Privacy in the Modern Era: The Cases of Online Social Network and Machine Learning Model

Yang Zhang

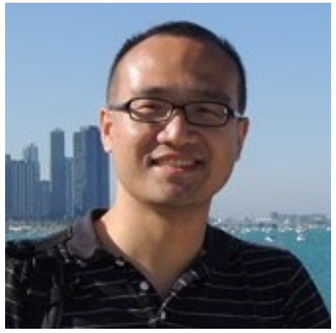
CISPA
HELMHOLTZ-ZENTRUM i. G.

Michael Backes

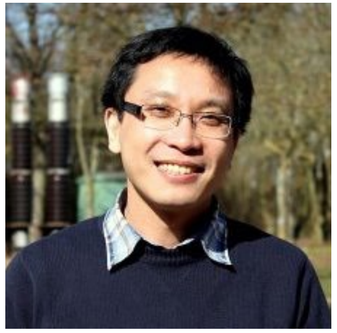
Pascal Berrang

Mario Fritz

Mathias Humbert



Cheng-Te Li



Jun Pang

Tahleen Rahman

Ahmed Salem

About Me

- Postdoc at CISPA Helmholtz Center i.G. working with Michael Backes
- From January 2019, research group leader at CISPA Helmholtz Center for Information Security
- Data privacy
  - Biomedical data, machine learning models, social network
- Ph.D. positions and summer interns available

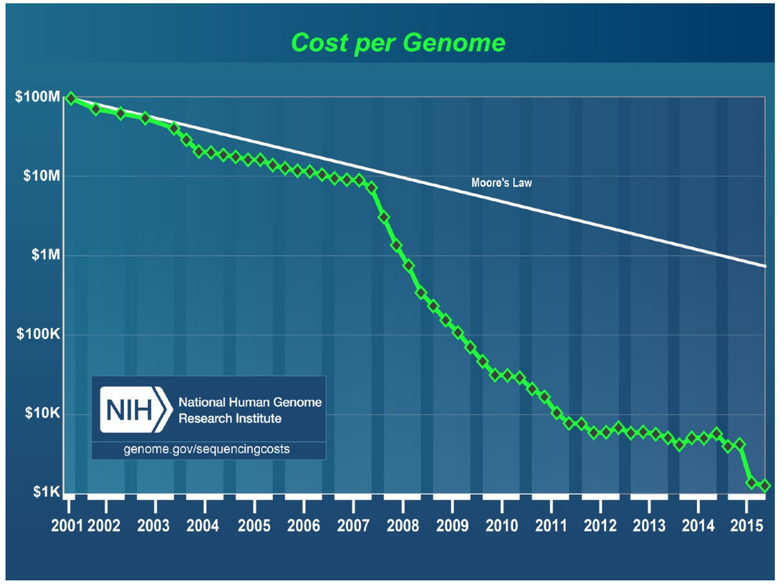
CISPA-Stanford Program

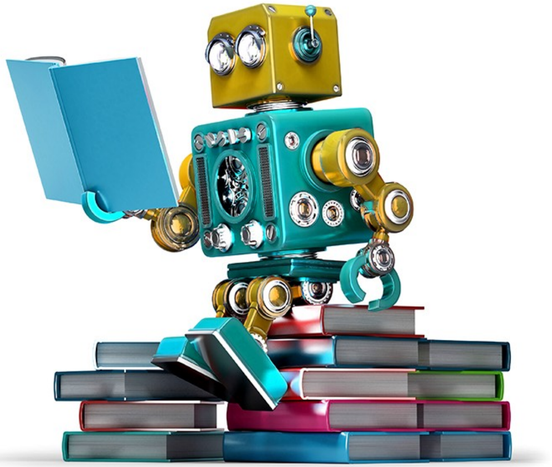
CISPA HELMHOLTZ-ZENTRUM i. G.

- https://www.cispa-stanford.org/
- Elite scholar program for doctors
  - 1 or 2 years at CISPA
  - 2 years at Stanford University as a visiting professor
  - 3 years at CISPA as a research group leader
  - Drop me an email if you are interested

The Advancement of ICT

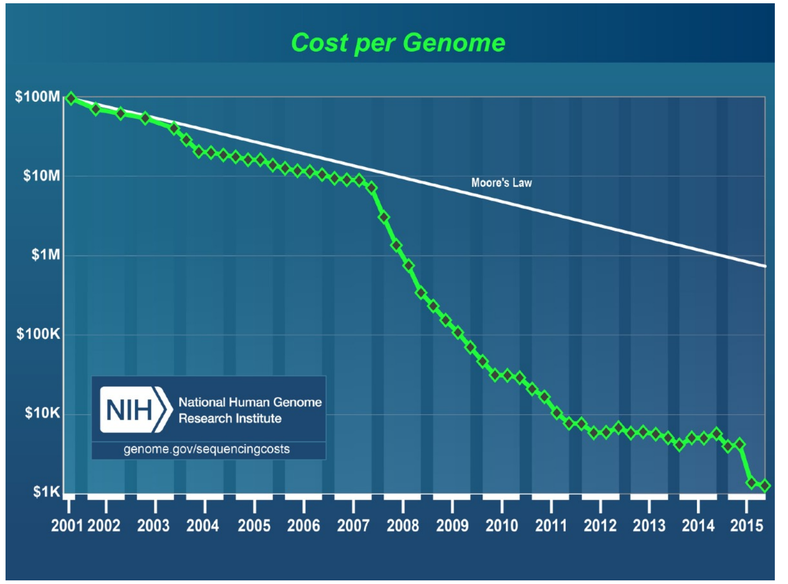
CISPA
HELMHOLTZ-ZENTRUM i. G.

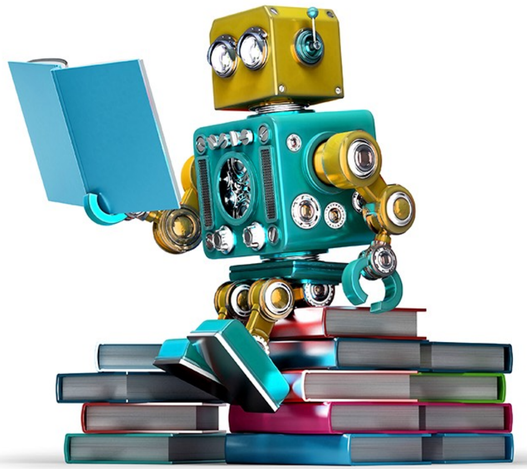




Privacy!!!

CISPA
HELMHOLTZ-ZENTRUM i. G.





Outline

CISPA
HELMHOLTZ-ZENTRUM i. G.

- Social network privacy
- Machine learning privacy

Outline

CISPA
HELMHOLTZ-ZENTRUM i. G.

- Social network privacy
- Machine learning privacy

Social Network Privacy

CISPA HELMHOLTZ-ZENTRUM i. G.

- In basic setting, users
  - Articulate their personal attributes and their social relations -> many attacks exist
- De-facto way for communication
  - Texts, images… -> some attacks exist
  - Cooler information
    - Location check-in, hashtags… -> privacy?

Social Network Privacy

CISPA HELMHOLTZ-ZENTRUM i. G.

- Location check-in to infer social relation
  - walk2friends: Inferring Social Links from Mobility Profiles (CCS 2017)
- Hashtag to infer location
  - Tagvisor: A Privacy Advisor for Sharing Hashtags (WWW 2018)

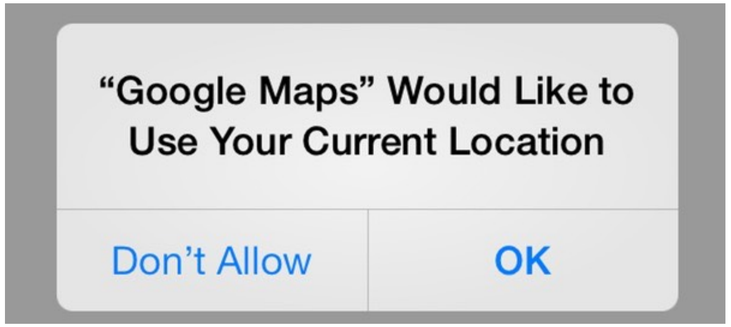
Social Network Privacy

CISPA HELMHOLTZ-ZENTRUM i. G.

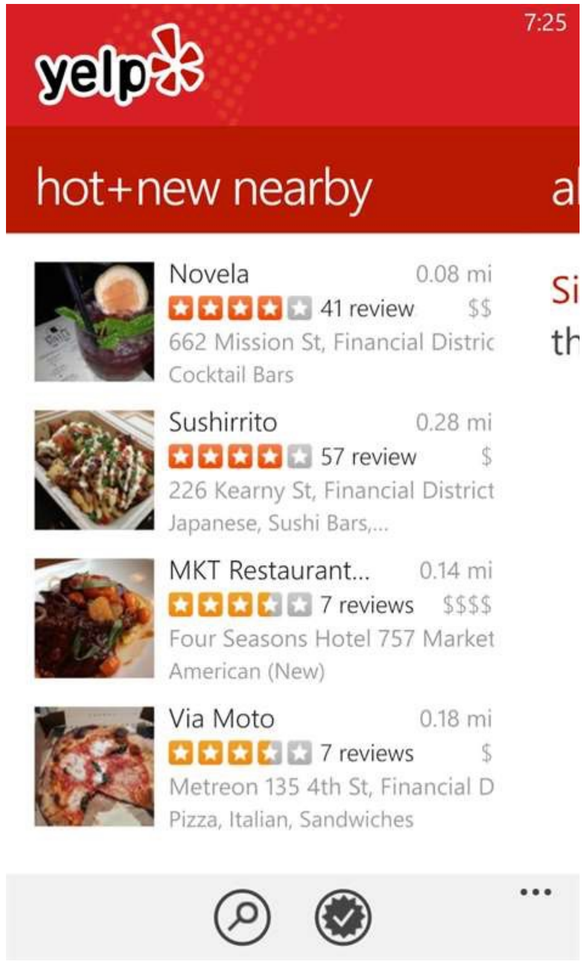
- Location check-in to infer social relation
  - walk2friends: Inferring Social Links from Mobility Profiles (CCS 2017)
- Hashtag to infer location
  - Tagvisor: A Privacy Advisor for Sharing Hashtags (WWW 2018)

Location Check-in

CISPA
HELMHOLTZ-ZENTRUM i. G.



Location Check-in



Location Privacy

CISPA HELMHOLTZ-ZENTRUM i. G.

- 4 spatial-temporal points can identify 95% of the individuals
- Mobility traces can be effectively de-anonymized
- You are where you go
  - Demographics
  - Social relations

Location Privacy

CISPA HELMHOLTZ-ZENTRUM i. G.

- 4 spatial-temporal points can identify 95% of the individuals
- Mobility traces can be effectively de-anonymized
- You are where you go
  - Demographics
  - Social relations

Social Relation is Private?

CISPA HELMHOLTZ-ZENTRUM i. G.

- Social relations can be sensitive, e.g., office romance
- 17.2% -> 56.2% (Facebook users in New York)
- NSA’s co-traveler program

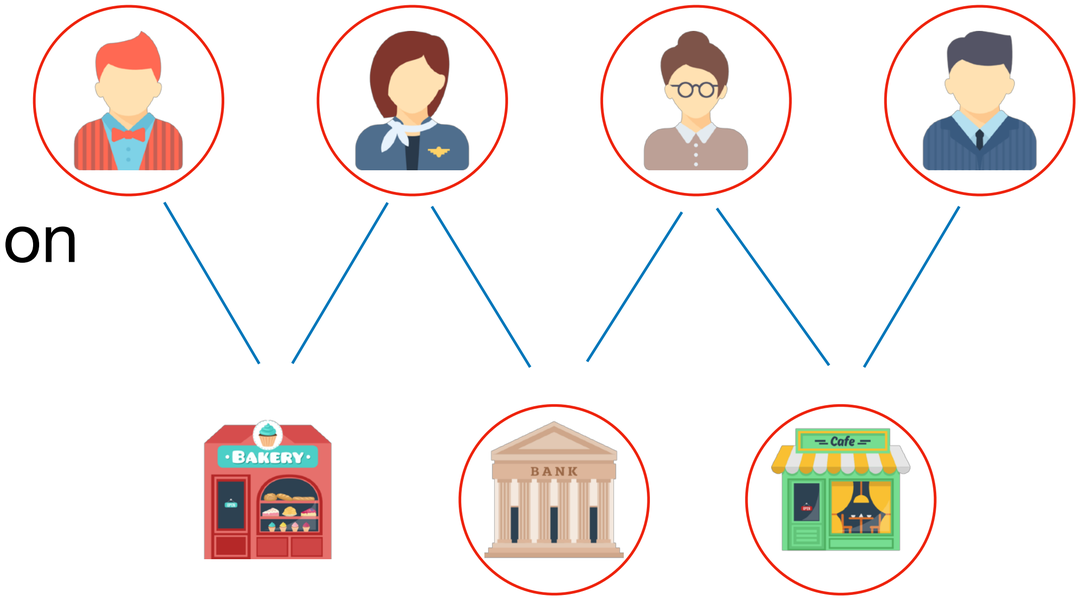
Research Question

CISPA
HELMHOLTZ-ZENTRUM i. G.

Can two users' check-ins be used to predict their social relations?

Existing Approach

- Solution 1: common locations two users have visited
- Almost all data mining approaches take this way
- Location entropy
- Can’t work when two users share no common locations



Existing Approach

CISPA HELMHOLTZ-ZENTRUM i. G.

- Solution 2: mobility profiles/features
- Summarize each user's mobility profiles
  - Friends share similar mobility profiles than strangers
- Feature engineering
  - Tedious efforts and domain expert knowledge
  - Time consuming

Every Single Time!!!

Representation Learning

CISPA HELMHOLTZ-ZENTRUM i. G.

- Learning features (representation/deep learning)
  - Follow a general object (unsupervised)
- Graph representation learning (graph embedding)
  - Preserve each user’s neighbors in a social network
- Mobility feature learning

walk2friends

CISPA HELMHOLTZ-ZENTRUM i. G.

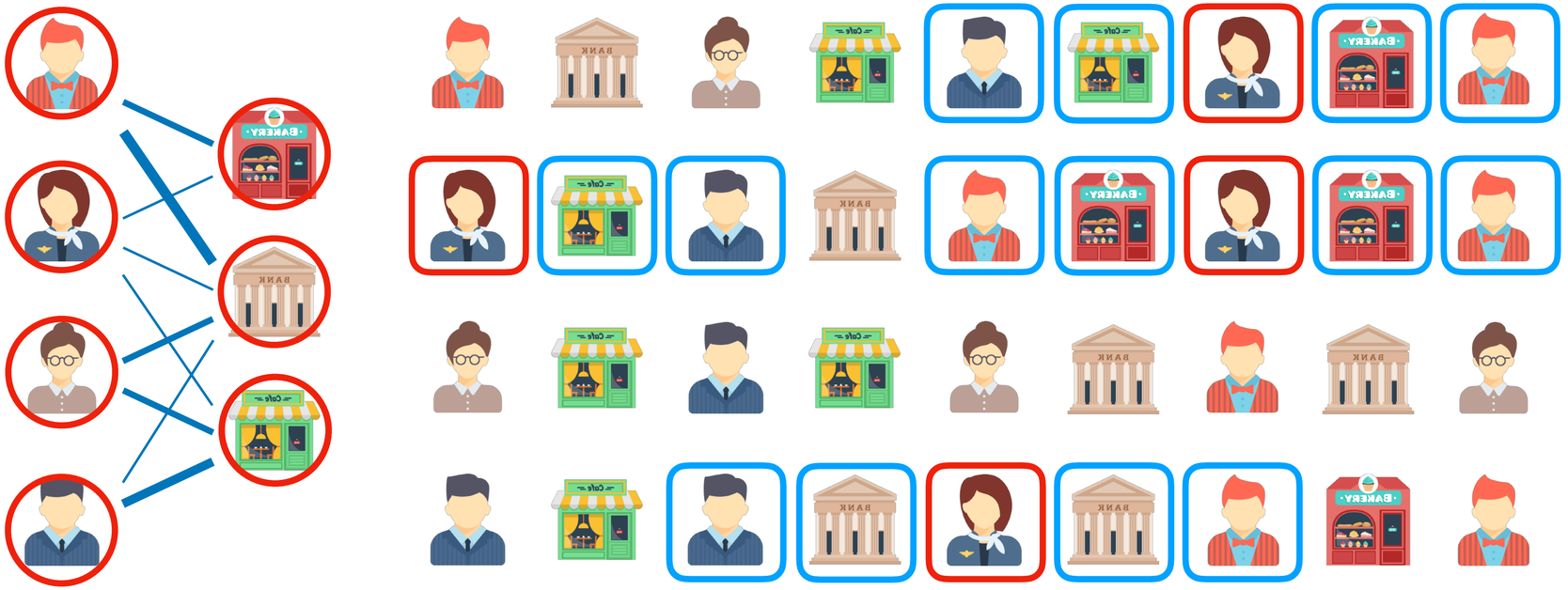
- Assumption: A user's mobility neighbors can reflect her mobility profile/features

