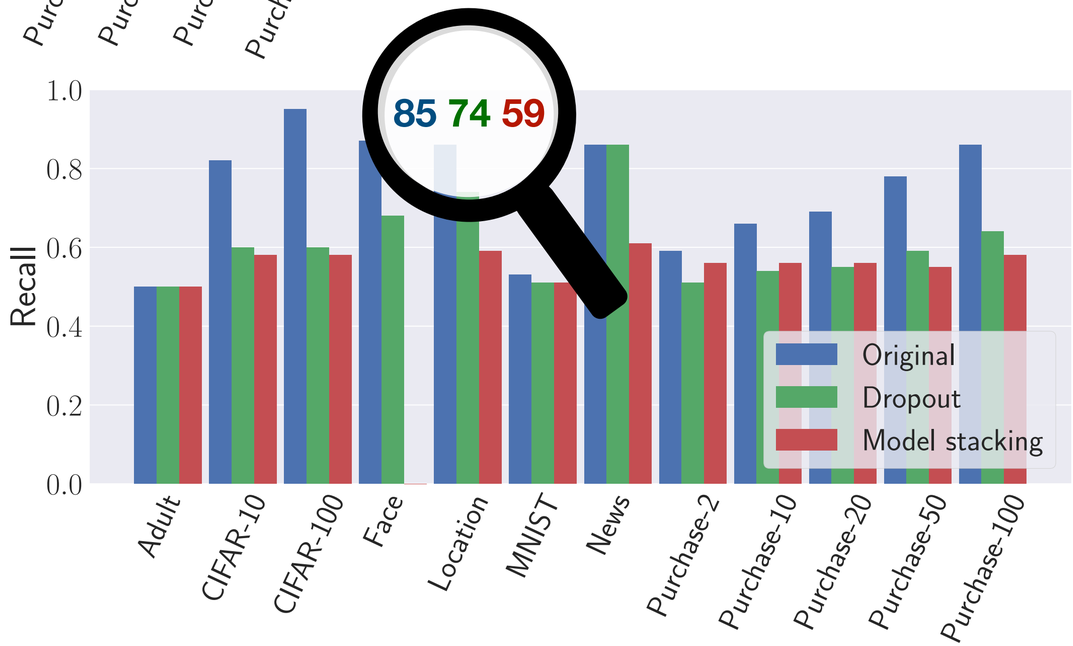


Performance

	Accuracy	Precision	Recall
Dropout	**2 %**	**16 %**	**13 %**
Model stacking	**19 %**	**32 %**	**31 %**







Conclusion

Conclusion

- Membership inference attack simpler

Conclusion

- Membership inference attack simpler
- Overfitting is a common enemy

Conclusion

CISPA
HELMHOLTZ CENTER FOR INFORMATION SECURITY

- Membership inference attack simpler
- Overfitting is a common enemy
- Defenses against membership inference

Conclusion

- Membership inference attack simpler
- Overfitting is a common enemy
- Defenses against membership inference

Thank you for your attention!
Questions?

ahmed.salem@cispa.saarland
https://ahmedsalem2.github.io/
@AhmedGaSalem