ML Privacy Meter

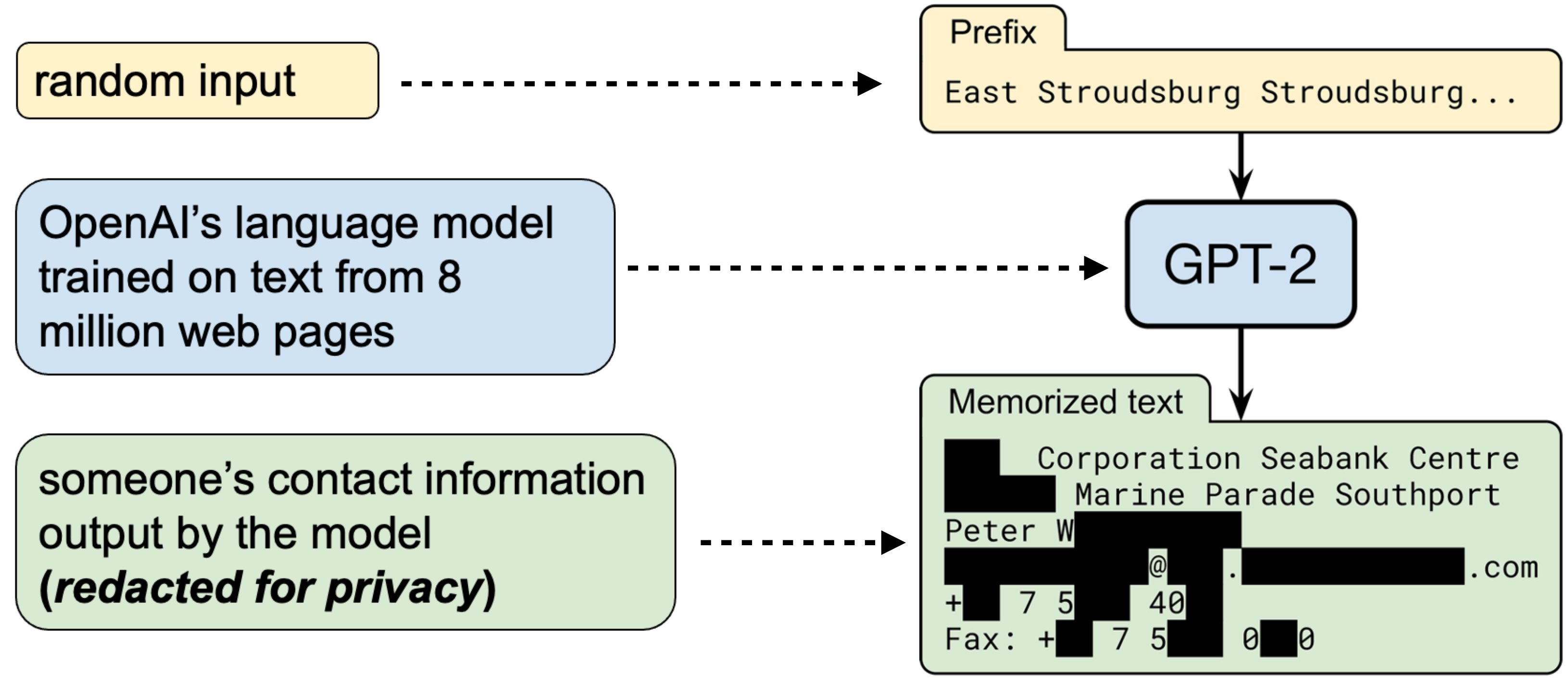
Reza Shokri

Data Privacy and Trustworthy ML Research Lab
National University of Singapore

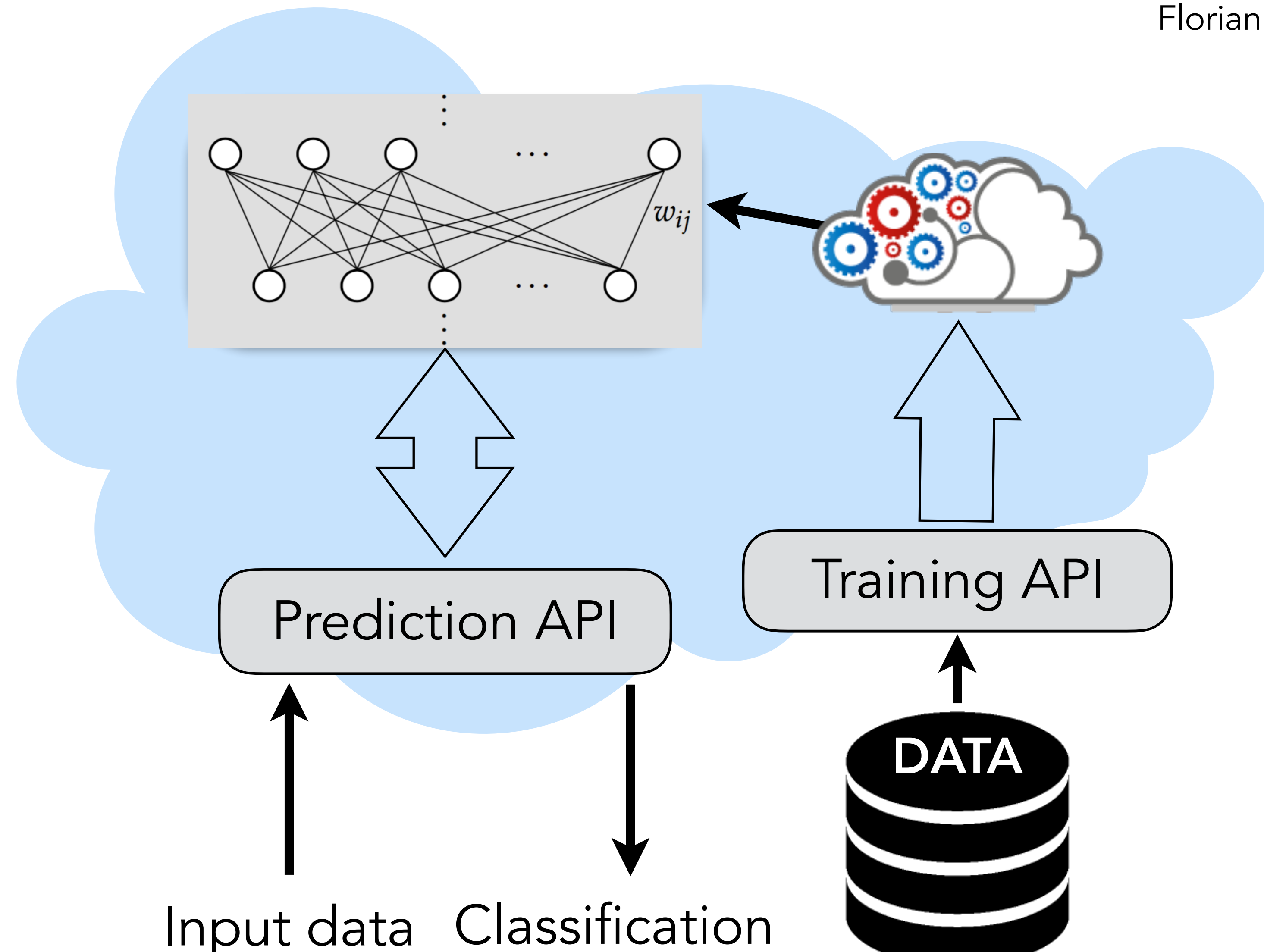
reza@comp.nus.edu.sg @rzshokri

Machine Learning Under Attack

Membership Inference Attack Accuracy: 90%



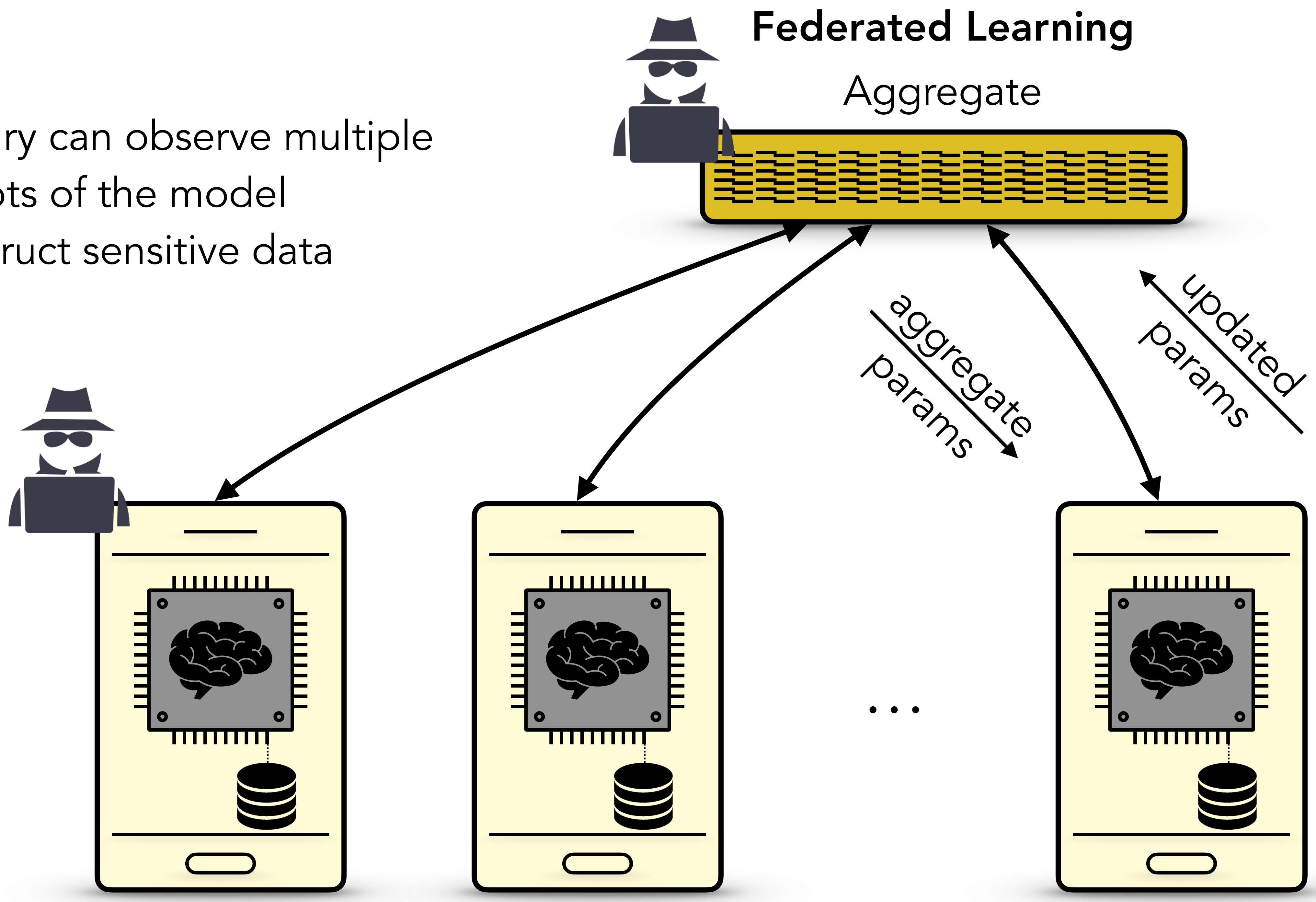
Florian Tramer



[Shokri, Stronati, Song, Shmatikov] Membership Inference Attacks against Machine Learning Models, SP'17
[Carlini, Tramer, et al.] Extracting Training Data from Large Language Models, Usenix security'21

Machine Learning Under Attack

- Adversary can observe multiple snapshots of the model
- Reconstruct sensitive data



[Shokri and Shmatikov] Privacy-Preserving Deep Learning, CCS'15

[Nasr, Shokri, Houmansadr] Comprehensive Privacy Analysis of Deep Learning: Passive and Active White-box Inference Attacks against Centralized and Federated Learning, SP'19

[Melis, Song, De Cristofaro, Shmatikov] Exploiting Unintended Feature Leakage in Collaborative Learning, SP'19

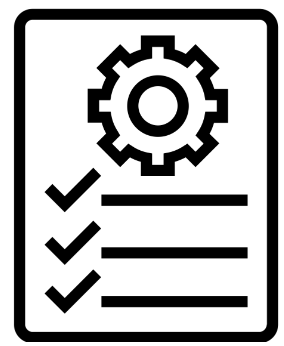
AI Regulations - Data Protection

- "… membership inferences show that AI models can inadvertently contain personal data" ico. Information Commissioner's Office
- "Attacks that reveal confidential information about the data include membership inference whereby …" NIST National Institute of Standards and Technology U.S. Department of Commerce
- ''….. ensuring that privacy and personal data are adequately protected during the use of AI''
- "……. ensuring that AI systems are resilient to overt attacks and subtle attacks that manipulate data or algorithms…."
- ''…should consider the risks to data throughout the design, development, and operation of an AI system''