1. Define each user's mobility neighbors
2. Learn mobility features/profiles
3. Infer two users' social relation

Mobility Neighbor

- A user’s mobility neighbors include
  - Locations a user has visited
  - Others who have visited similar locations and their locations
- Breadth first search
  - Not considering the visiting frequencies
- Random walk sampling

Mobility Neighbor

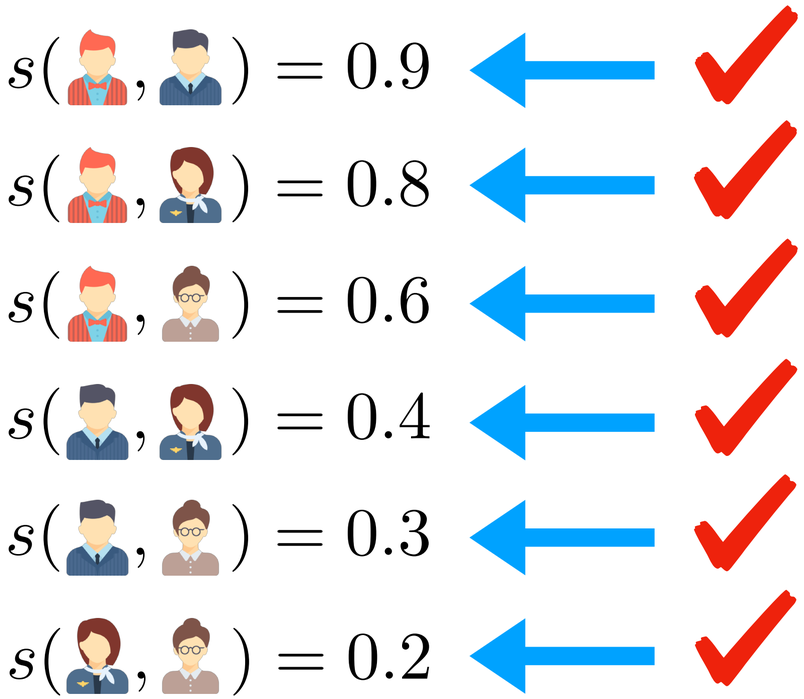


Mobility Feature Learning

- Learn a function:
- Each node to predict it's neighbors
- $p(\cdot \mid \cdot\,; \theta)$ Softmax

$$\arg\max_{\theta} p(\text{🏛}|\text{👩‍💼};\theta)\cdot p(\text{👩}|\text{👩‍💼};\theta)\cdot p(\text{👨}|\text{👩‍💼};\theta)\cdot$$
$$p(\text{👩‍💼}|\text{👨‍💼};\theta)\cdot p(\text{🏛}|\text{👨‍💼};\theta)\cdot p(\text{🏪}|\text{👨‍💼};\theta)\cdot p(\text{👩}|\text{👨‍💼};\theta)\cdot$$
$$p(\text{👩‍💼}|\text{👩};\theta)\cdot p(\text{🏛}|\text{👩};\theta)\cdot p(\text{🏪}|\text{👩};\theta)\cdot p(\text{👩}|\text{👩};\theta)\cdot$$
$$p(\text{🏪}|\text{👩‍🦰};\theta)\cdot p(\text{👨}|\text{👩‍🦰};\theta)\cdot$$
$$p(\text{👩‍💼}|\text{🏛};\theta)\cdot p(\text{👩}|\text{🏛};\theta)\cdot p(\text{🏪}|\text{🏛};\theta)\cdot p(\text{👨}|\text{🏛};\theta)\cdot$$
$$p(\text{🏛}|\text{🏪};\theta)\cdot p(\text{👩}|\text{🏪};\theta)\cdot p(\text{👨}|\text{🏪};\theta)\cdot p(\text{👩‍🦰}|\text{🏪};\theta)$$

Social Relation Inference



- Cosine similarity
- Unsupervised
- Predict any social relation

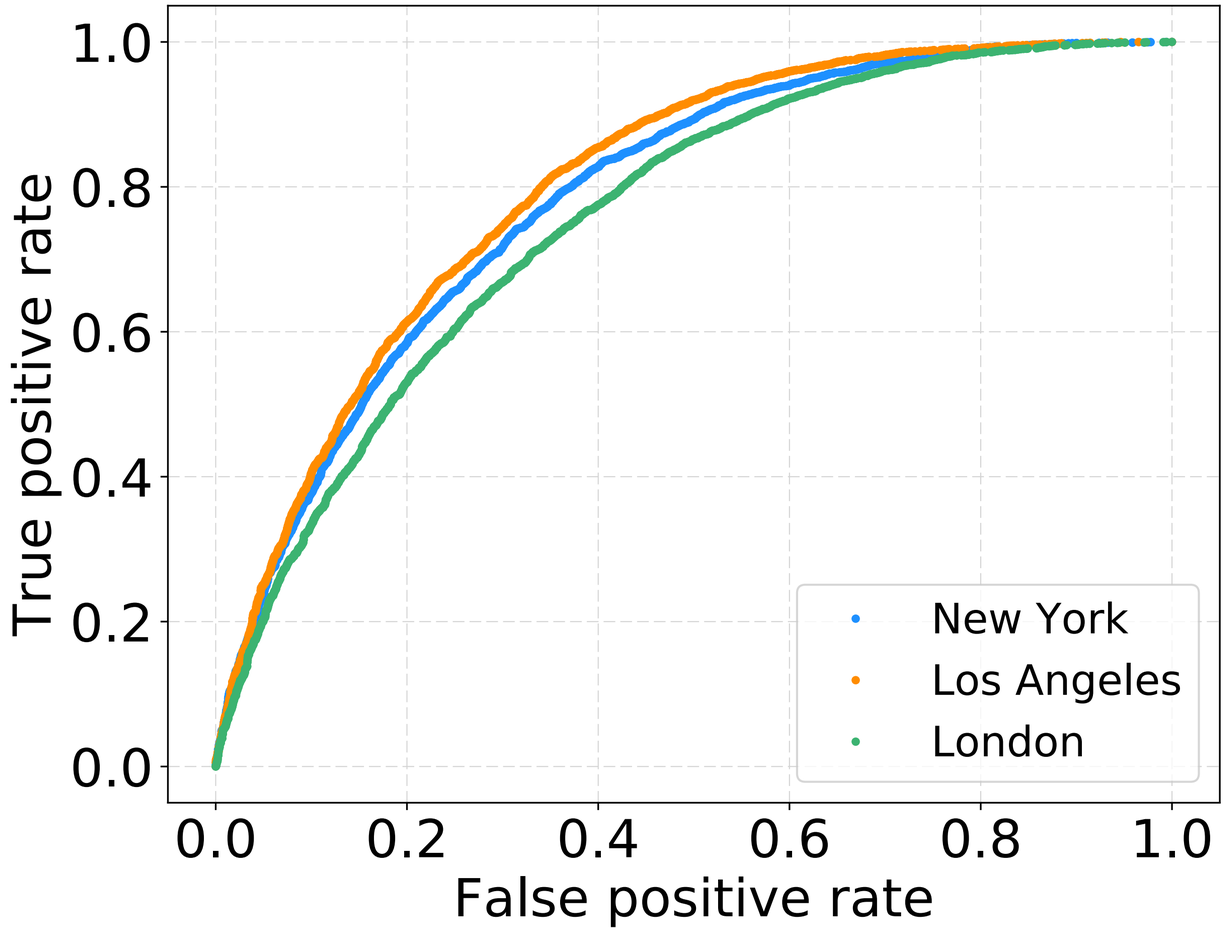
Dataset

CISPA HELMHOLTZ-ZENTRUM i. G.

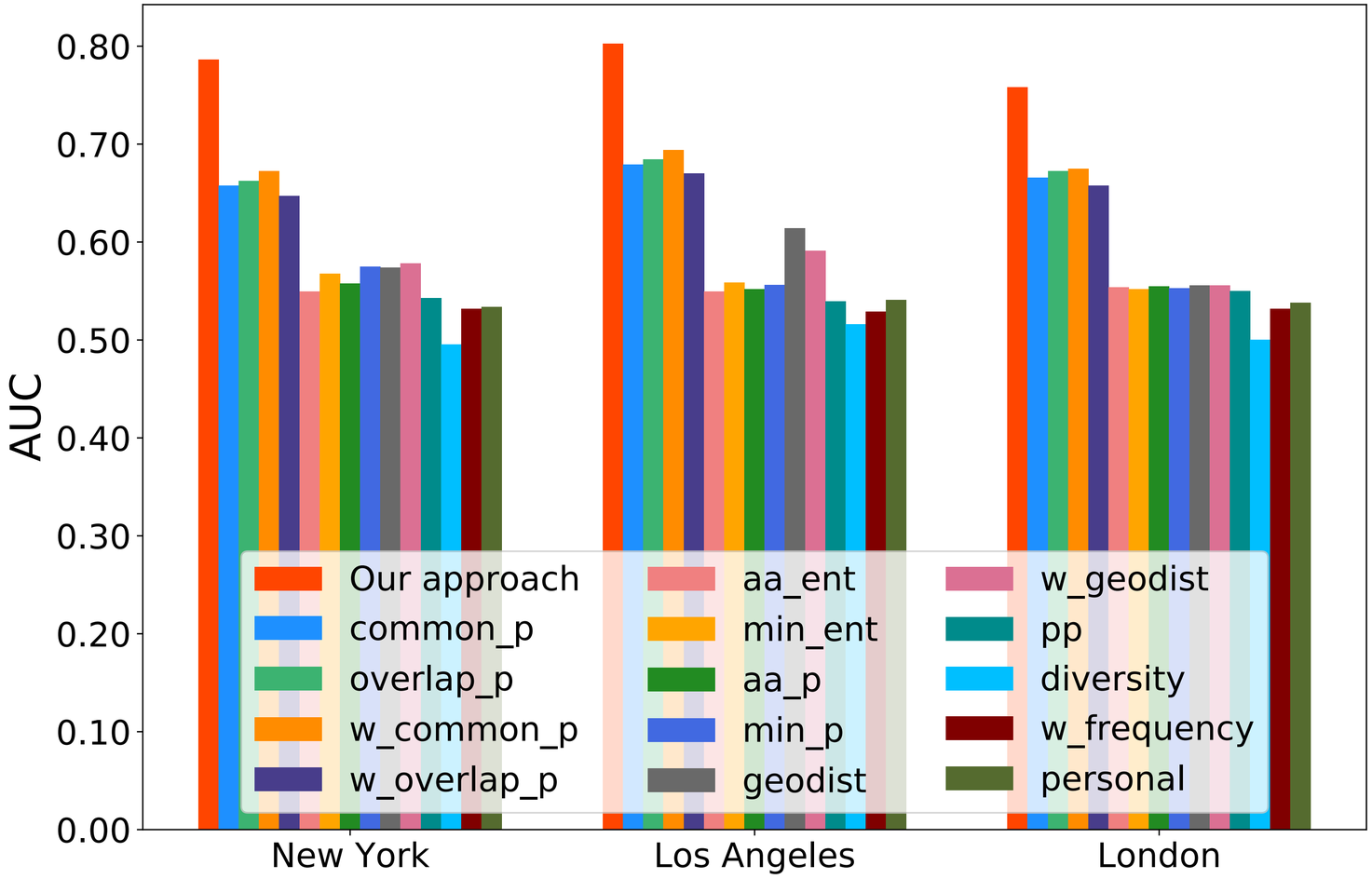
- Instagram users' check-ins
  - New York, Los Angeles and London
  - Foursquare (location semantics)
- Social relations (two users follow each other)
- Dataset available!

	New York	Los Angeles	London
No. check-ins	1,843,187	1,301,991	500,776
No. locations	25,868	22,260	10,693
No. users	44,371	30,679	13,187
No. social links	193,995	129,004	25,413

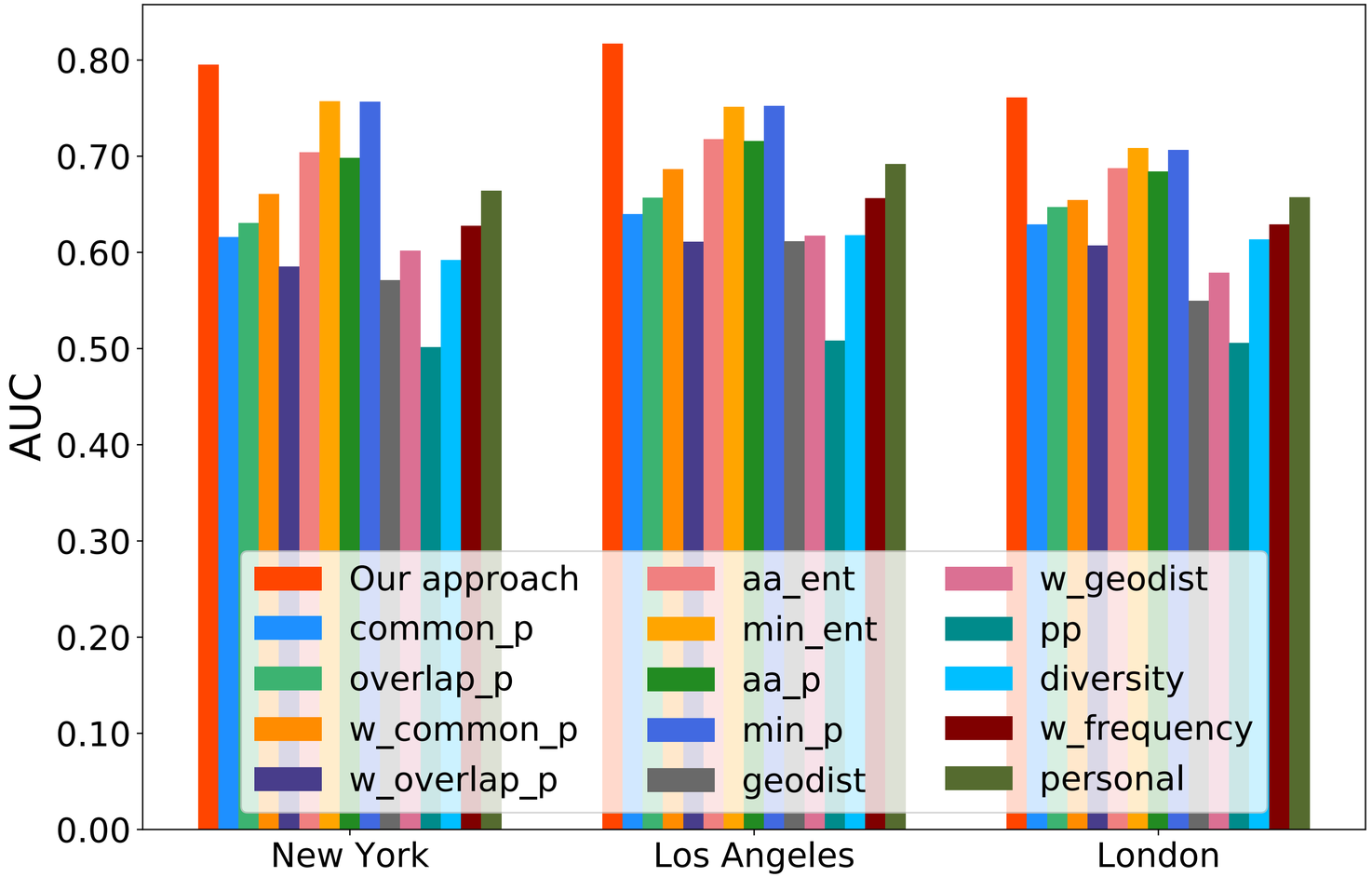
Evaluation


True positive rate
False positive rate
0.0
0.2
0.4
0.6
0.8
1.0
New York
Los Angeles
London

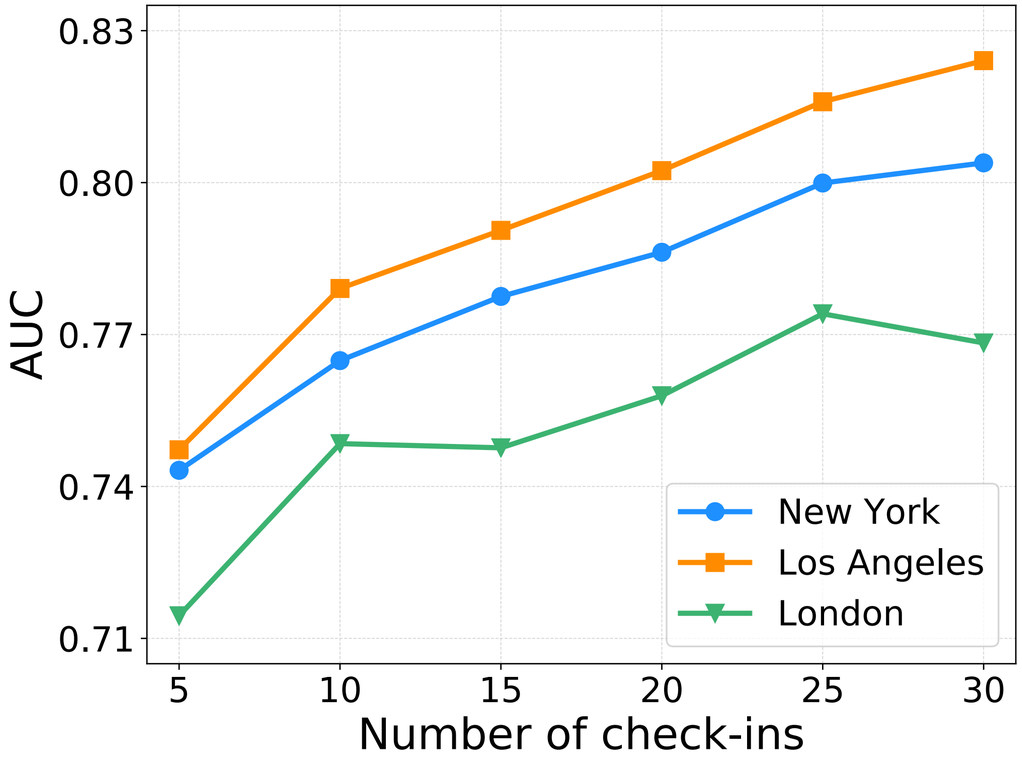
Evaluation

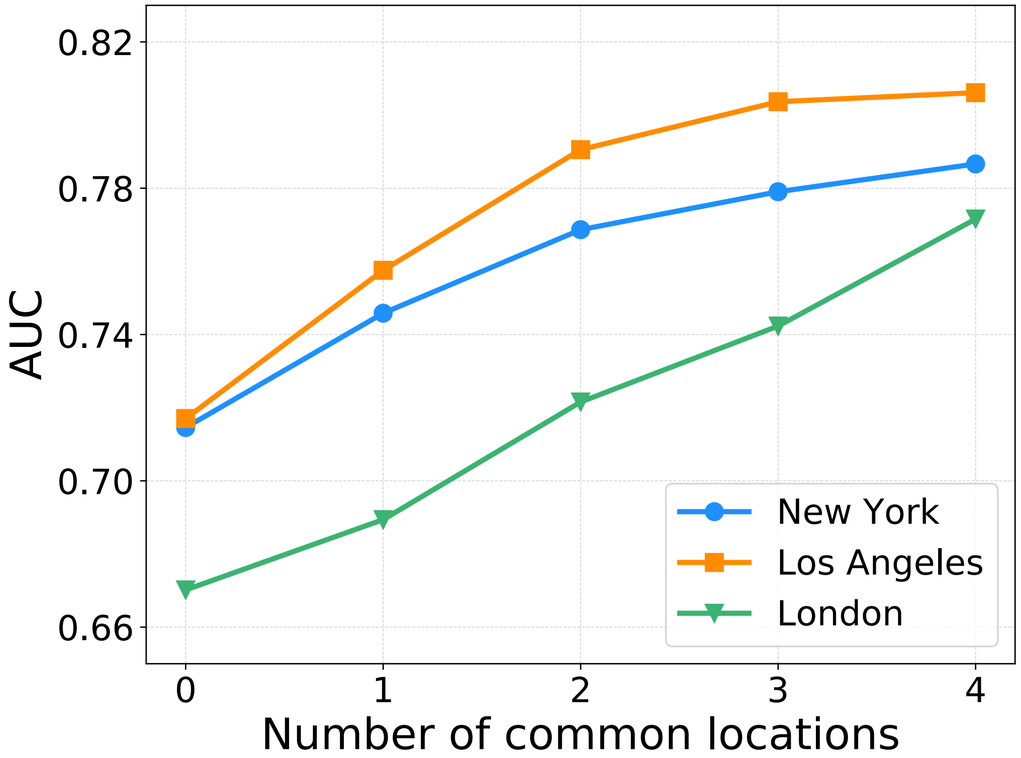

AUC
0.80
0.70
0.60
0.50
0.40
0.30
0.20
0.10
0.00
New York
Los Angeles
London
Our approach
common_p
overlap_p
w_common_p
w_overlap_p
aa_ent
min_ent
aa_p
min_p
geodist
w_geodist
pp
diversity
w_frequency
personal

Evaluation

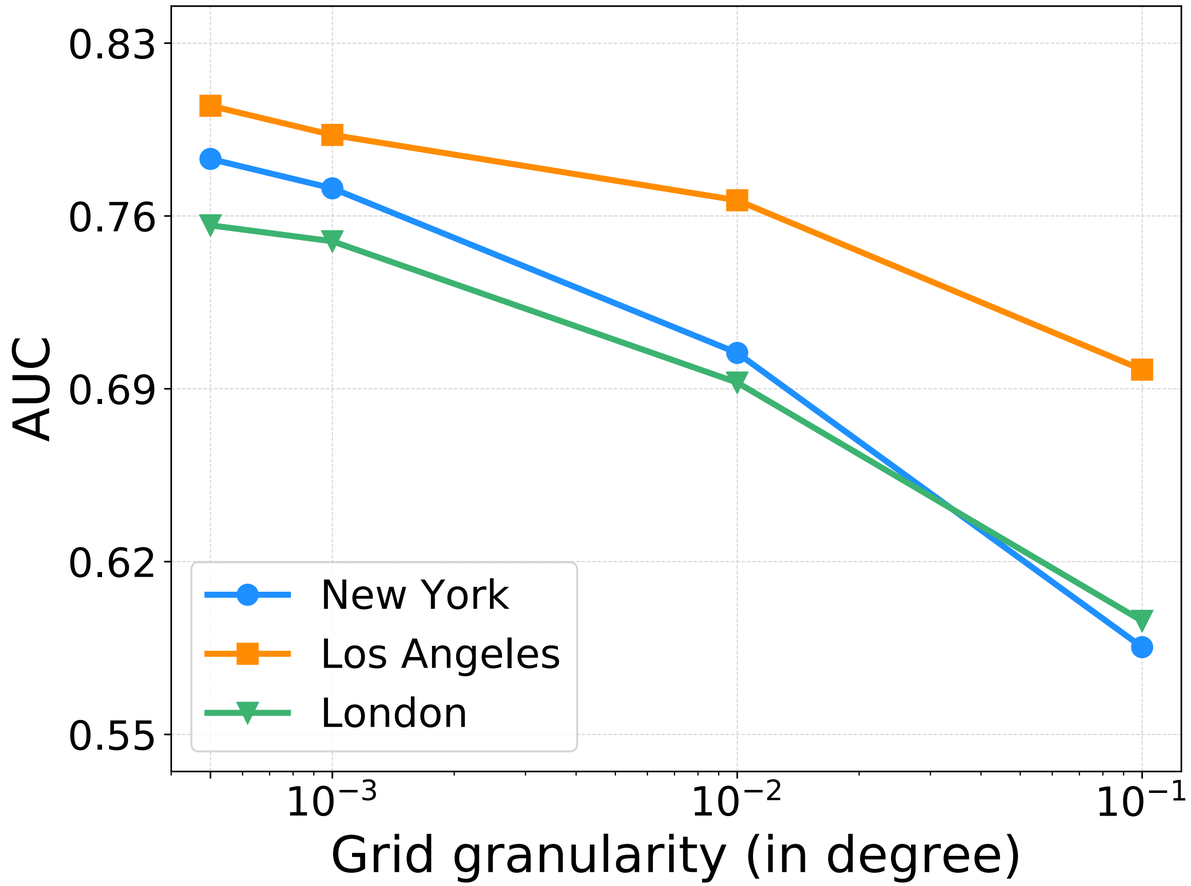

AUC
0.00
0.10
0.20
0.30
0.40
0.50
0.60
0.70
0.80
New York
Los Angeles
London
Our approach
common_p
overlap_p
w_common_p
w_overlap_p
aa_ent
min_ent
aa_p
min_p
geodist
w_geodist
pp
diversity
w_frequency
personal

Evaluation


0.83
0.80
0.77
0.74
0.71
AUC
5
10
15
20
25
30
Number of check-ins
New York
Los Angeles
London


0.82
0.78
0.74
0.70
0.66
AUC
0
1
2
3
4
Number of common locations
New York
Los Angeles
London

Evaluation


0.83
0.76
0.69
0.62
0.55
AUC
New York
Los Angeles
London
10^{-3}
10^{-2}
10^{-1}
Grid granularity (in degree)

Defense

- Hiding
  - Delete certain proportion of check-ins
- Replacement
  - Random walk to replace locations

Defense

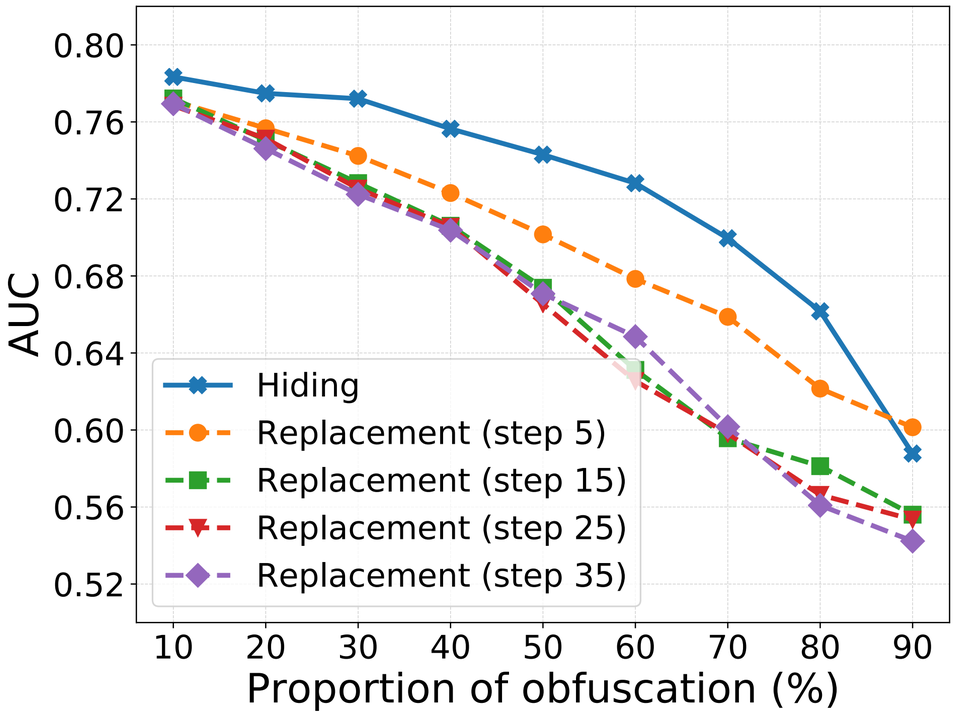
CISPA HELMHOLTZ-ZENTRUM i. G.

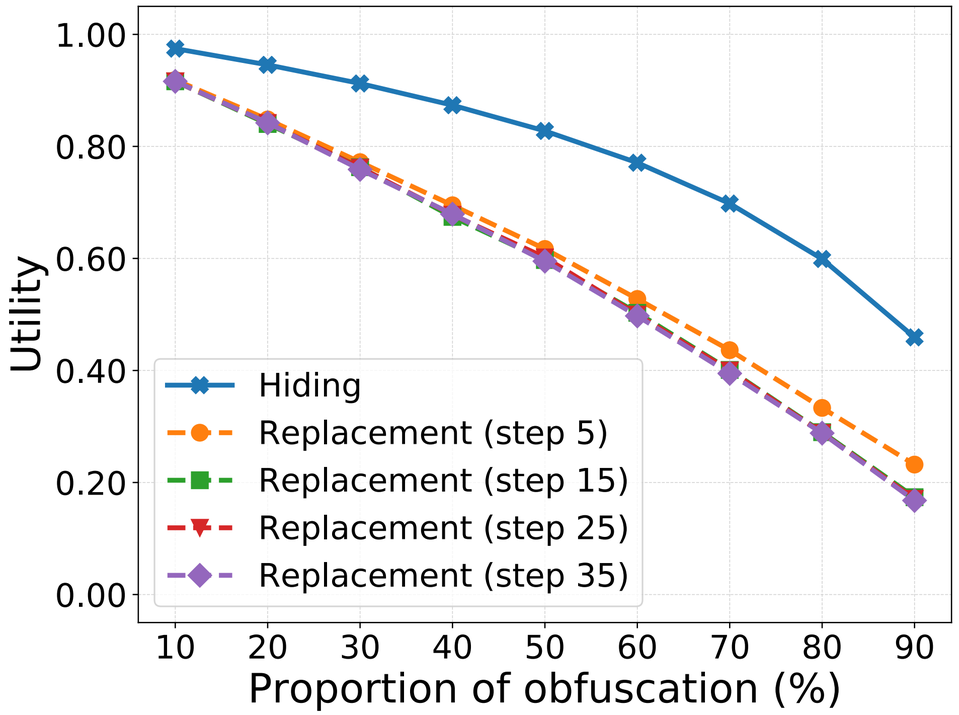
- Generalization
  - Geo-coordinate and location semantics
  - MoMA -> art (40.76N, -73.97W)
- Recover location first
  - art (40.76N, -73.97W) -> MoMA or Tom Otterness Frog?

Utility Metric

- Each user’s check-in distribution
  - Both original and obfuscated
- Jensen-Shannon divergence
- Average over all users

Defense


AUC
0.80
0.76
0.72
0.68
0.64
0.60
0.56
0.52
10
20
30
40
50
60
70
80
90
Proportion of obfuscation (%)
Hiding
Replacement (step 5)
Replacement (step 15)
Replacement (step 25)
Replacement (step 35)

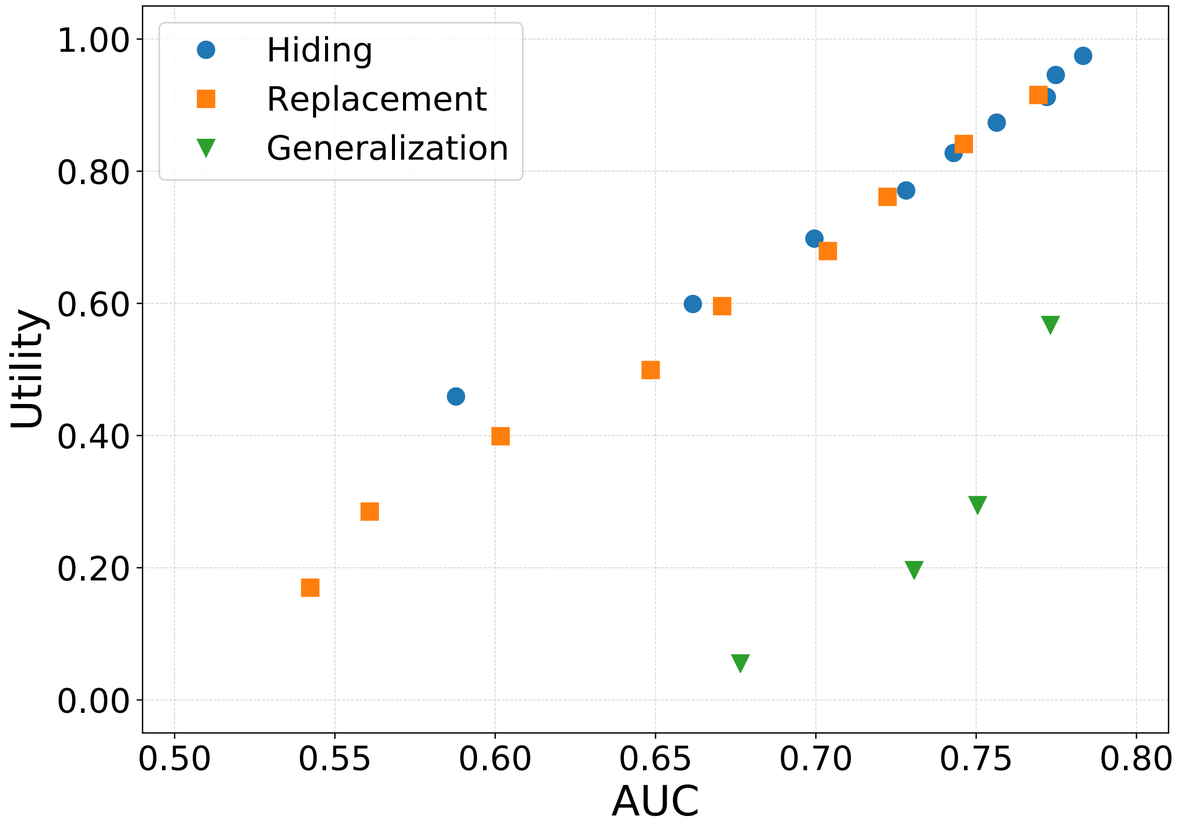

Utility
1.00
0.80
0.60
0.40
0.20
0.00
10
20
30
40
50
60
70
80
90
Proportion of obfuscation (%)
Hiding
Replacement (step 5)
Replacement (step 15)
Replacement (step 25)
Replacement (step 35)

Defense

CISPA HELMHOLTZ-ZENTRUM i. G.

	AUC		Utility		Recovery rate	
	ls	*hs*	*ls*	*hs*	*ls*	*hs*
lg	0.77	0.75	0.57	0.30	52%	23%
hg	0.73	0.67	0.20	0.06	14%	2%

Defense


Hiding
Replacement
Generalization
Utility
AUC
1.00
0.80
0.60
0.40
0.20
0.00
0.50
0.55
0.60
0.65
0.70
0.75
0.80

Summary on Walk2friends

CISPA HELMHOLTZ-ZENTRUM i. G.