On Artificial Intelligence - A European Approach to excellence and trust - Feb 2020
The White House Memo on Guidance for Regulation of Artificial Intelligence Applications - Jan 2020
Guidance on the AI auditing framework Draft guidance for consultation. Information Commissioner's Office
A Taxonomy and Terminology of Adversarial Machine Learning. Draft NISTIR 8269

Data Protection Impact Assessment

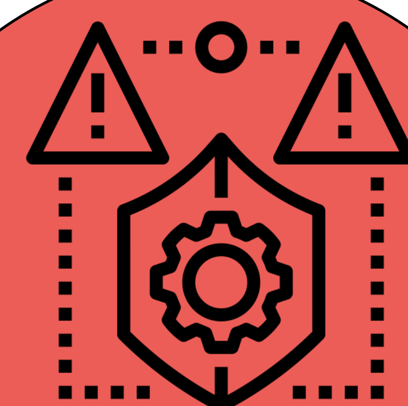
Systematic description of data collection, storage and processing



Assess necessity and proportionality

Likelihood and impact of the threats on individuals

Assess potential threats to the data



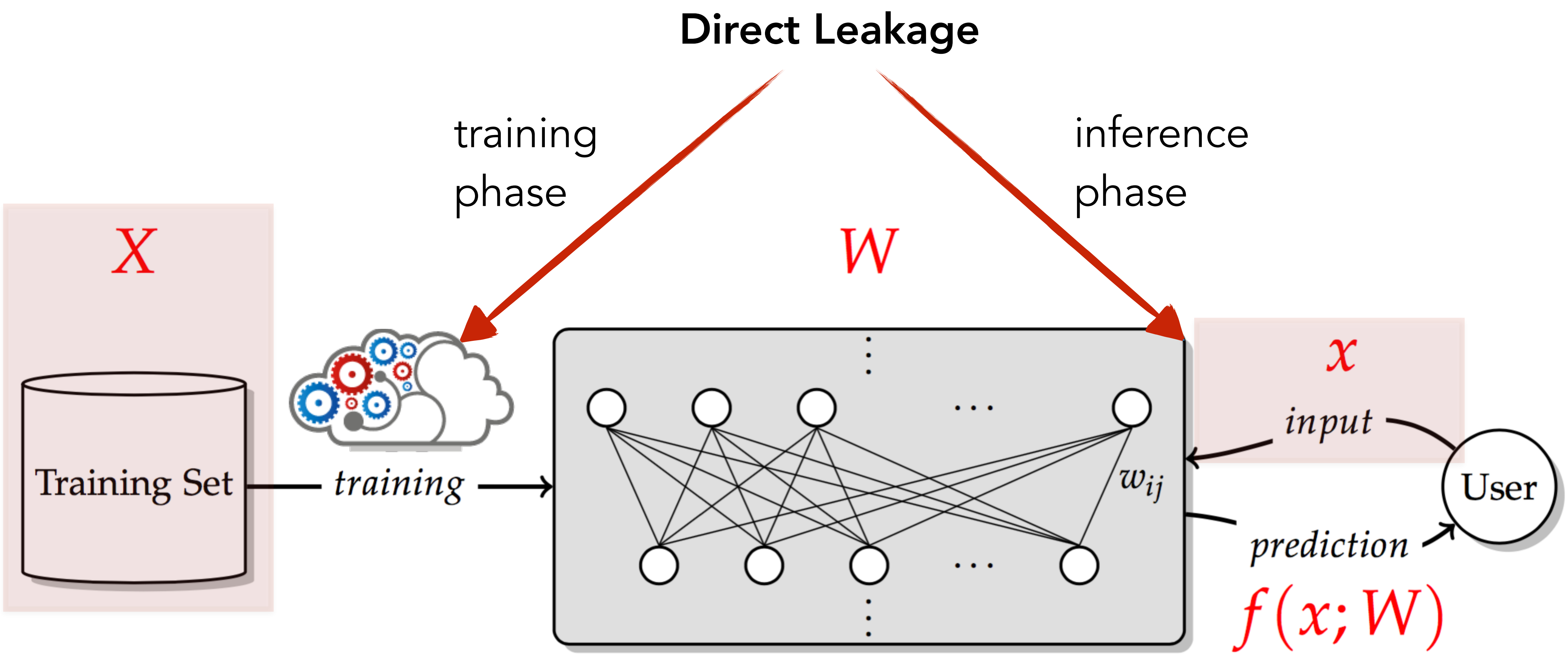
Identify and analyze possible risk mitigation measures