- A novel social relation inference attack with mobility profiles
  - Unsupervised and predict any social relations
  - Outperforms baseline models
- Three general defense mechanisms
  - Replacement and hiding outperform generalization

Social Network Privacy

CISPA HELMHOLTZ-ZENTRUM i. G.

- Location check-in to infer social relation
  - walk2friends: Inferring Social Links from Mobility Profiles (CCS 2017)
- Hashtag to infer location
  - Tagvisor: A Privacy Advisor for Sharing Hashtags (WWW 2018)

#hashtag

CISPA
HELMHOLTZ-ZENTRUM i. G.

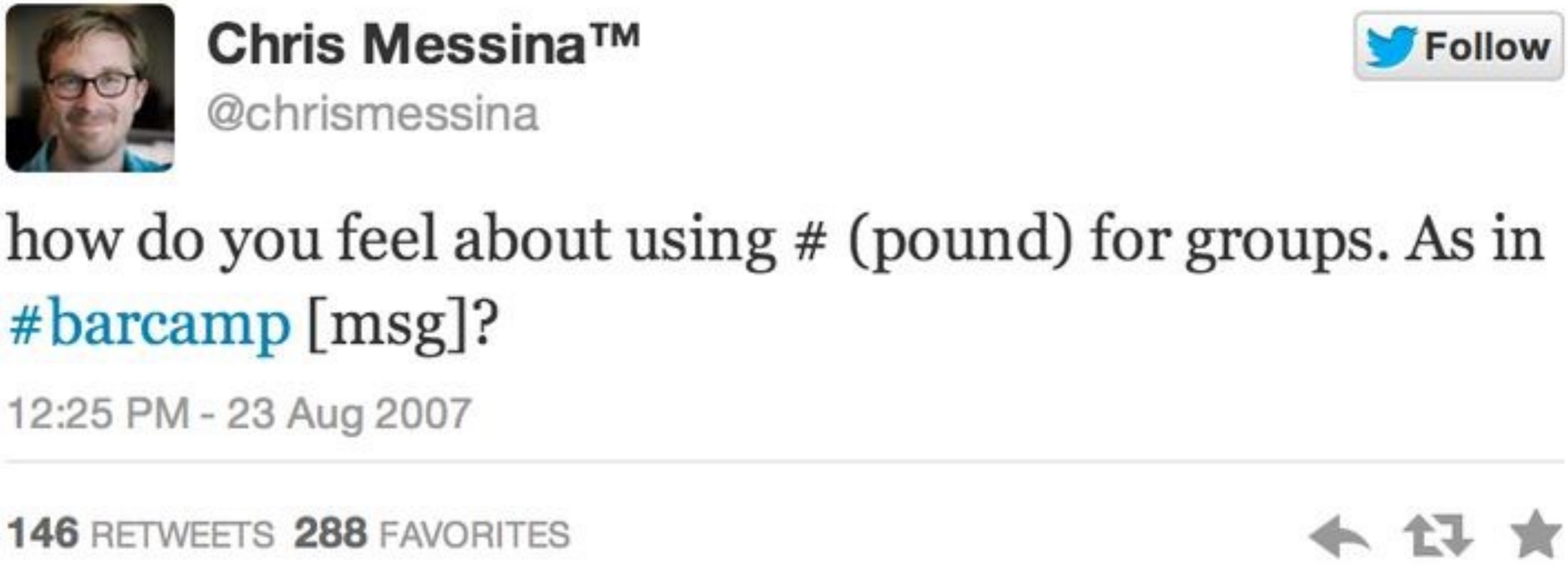
Follow

SHARE

#

#hashtag

CISPA
HELMHOLTZ-ZENTRUM i. G.



#hashtag

CISPA
HELMHOLTZ-ZENTRUM i. G.

#hashtag

#hashtag

#like4like

#foodporn

#tbt

#hashtag

CISPA
HELMHOLTZ-ZENTRUM i. G.

#privacy

#locationprivacy

Research Question

CISPA
HELMHOLTZ-ZENTRUM i. G.

Can hashtags a user shares be used to infer her location?

Tagvisor

CISPA HELMHOLTZ-ZENTRUM i. G.

- Attack: location inference with hashtags
- Defense: Tagvisor, a privacy advisor to mitigate the privacy threat by hashtags

#dataset

CISPA HELMHOLTZ-ZENTRUM i. G.

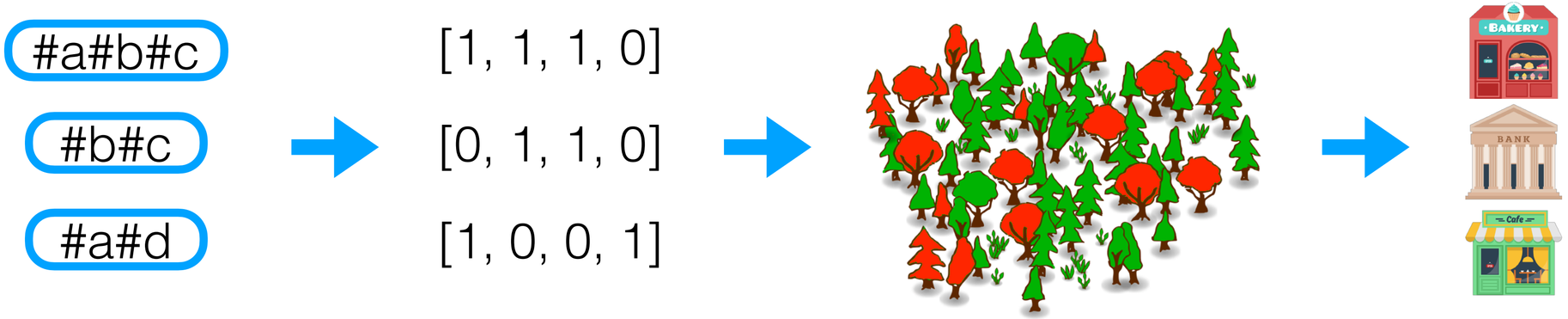
- Collected through Instagram's APIs
- New York, Los Angeles, and London
- Hashtags + locations (check-ins)

	New York	Los Angeles	London
No. of posts	144,263	61,767	34,018
No. of hashtags	8,552	4,600	2,395
No. of users	3,911	1,625	992
No. of locations	498	268	141

17 likes

[redacted] #sunday #sun #reading #rva #tan #light #relax #girl #me #outside #spring #warm #instagood #photooftheday #iphonesia #instamood #igers #instagramhub #picoftheday #instadaily #bestoftheday #igdaily #instagramers #webstagram #all_shots #statigram #popular #photography #art #iphoneography

#attack



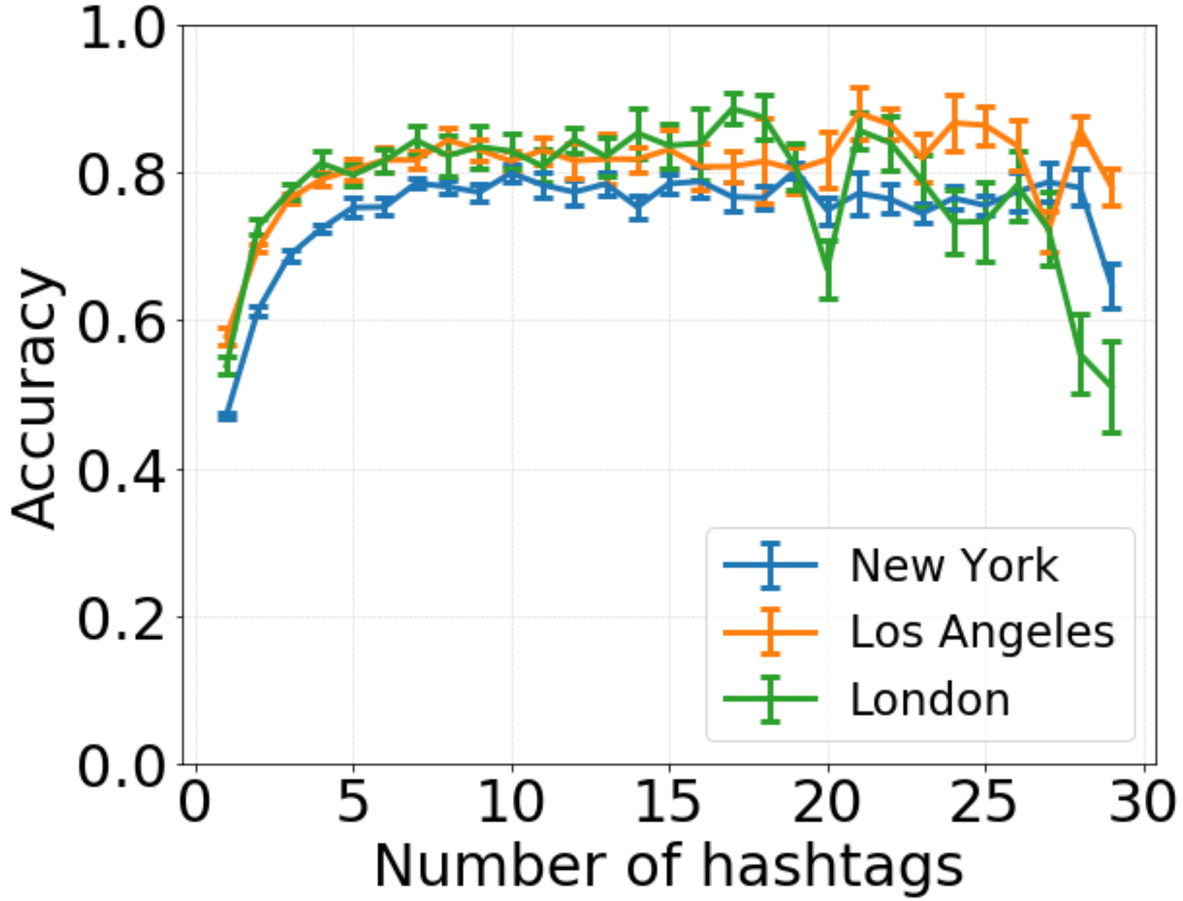
- Bag-of-words for feature representation
- Random forest classifier
- Multiple-class classification, e.g., 498 classes (locations) in New York
- All posts are trained together

#attack

CISPA HELMHOLTZ-ZENTRUM i. G.

	New York		Los Angeles		London		All cities	
	attack	baseline	attack	baseline	attack	baseline	attack	baseline
Correctness	0.613	0.015	0.685	0.015	0.686	0.020	0.624	0.010
Distance (km)	0.917	4.198	1.870	11.275	0.857	4.518	211.471	3563.082
Accuracy	0.697	0.053	0.758	0.048	0.761	0.051	0.712	0.045

#attack


1.0
0.8
0.6
0.4
0.2
0.0
Accuracy
0
5
10
15
20
25
30
Number of hashtags
New York
Los Angeles
London

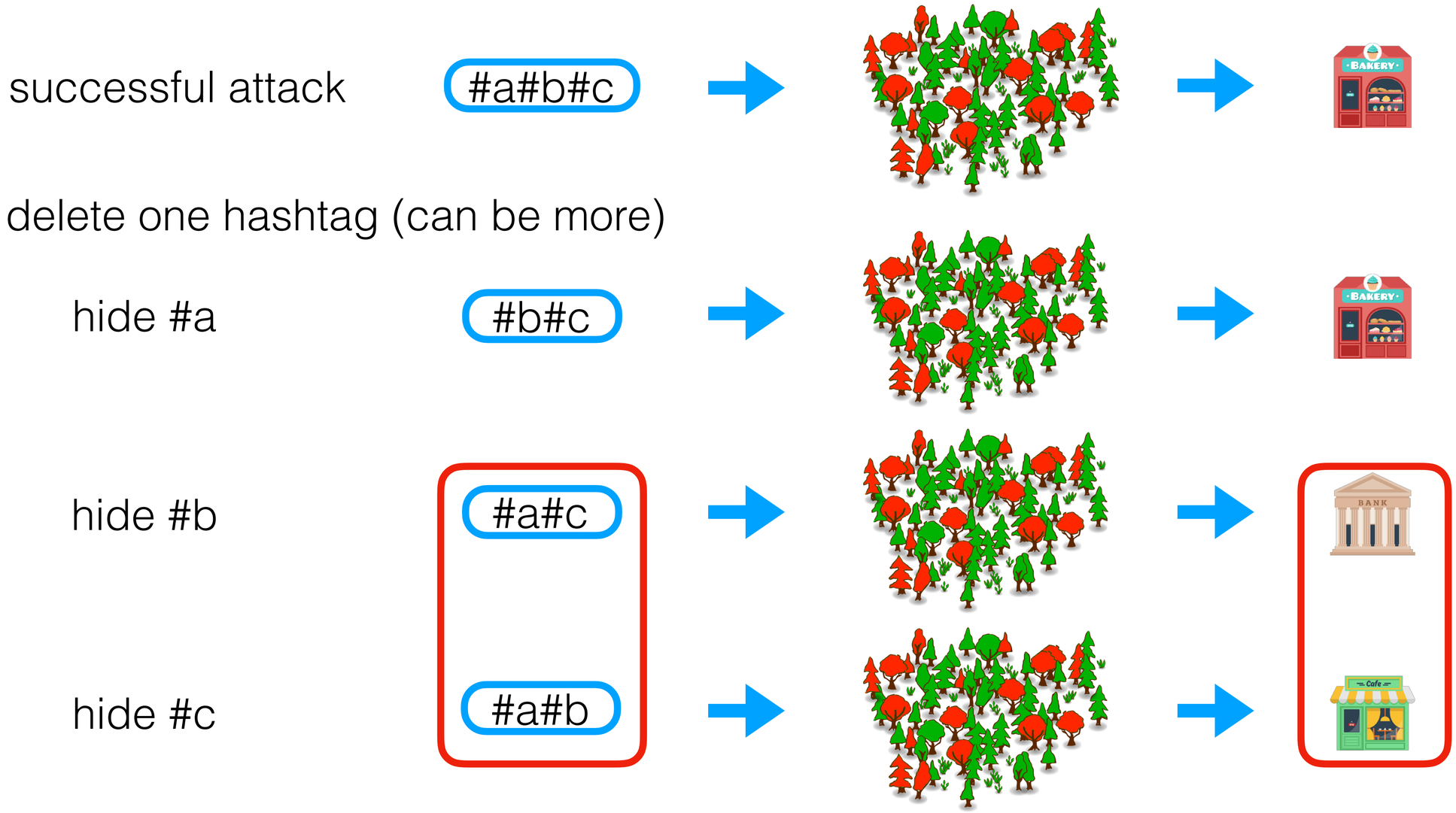
#tagvisor

CISPA HELMHOLTZ-ZENTRUM i. G.

- A privacy advisor for sharing hashtags
- Fool the attacker’s location inferencer (ML classifier)
- Three defense mechanisms
  - Hiding
  - Replacement
  - Generalization (location category)
- Utility: preserving the semantical meaning of hashtags

#hiding

CISPA HELMHOLTZ-ZENTRUM i. G.



#utility

#a#b ?

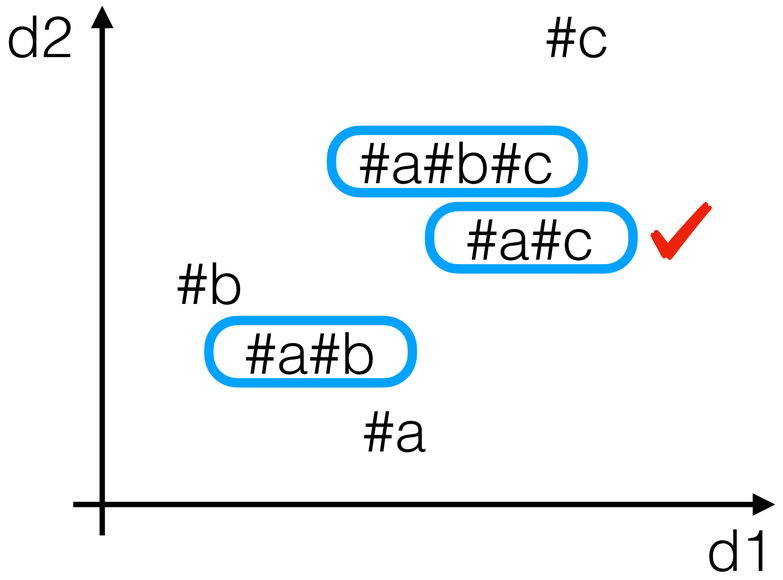
#a#b#c

#a#c ?