https://gdpr-info.eu/art-35-gdpr/

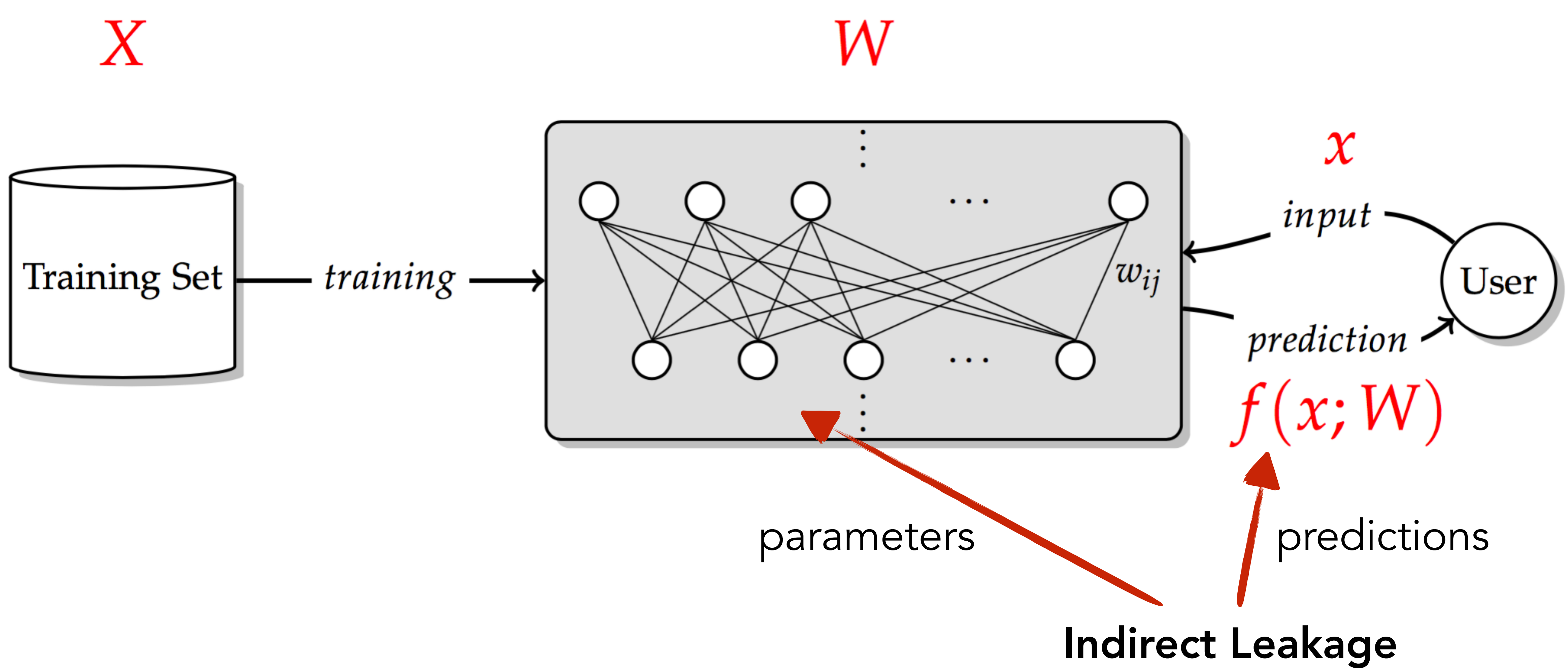
How to Quantitatively Measure the Privacy Risk of ML?

How to Check Compliance with Privacy Regulations?

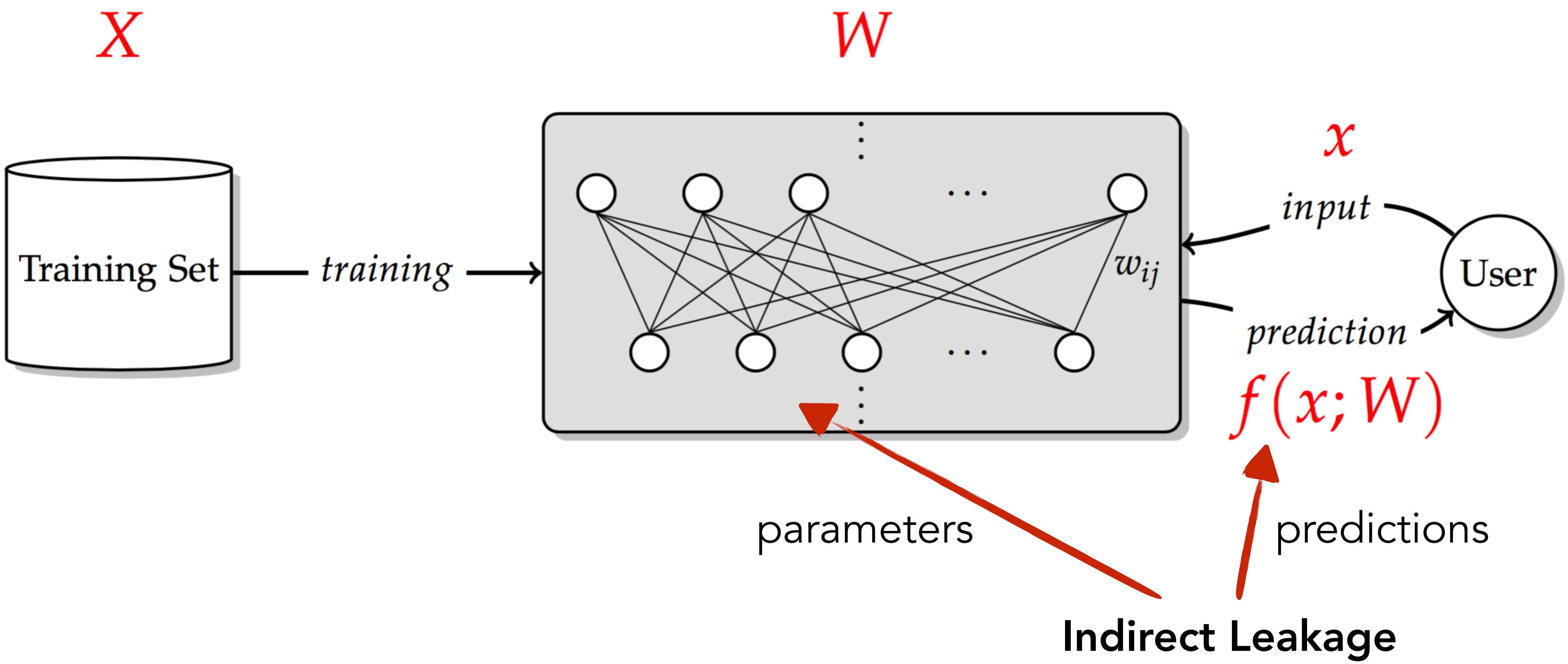
Privacy Risks in Machine Learning



Privacy Risks in Machine Learning

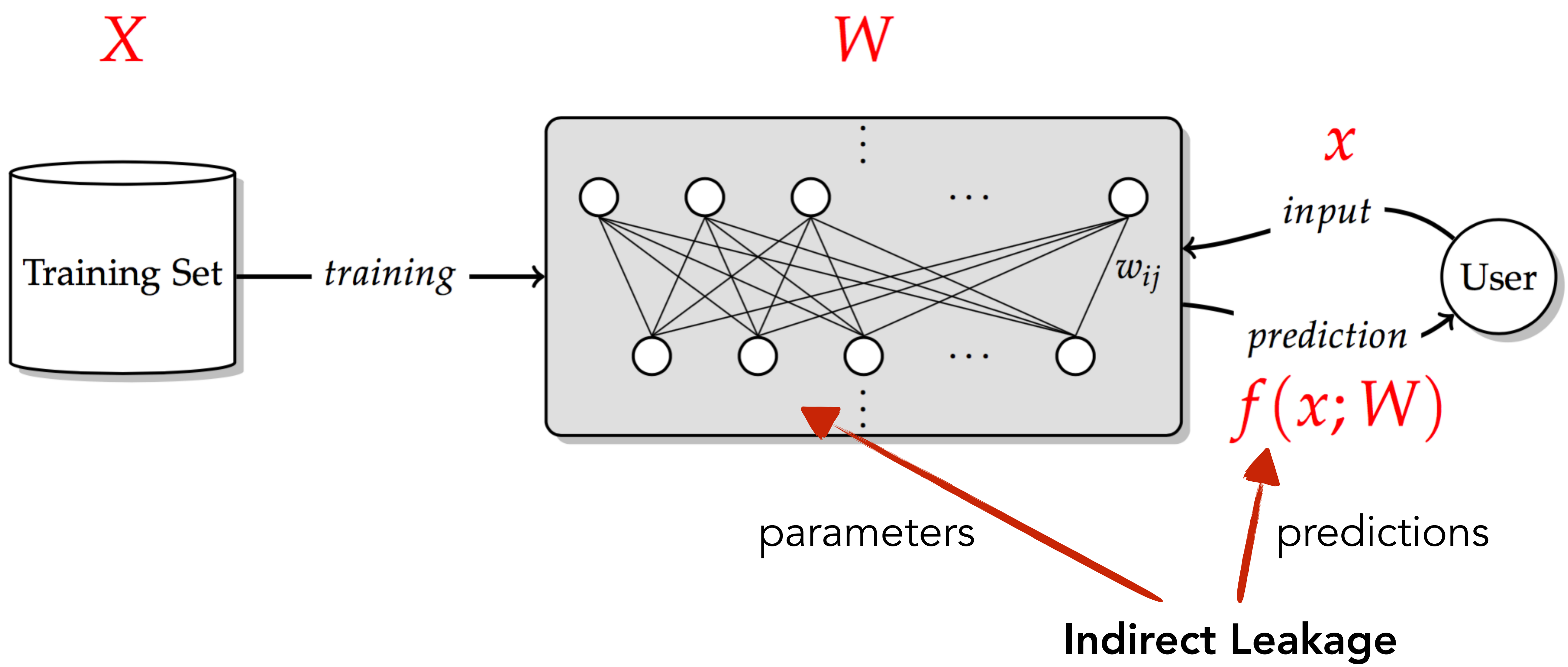


Privacy Risks in Machine Learning



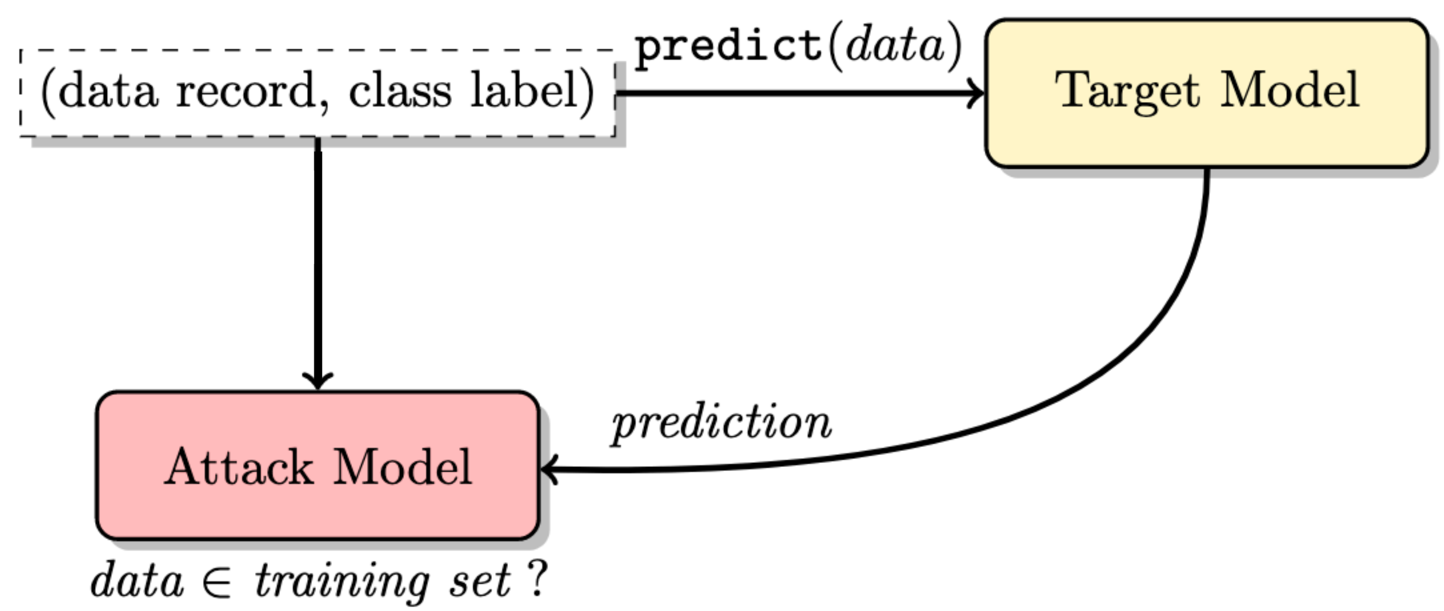
Privacy Risks in Machine Learning

What is leakage? Inferring information about members of X, beyond what can be learned about its underlying distribution



How to Quantify the Leakage?

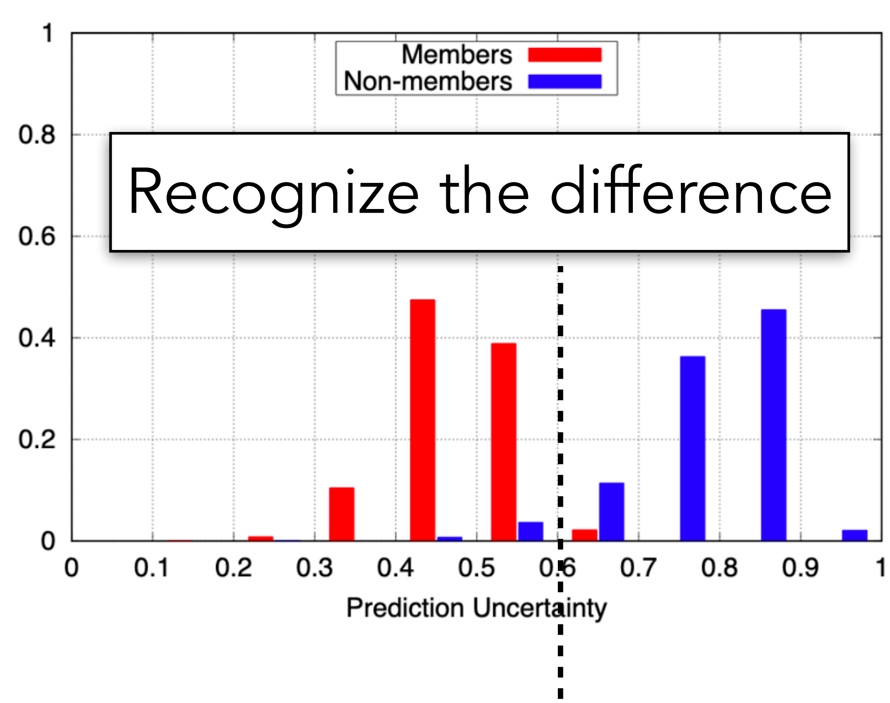
- Indistinguishability game: Can an adversary distinguish between two models that are trained on two neighboring datasets (one includes an extra data point x)?
  - Membership inference: Given a model, can an adversary infer whether data point x is part of its training set?