- Semantical meaning
- Skip-gram, aka word2vec
- Skip-gram over all posts' hashtags

Hashtag vectors

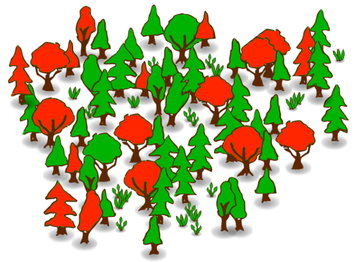
	d1	d2
#a:	3.1,	1.3]
#b:	2.5,	1.9]
#c:	4.0,	5.1]



#replacement

CISPA HELMHOLTZ-ZENTRUM i. G.

successful attack #a#b#c



- Replace each hashtag with all the possible hashtag
  - Search space is very large
- Bound to the most closest hashtags (with word2vec)
  - Reduce the search space
  - Semantical meaning can be preserved

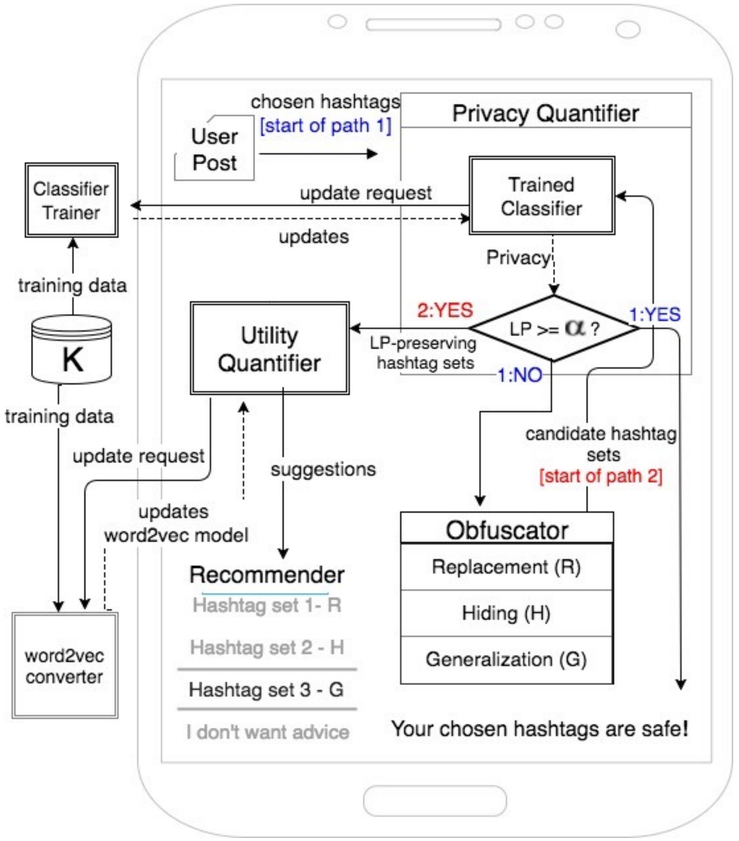
#generalization

- Location category from foursquare
  - #centralpark -> #park
- Do not apply to all hashtags
  - e.g., #tbt #love

#tagvisor

CISPA HELMHOLTZ-ZENTRUM i. G.

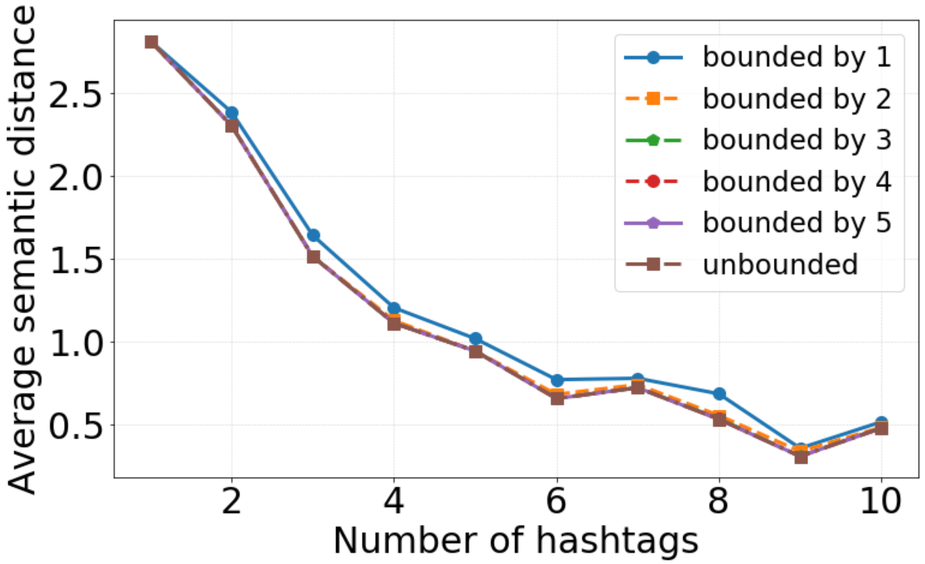
- Check whether the post's location is inferred correctly
  - If no, then publish
  - Else, consider the three defense mechanisms
    - Pick the hashtag set with the highest utility

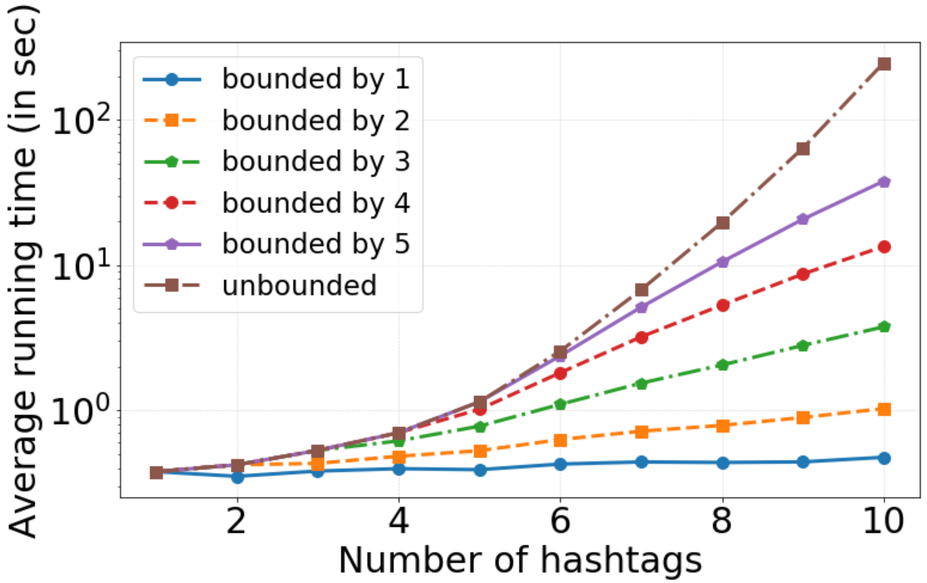


#tagvisor

CISPA HELMHOLTZ-ZENTRUM i. G.

Obfuscating bounded number of hashtags





Obfuscating 2 hashtags is enough!

Summary on Tagvisor

CISPA HELMHOLTZ-ZENTRUM i. G.

- First location inference attack with hashtags
  - Sharing hashtags is not safe!!!
- A privacy advisor to mitigate this risk
  - Minimal risk and maximal utility
  - Fit for the real-world setting

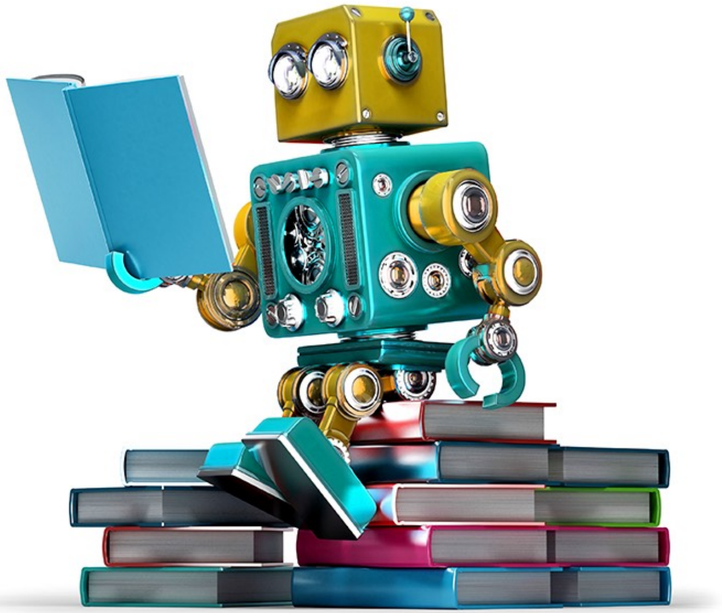
Outline

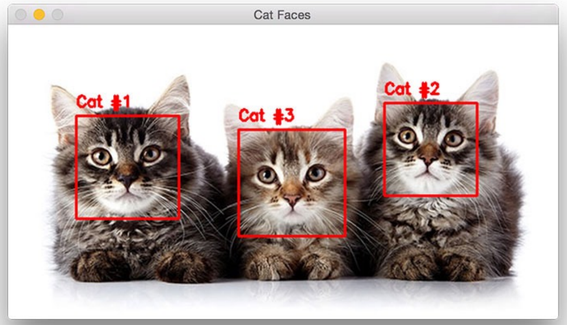
CISPA HELMHOLTZ-ZENTRUM i. G.

- Social network privacy
- Machine learning privacy

Machine Learning

CISPA
HELMHOLTZ-ZENTRUM i. G.





ML Security and Privacy

CISPA
HELMHOLTZ-ZENTRUM i. G.

- Many ML models are used in critical infrastructures
- ML models are trained on sensitive data
  - Biomedical data, emojis (Apple’s differential privacy)
- Largely overlooked

Research Question

CISPA
HELMHOLTZ-ZENTRUM i. G.

Does an ML model trained on privacy-sensitive data leak information of the data?

ML-Leaks: Model and Data Independent Membership Inference Attacks and Defenses on Machine Learning Models (to appear in NDSS 2019)

Membership Inference

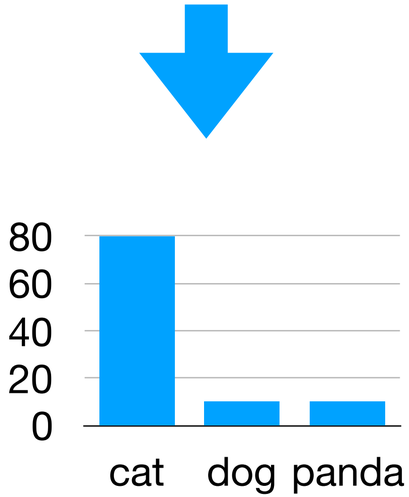
CISPA HELMHOLTZ-ZENTRUM i. G.

- Determine whether a data point is inside something
  - Biomedical data, case and reference group
  - Location data, NDSS 18’
- Machine learning models
  - Oakland 17’
- One of the most “popular” attacks in the community

ML Routine

Get some data

Train the model



Membership Inference

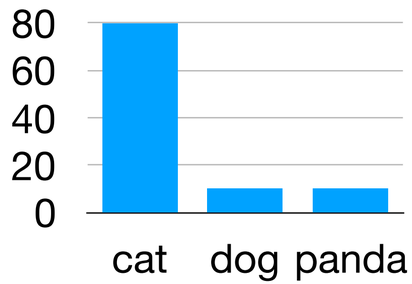
CISPA HELMHOLTZ-ZENTRUM i. G.

Get some data

Train the model

ML model

?

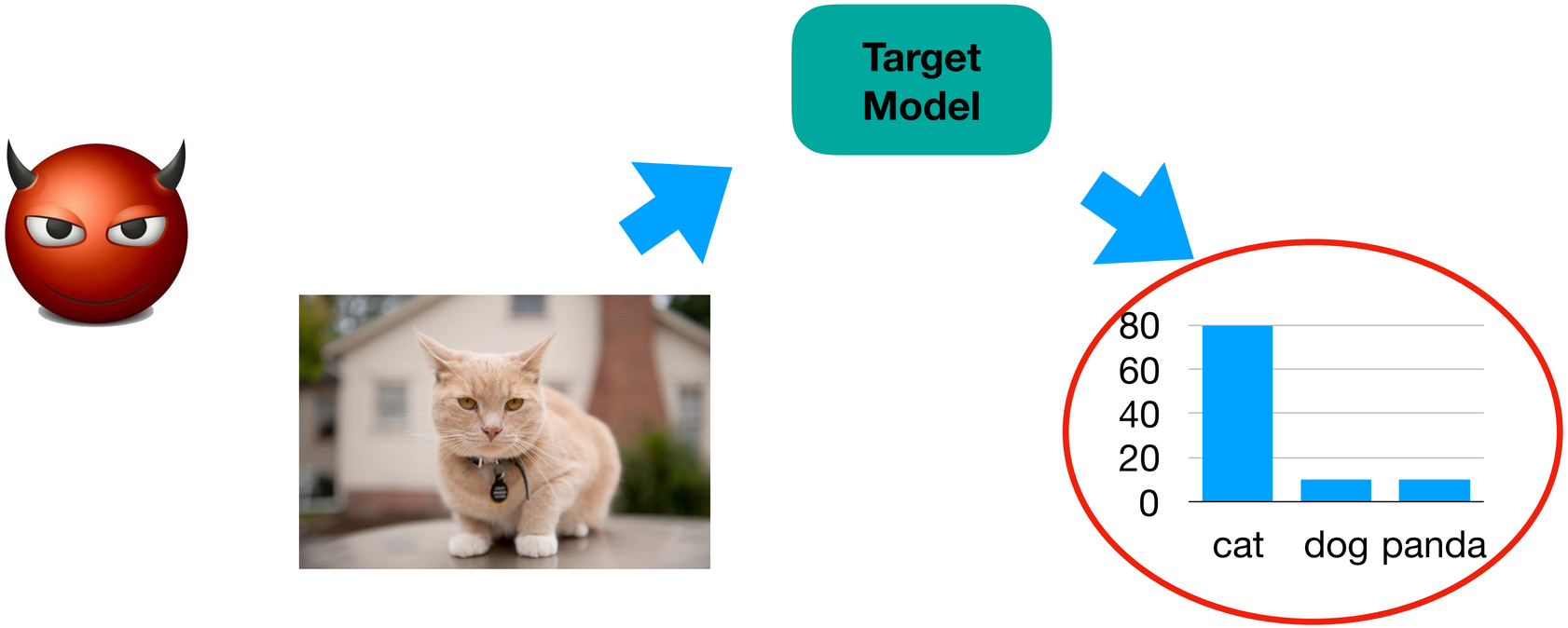


Membership Inference

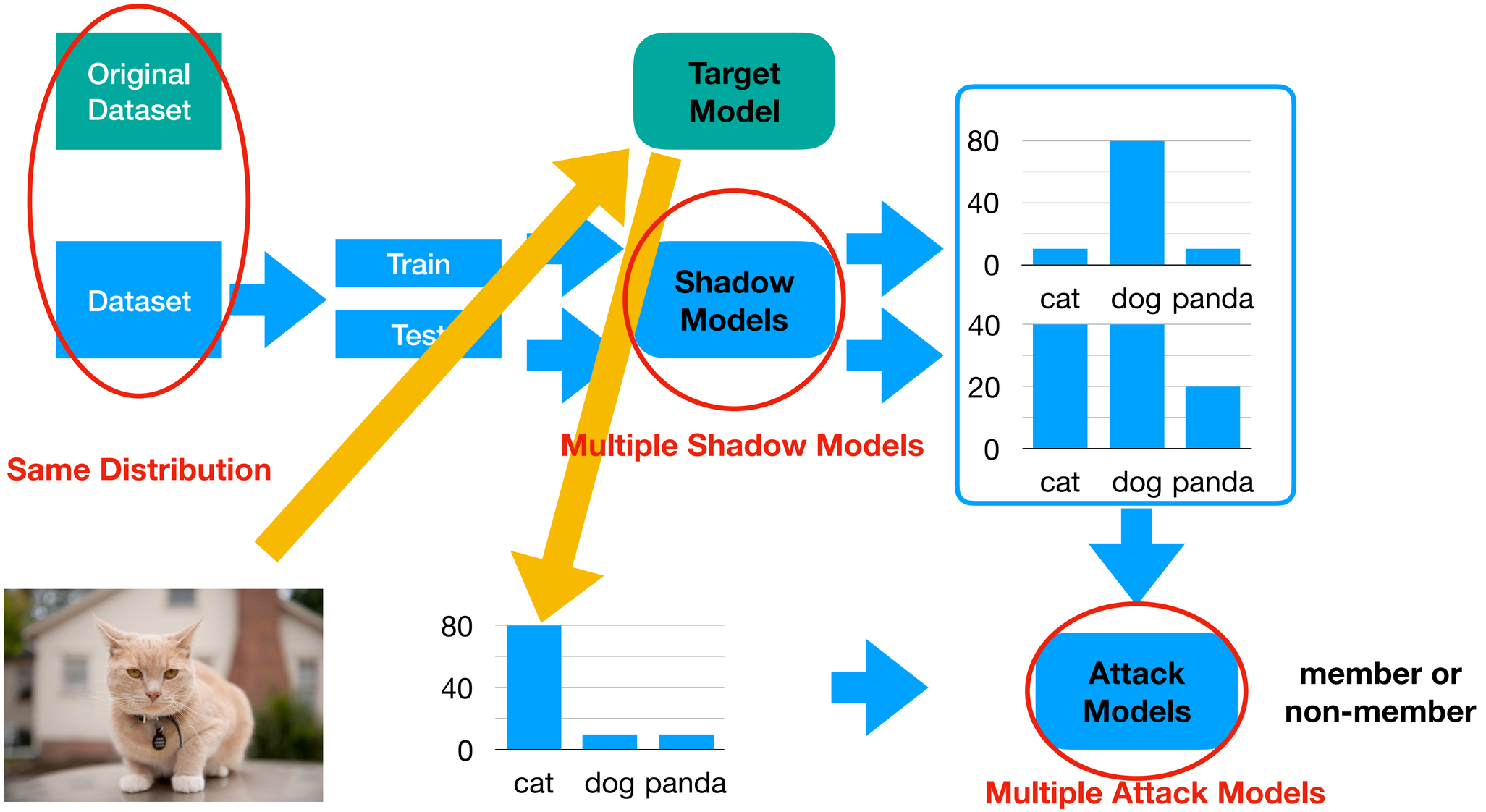
CISPA HELMHOLTZ-ZENTRUM i. G.