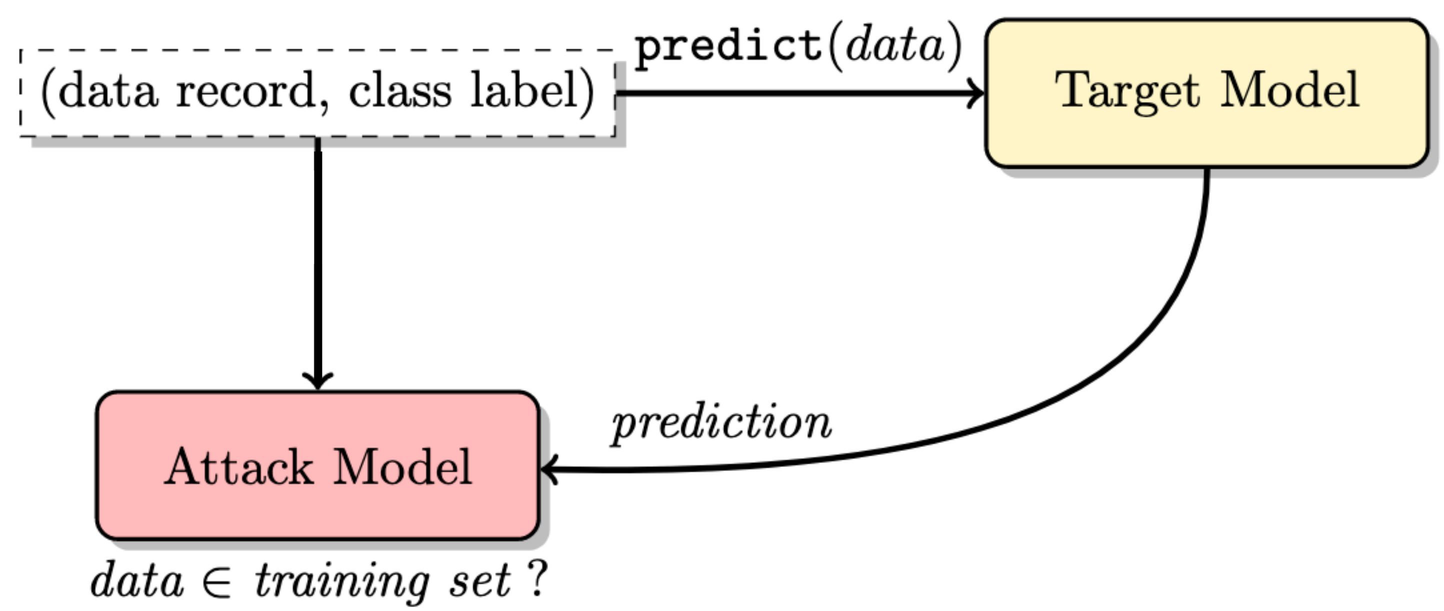


[Shokri, Stronati, Song, Shmatikov] Membership Inference Attacks against Machine Learning Models, SP'17

How to Quantify the Leakage?

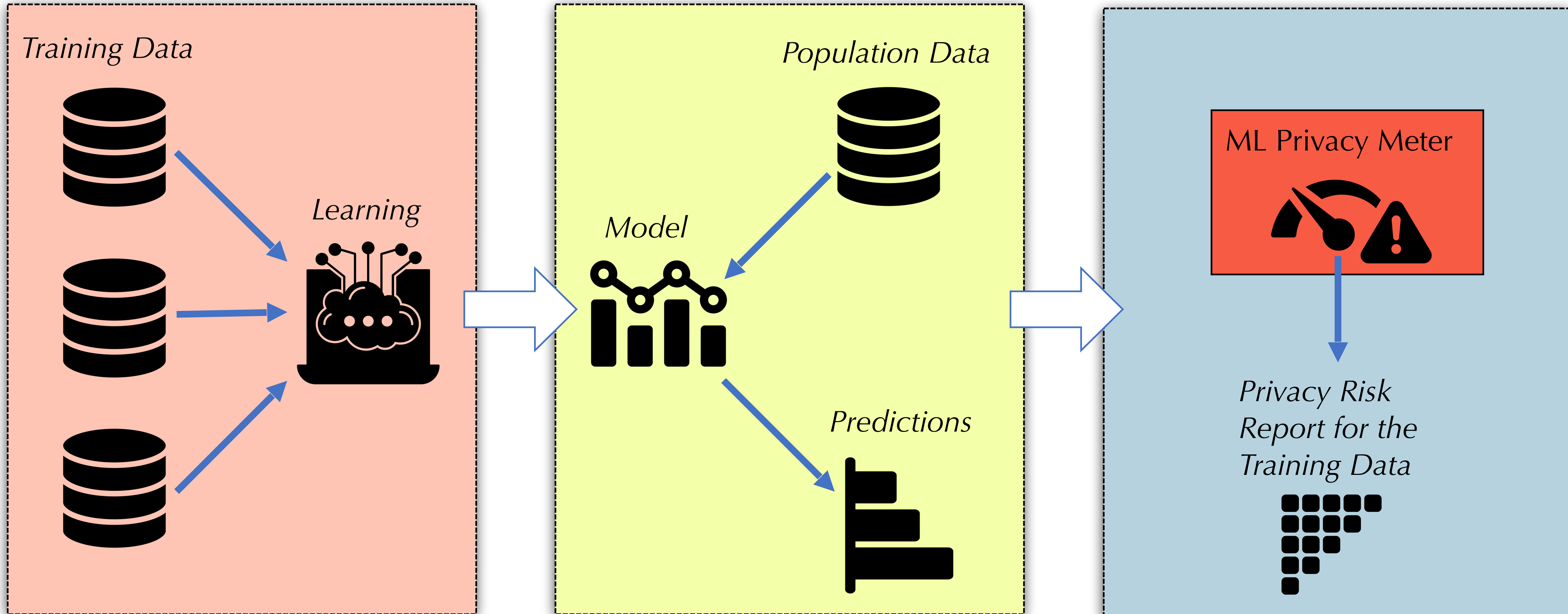
- Indistinguishability game: Can an adversary distinguish between two models that are trained on two neighboring datasets (one includes an extra data point x)?
    - <u>Membership inference</u>: Given a model, can an adversary infer whether data point x is part of its training set?