- Why membership matters?
- A cliché example: a ML model for medical diagnosis, if a person is in the training set, then she has the corresponding disease
- Security implications, IP implications

Threat Model



Attack by Shokri et al.

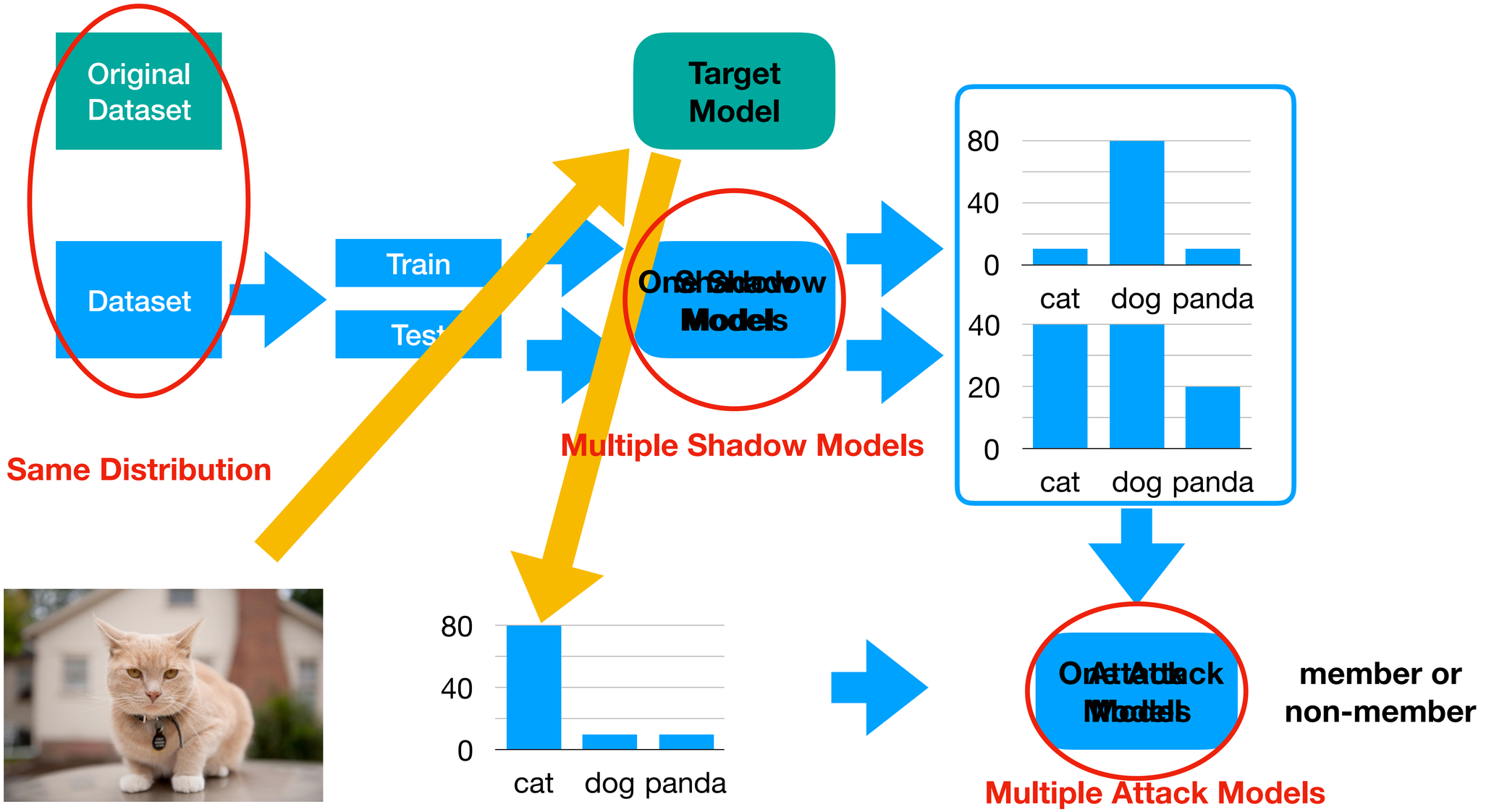


Our Attack 1

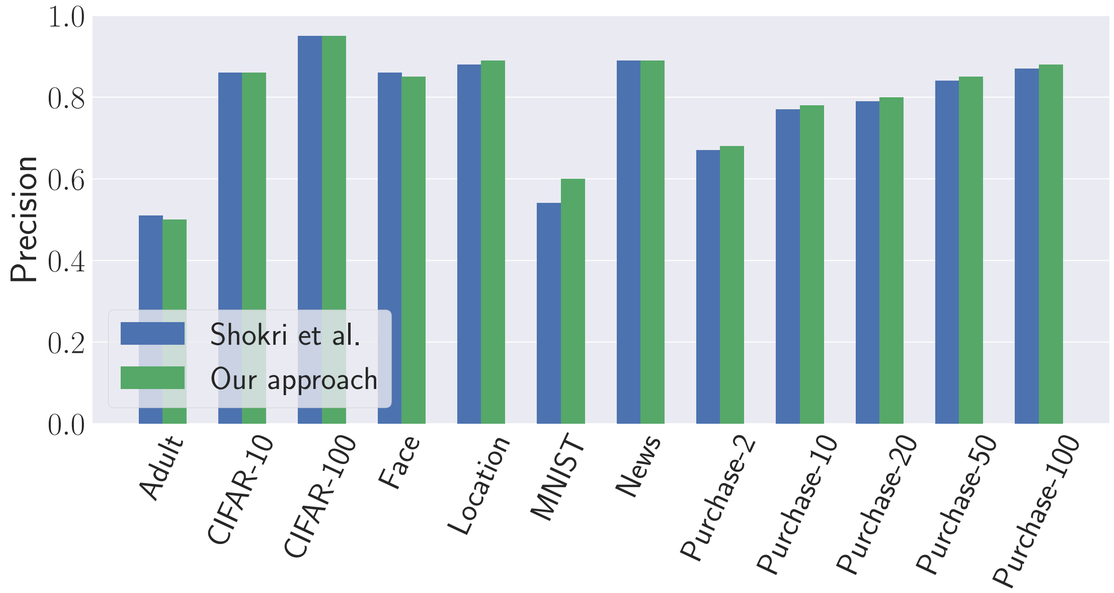
CISPA
HELMHOLTZ-ZENTRUM i. G.

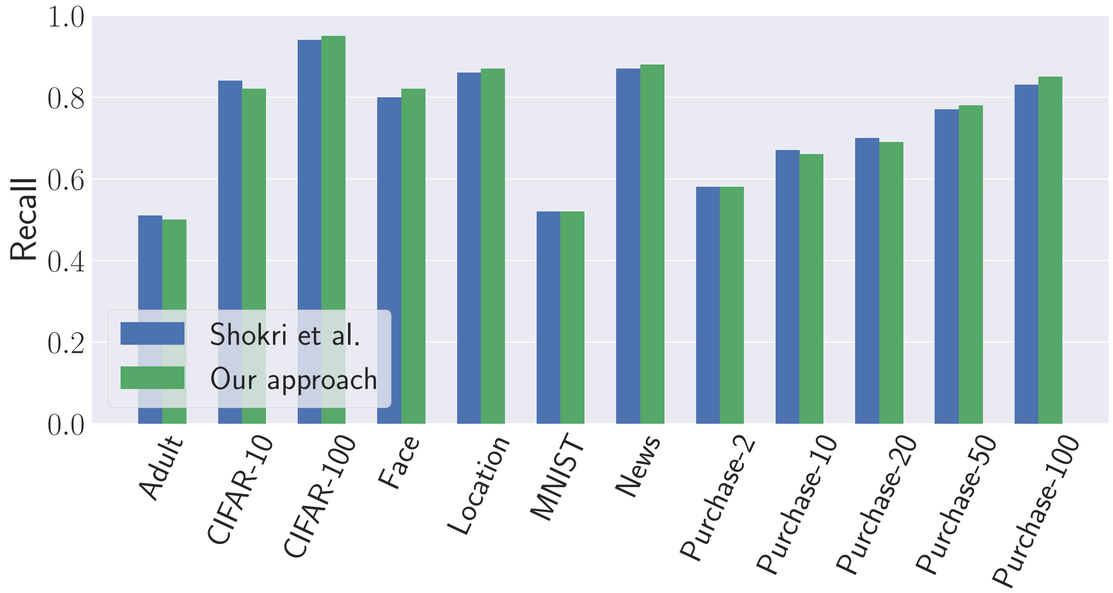
- One shadow model
- One attack model
- Same data distribution

Attack 1


Original Dataset
Dataset
Train
Test
Target Model
Shadow Models
Same Distribution
Multiple Shadow Models
80
40
0
cat dog panda
40
20
0
cat dog panda
80
40
0
cat dog panda
Attack Models
member or non-member
Multiple Attack Models

Our Attack 1


Precision
1.0
0.8
0.6
0.4
0.2
0.0
Shokri et al.
Our approach
Adult
CIFAR-10
CIFAR-100
Face
Location
MNIST
News
Purchase-2
Purchase-10
Purchase-20
Purchase-50
Purchase-100

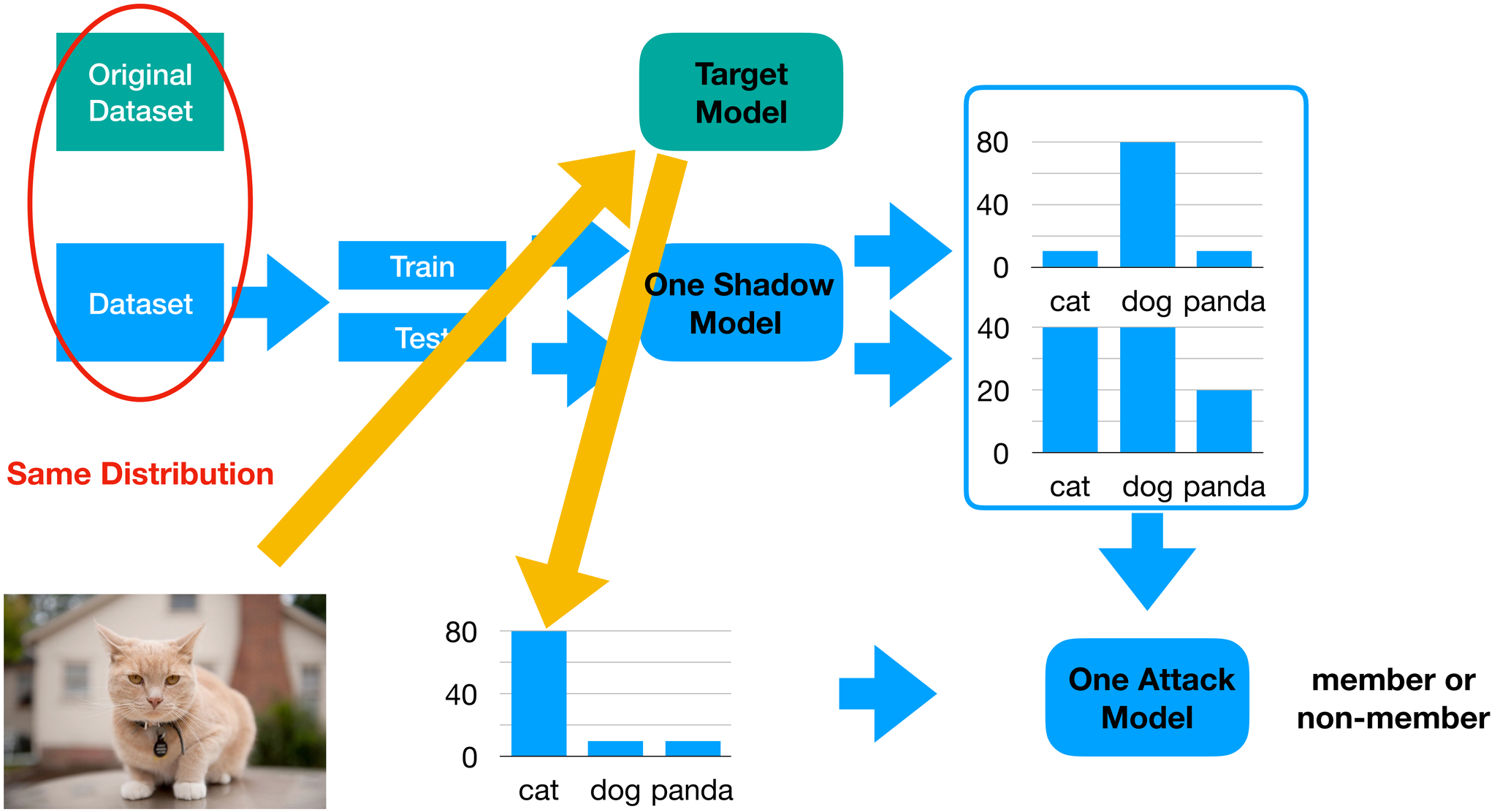

Recall
1.0
0.8
0.6
0.4
0.2
0.0
Shokri et al.
Our approach
Adult
CIFAR-10
CIFAR-100
Face
Location
MNIST
News
Purchase-2
Purchase-10
Purchase-20
Purchase-50
Purchase-100

Our Attack 2

Can we do better?

- No assumption on the dataset
- Data transferring attack
- Train shadow model on a different dataset, and attack on the target model

Attack 1


Original Dataset
Dataset
Train
Test
Target Model
One Shadow Model
Same Distribution
80
40
0
cat dog panda
40
20
0
cat dog panda
80
40
0
cat dog panda
One Attack Model
member or non-member

Attack 2

Image dataset

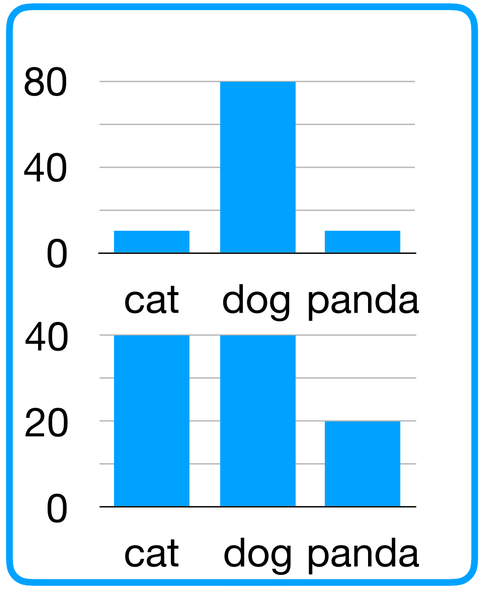
Text dataset

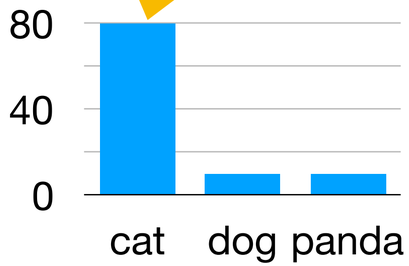
Train

Test

Target Model

One Shadow Model

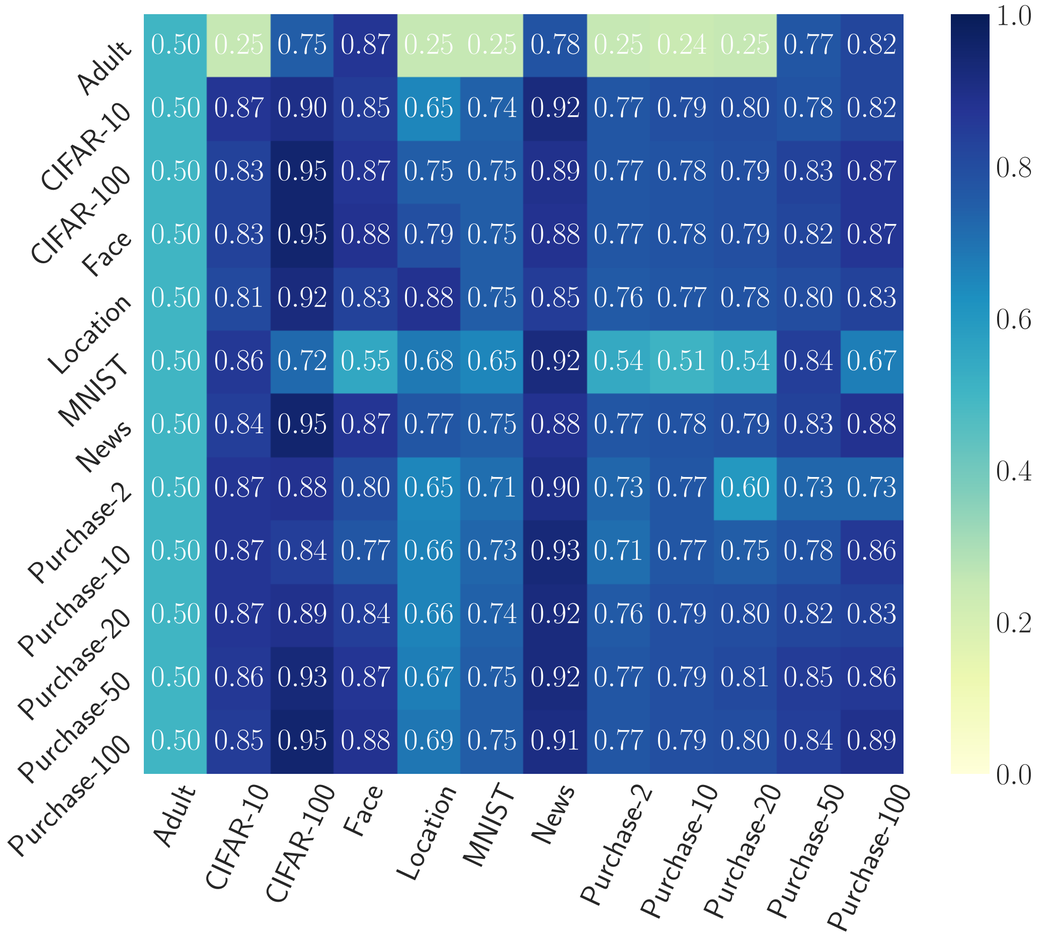




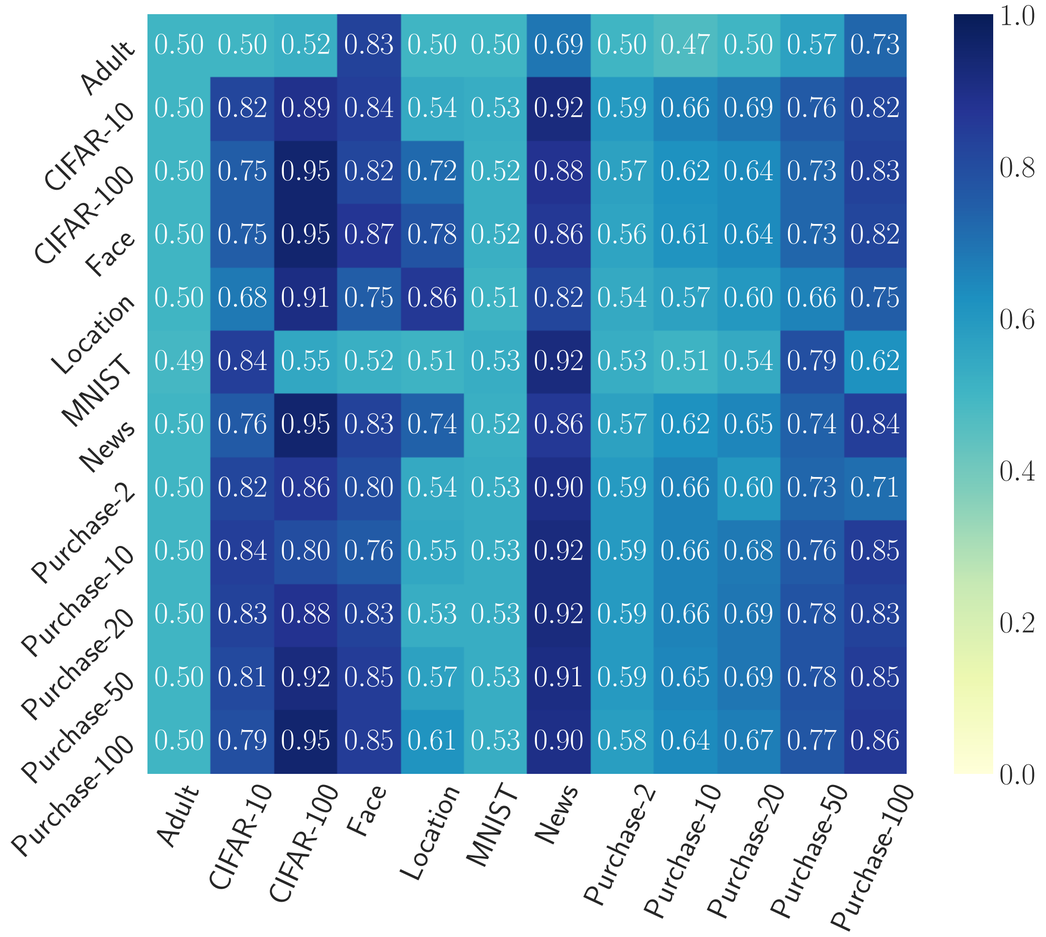
One Attack Model

member or non-member

Our Attack 2

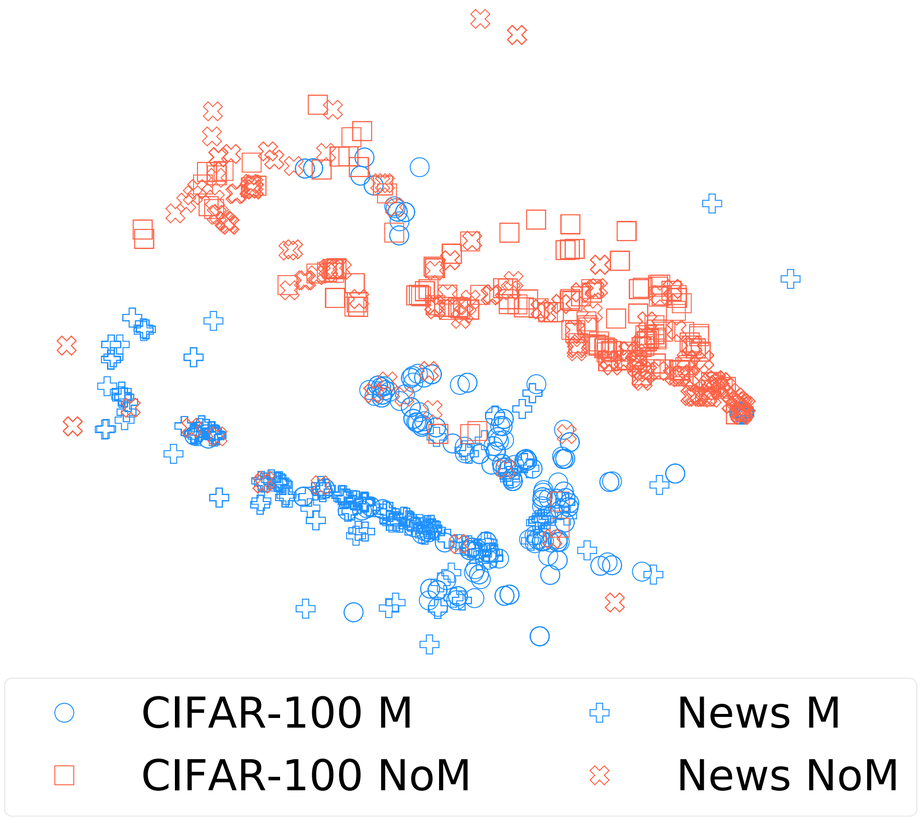


Precision



Recall

Sounds Magic, Why?



Attack 3

CISPA HELMHOLTZ-ZENTRUM i. G.

Can we do better?

- Get rid of the shadow model
- Take the maximum, std, or entropy of the posterior as the score
  - The simplest attack
  - Unsupervised
  - Reason: overfitting

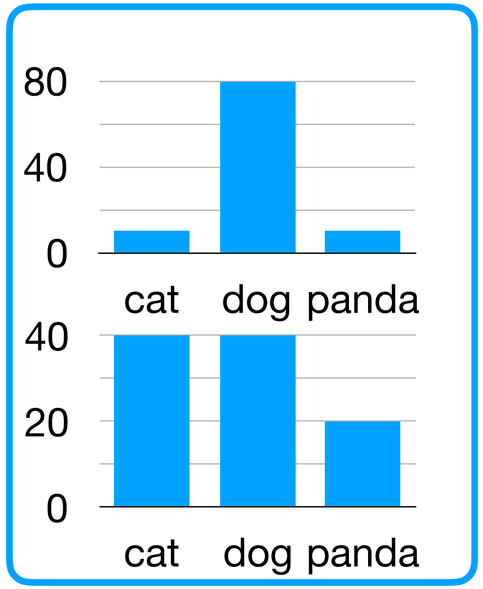
Attack 2

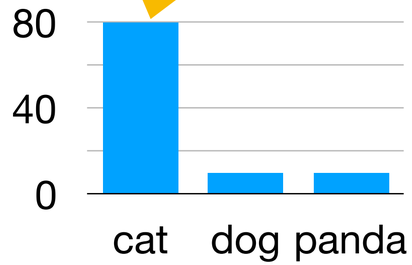
Train

Test

Target Model

One Shadow Model



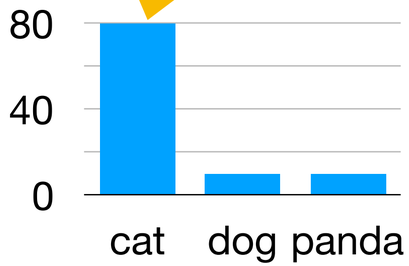


One Attack Model

member or non-member

Attack 3

Target Model



Unsupervised attack

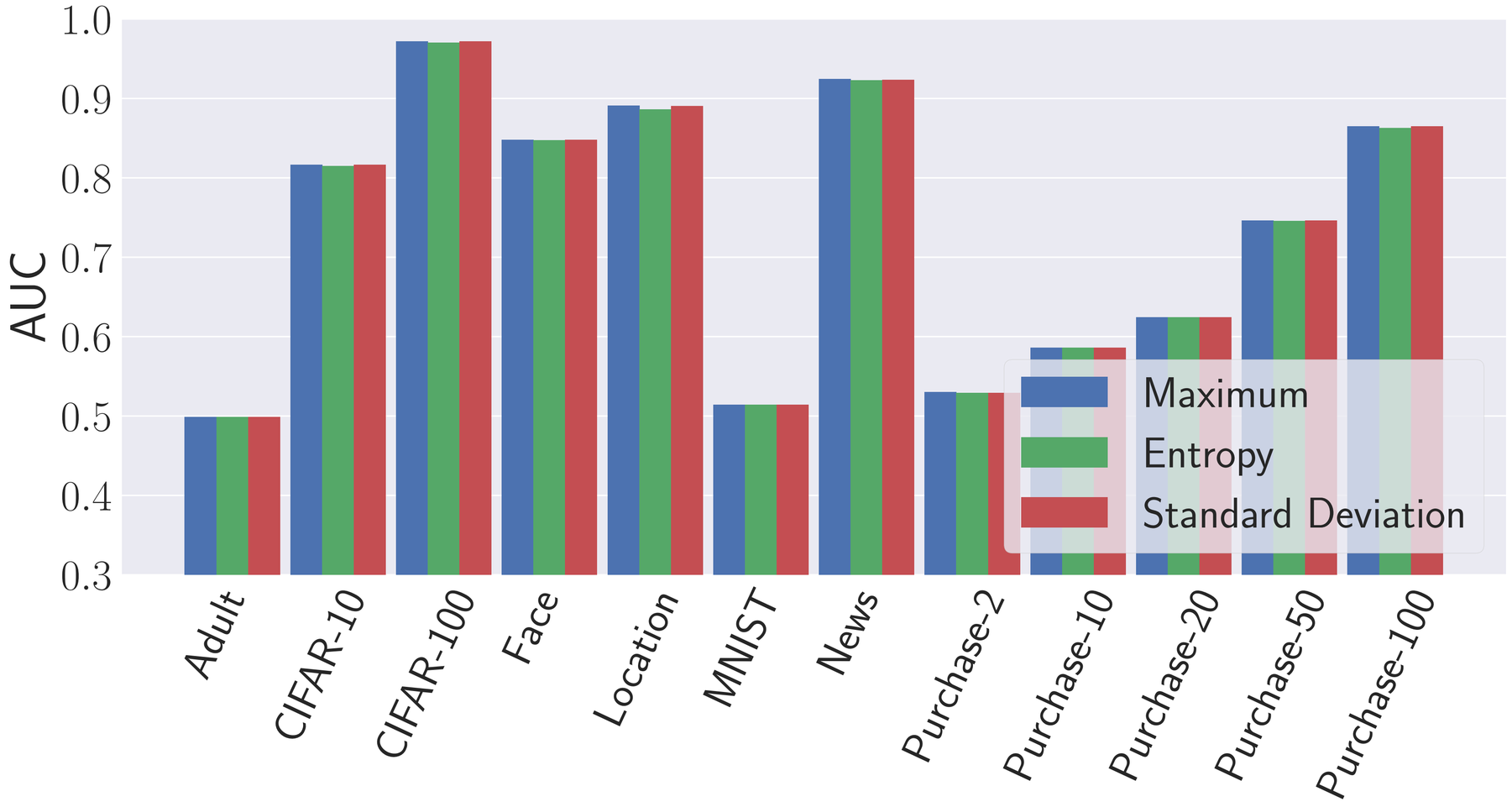
member or non-member

Attack 3

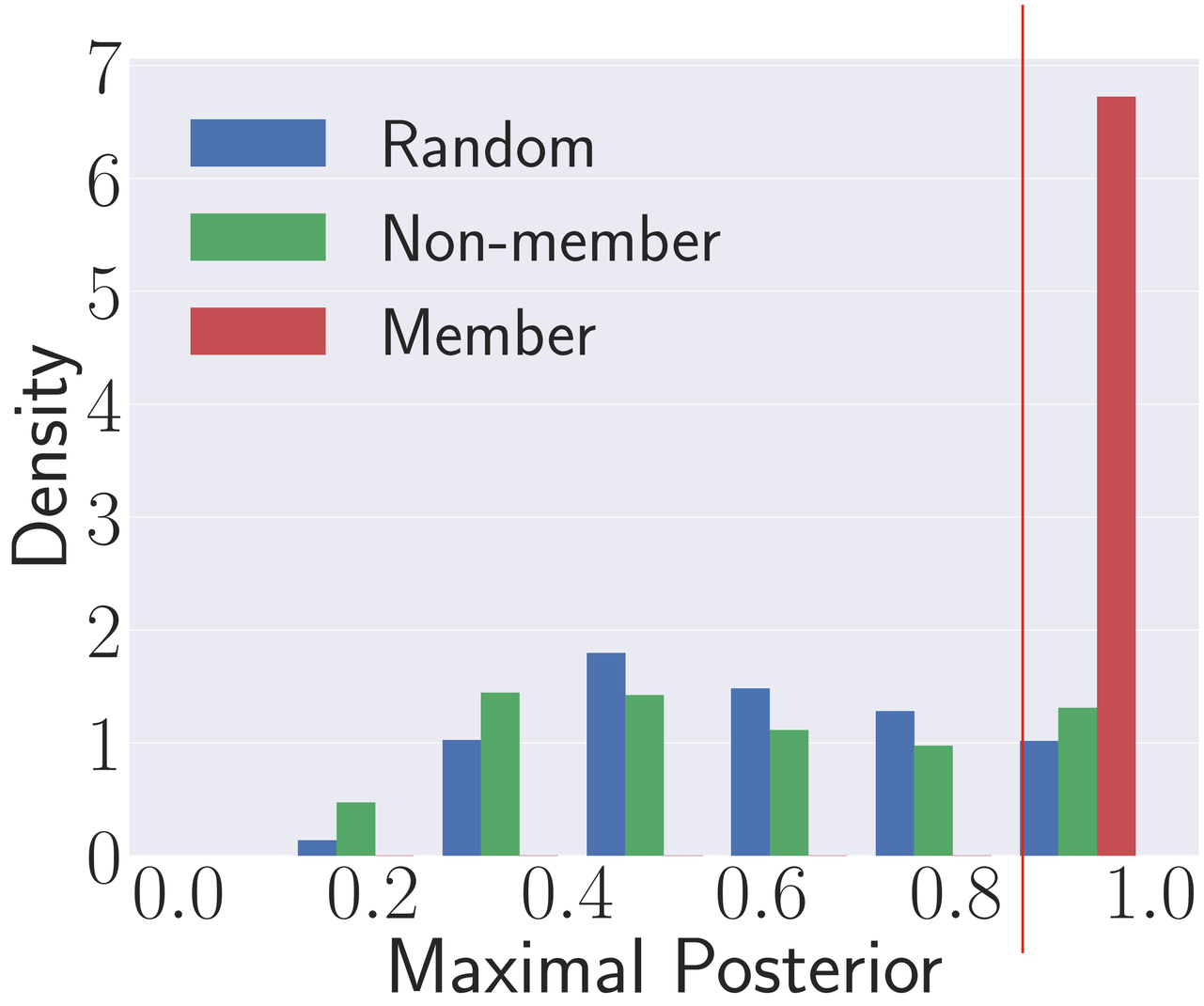
Unsupervised attack

1. Statistical measures over posterior
   - Maximum, Std, Entropy
2. Decide a threshold for the attack
   - Above ??% maximal posterior is member