[Shokri, Stronati, Song, Shmatikov] Membership Inference Attacks against Machine Learning Models, SP'17

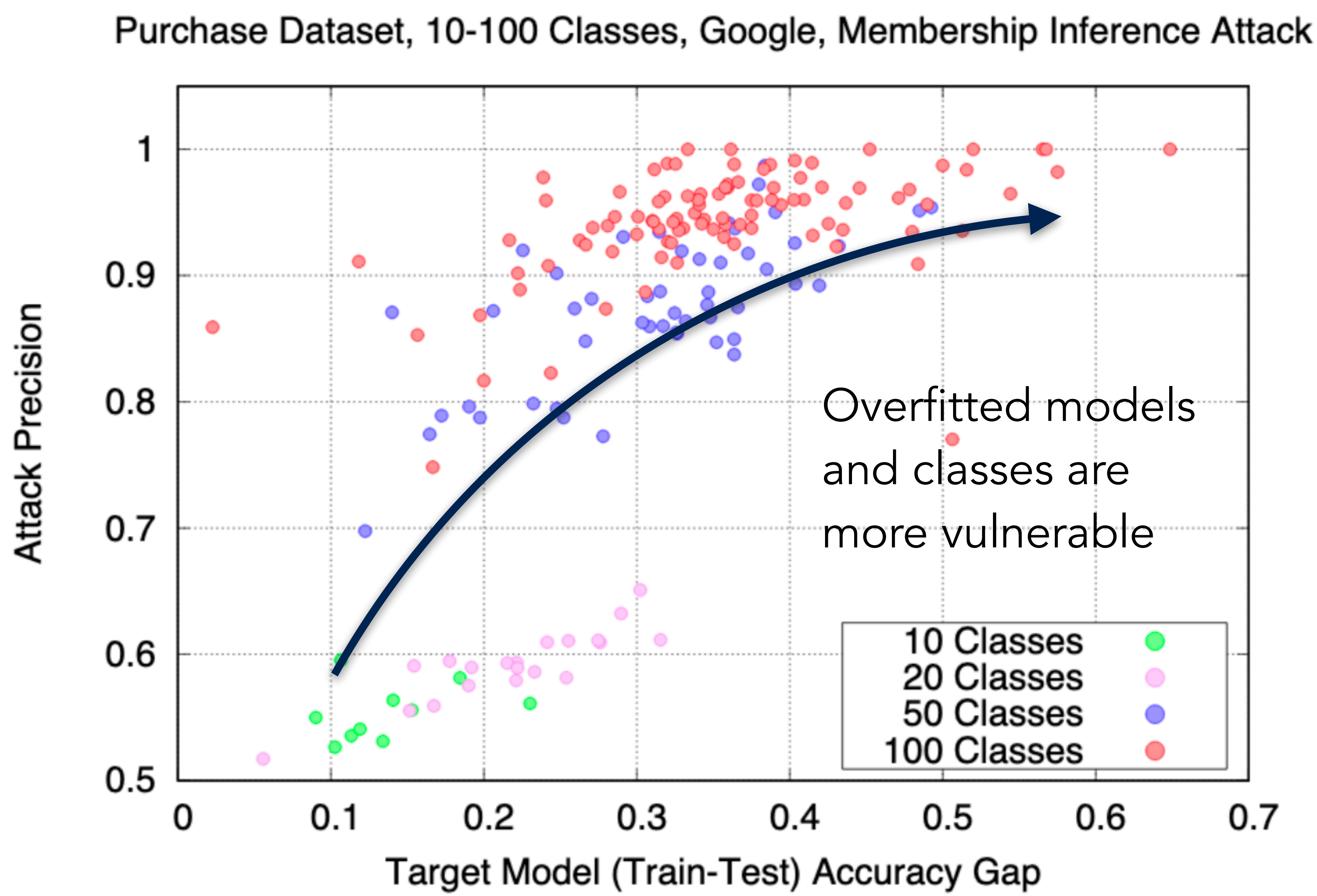
Tool: ML Privacy Meter



ML Privacy Meter is a Python library (`ml_privacy_meter`) that enables quantifying the privacy risks of machine learning models. https://github.com/privacytrustlab/ml_privacy_meter