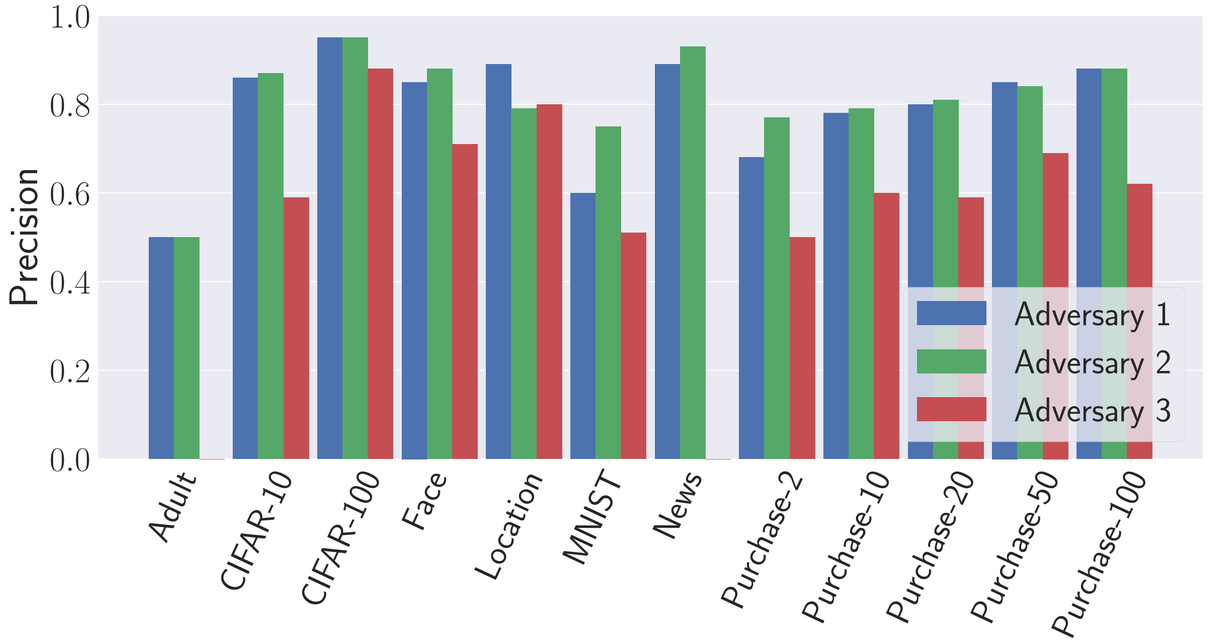
Attack 3

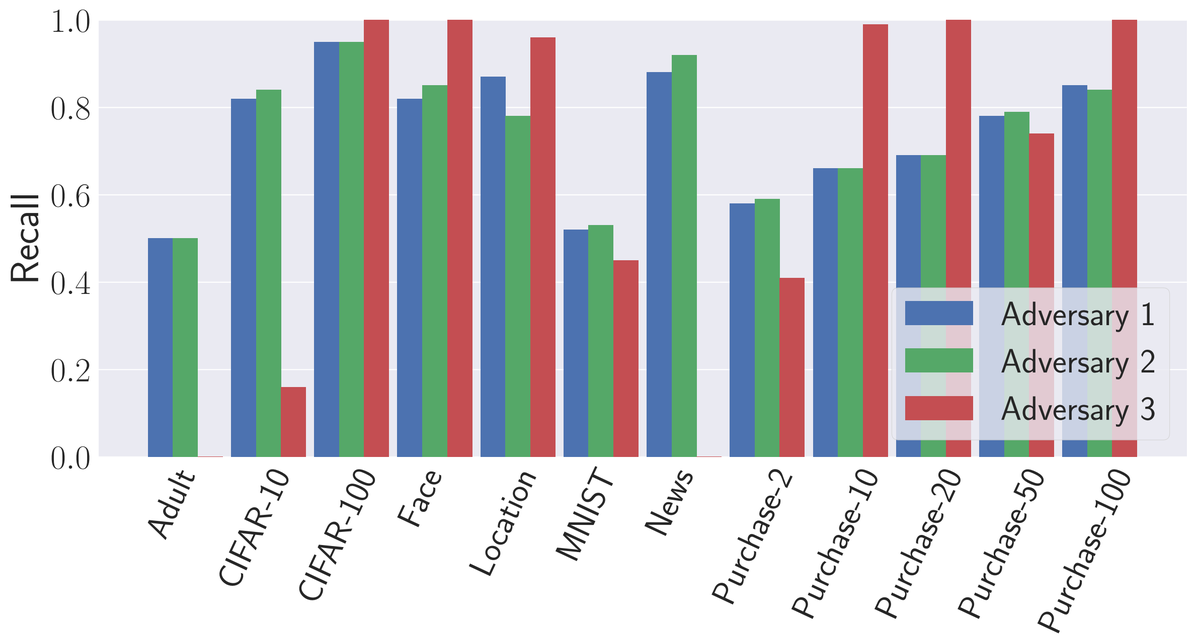

AUC
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
Adult
CIFAR-10
CIFAR-100
Face
Location
MNIST
News
Purchase-2
Purchase-10
Purchase-20
Purchase-50
Purchase-100
Maximum
Entropy
Standard Deviation

Threshold Picking



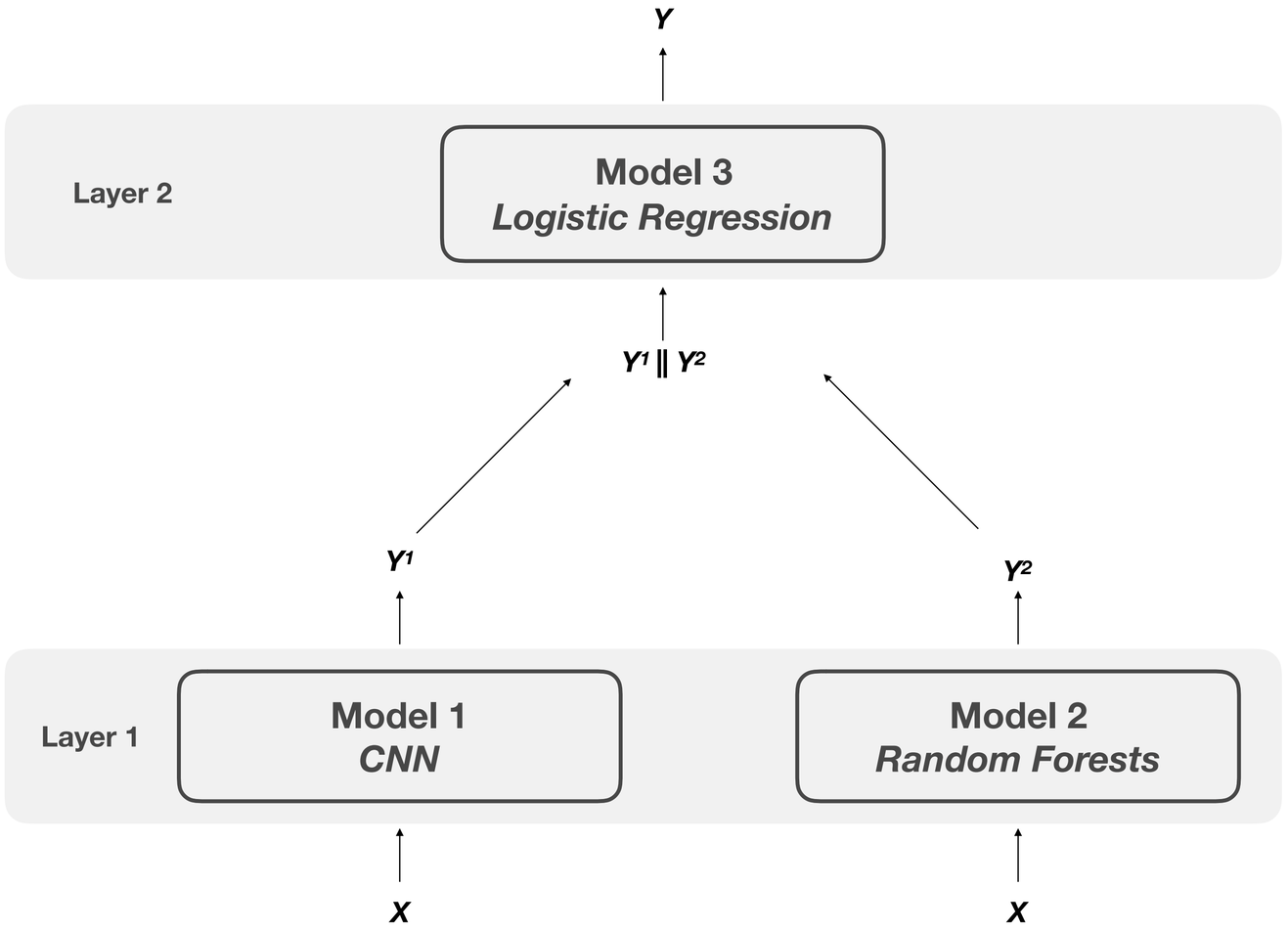
All Together



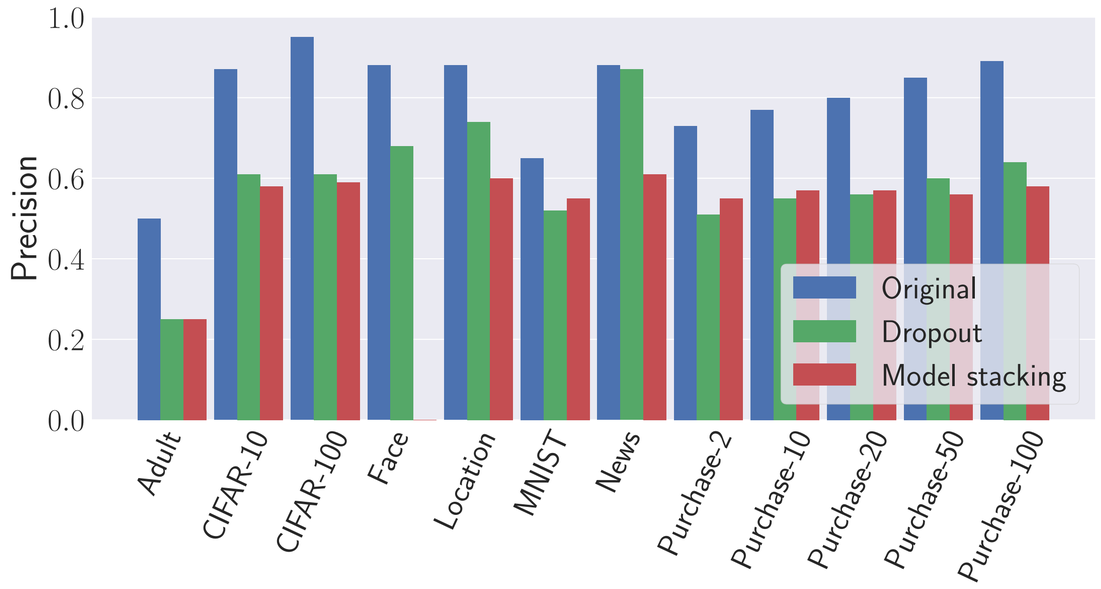


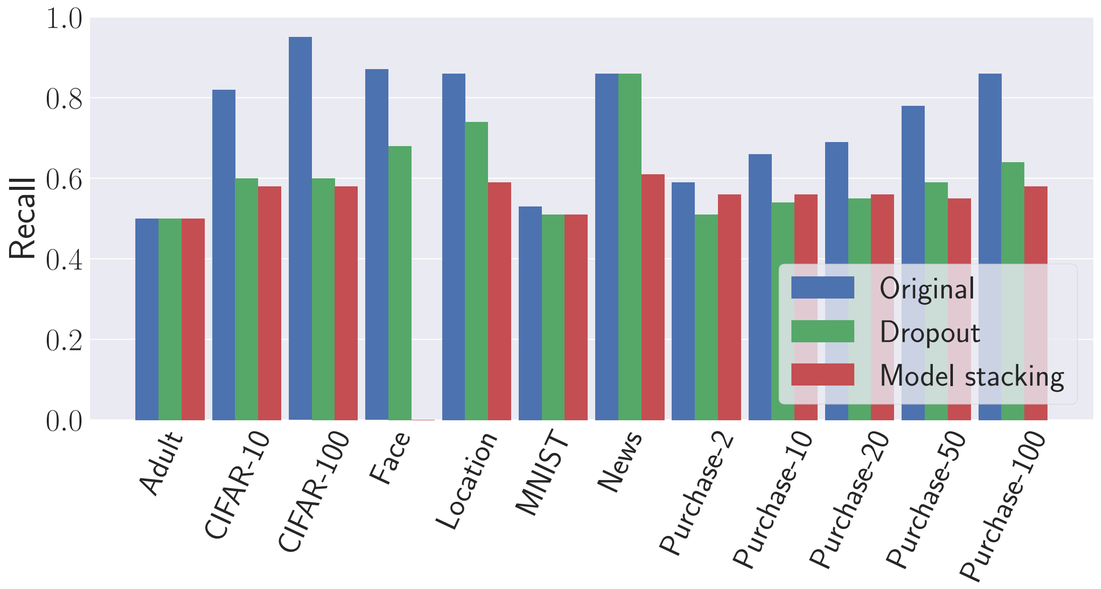
Defense

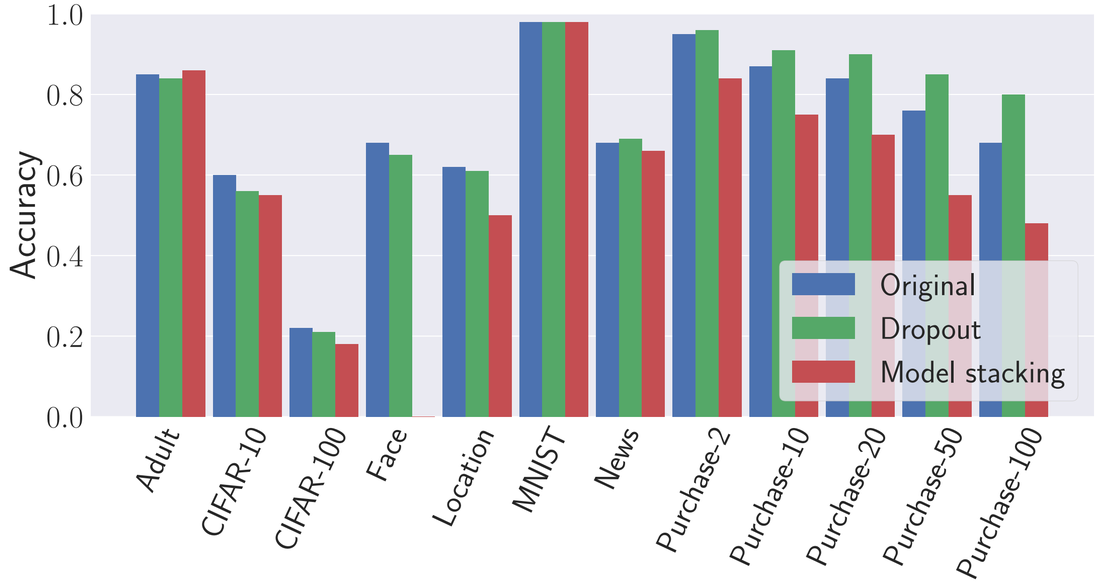
- Dropout
- Model stacking



Defense


Precision
1.0
0.8
0.6
0.4
0.2
0.0
Adult
CIFAR-10
CIFAR-100
Face
Location
MNIST
News
Purchase-2
Purchase-10
Purchase-20
Purchase-50
Purchase-100
Original
Dropout
Model stacking


Recall
1.0
0.8
0.6
0.4
0.2
0.0
Adult
CIFAR-10
CIFAR-100
Face
Location
MNIST
News
Purchase-2
Purchase-10
Purchase-20
Purchase-50
Purchase-100
Original
Dropout
Model stacking


Accuracy
1.0
0.8
0.6
0.4
0.2
0.0
Adult
CIFAR-10
CIFAR-100
Face
Location
MNIST
News
Purchase-2
Purchase-10
Purchase-20
Purchase-50
Purchase-100
Original
Dropout
Model stacking

Summary of ML-Leaks

CISPA HELMHOLTZ-ZENTRUM i. G.

- Machine learning models suffer from security and privacy attacks
  - Membership inference
- Three attacks with weak attacker assumption (more practical)
- How to evaluate a machine learning model?
  - Accuracy is enough?
  - Security and privacy matter
  - …. Just like buying a car
- A very promising direction for our research community!

Summary

- Social Network Privacy
  - walk2friends
  - Tagvisor
- Machine Learning Privacy
  - ML-Leaks

Thank you for your attention!
Questions?
http://yangzhangalmo.github.io/
@yangzhangalmo
yang.zhang@cispa.saarland

- Ph.D. positions and summer interns at CISPA
- CISPA-Stanford Program