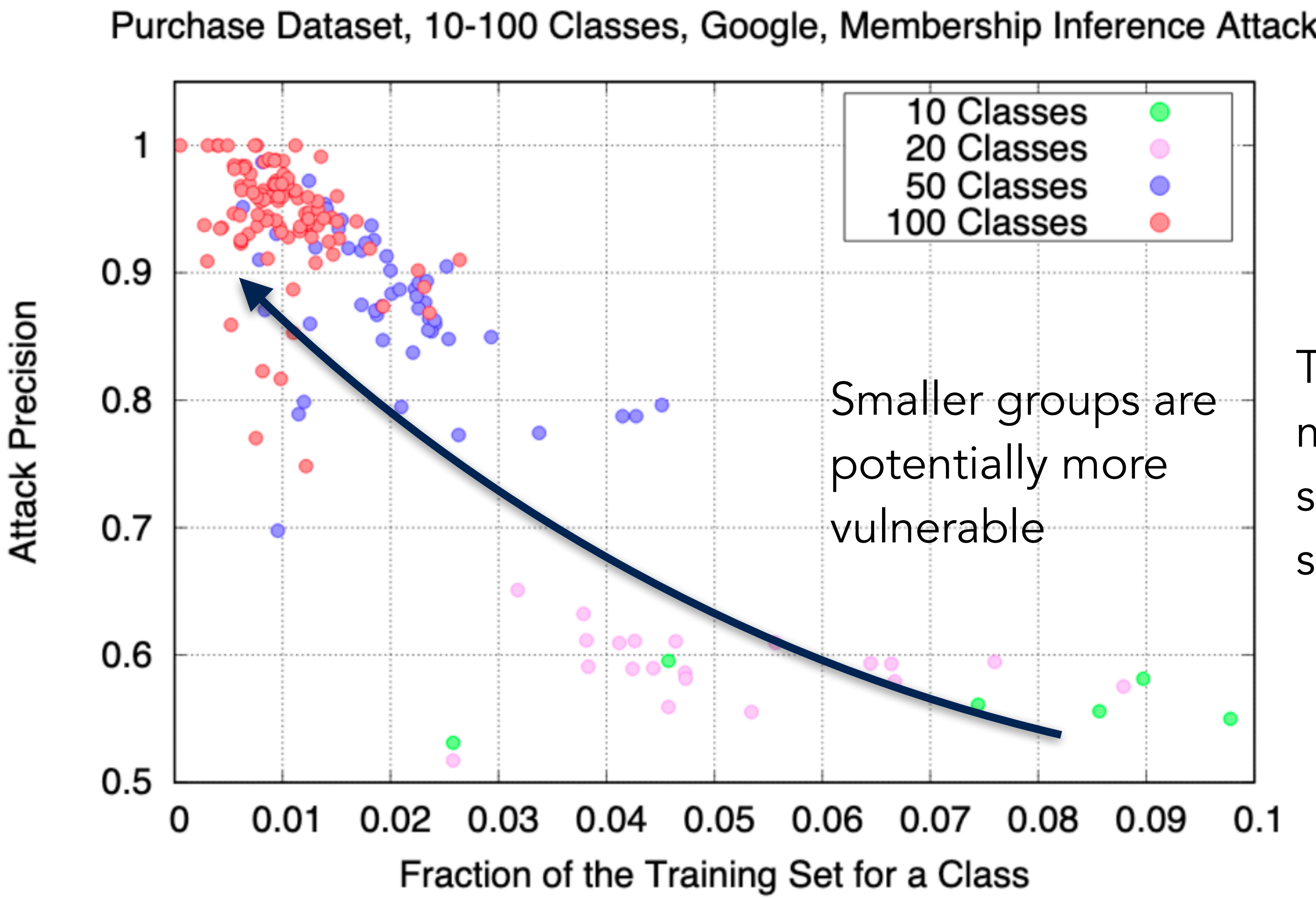
[Murakonda, Shokri] ML Privacy Meter: Aiding Regulatory Compliance by Quantifying the Privacy Risks of Machine Learning, HotPETs 2020

Privacy Leakage due to Overfitting



[Shokri, Stronati, Song, Shmatikov] Membership Inference Attacks against Machine Learning Models, SP'17

Disparate Privacy Vulnerability



The model memorizes samples from such groups

[Shokri, Stronati, Song, Shmatikov] Membership Inference Attacks against Machine Learning Models, SP'17

White-box Privacy Analysis

- Leakage through parameters (white-box) vs. predictions (black-box)

Most accurate pre-trained models				Mem inference attack accuracy		
Dataset	**Architecture**	**Train Accuracy**	**Test Accuracy**	**Black-box**	**White-box (Outputs)**	**White-box (Gradients)**
CIFAR100	Alexnet	99%	44%			
CIFAR100	ResNet	89%	73%			
CIFAR100	DenseNet	100%	82%			

[Nasr, Shokri, Houmansadr] Comprehensive Privacy Analysis of Deep Learning: Passive and Active White-box Inference Attacks against Centralized and Federated Learning, SP'19

White-box Privacy Analysis

- Leakage through parameters (white-box) vs. predictions (black-box)

Most accurate pre-trained models				Mem inference attack accuracy		
Dataset	**Architecture**	**Train Accuracy**	**Test Accuracy**	**Black-box**	**White-box (Outputs)**	**White-box (Gradients)**
CIFAR100	Alexnet	99%	44%			
CIFAR100	ResNet	89%	73%			
CIFAR100	DenseNet	$\boxed{100\%}$	$\boxed{82\%}$			

High generalizability
to test data

[Nasr, Shokri, Houmansadr] Comprehensive Privacy Analysis of Deep Learning: Passive and Active White-box Inference Attacks against Centralized and Federated Learning, SP'19

White-box Privacy Analysis

- Leakage through parameters (white-box) vs. predictions (black-box)

Most accurate pre-trained models				Mem inference attack accuracy		
Dataset	**Architecture**	**Train Accuracy**	**Test Accuracy**	**Black-box**	**White-box (Outputs)**	**White-box (Gradients)**
CIFAR100	Alexnet	99%	44%	74.2%	74.6%	75.1%
CIFAR100	ResNet	89%	73%	62.2%	62.2%	64.3%
CIFAR100	DenseNet	100%	82%	67.7%	67.7%	74.3%

High generalizability to test data

[Nasr, Shokri, Houmansadr] Comprehensive Privacy Analysis of Deep Learning: Passive and Active White-box Inference Attacks against Centralized and Federated Learning, SP'19

White-box Privacy Analysis

- Leakage through parameters (white-box) vs. predictions (black-box)

Most accurate pre-trained models				Mem inference attack accuracy		
Dataset	**Architecture**	**Train Accuracy**	**Test Accuracy**	**Black-box**	**White-box (Outputs)**	**White-box (Gradients)**
CIFAR100	Alexnet	99%	44%	74.2%	74.6%	75.1%
CIFAR100	ResNet	89%	73%	62.2%	62.2%	64.3%
CIFAR100	DenseNet	100%	82%	67.7%	67.7%	74.3%

High generalizability to test data

Low privacy (Significant leakage through parameters)

[Nasr, Shokri, Houmansadr] Comprehensive Privacy Analysis of Deep Learning: Passive and Active White-box Inference Attacks against Centralized and Federated Learning, SP'19

White-box Privacy Analysis

- Leakage through parameters (white-box) vs. predictions (black-box)

Most accurate pre-trained models				Mem inference attack accuracy		
Dataset	**Architecture**	**Train Accuracy**	**Test Accuracy**	**Black-box**	**White-box (Outputs)**	**White-box (Gradients)**
CIFAR100	Alexnet	99%	44%	74.2%	74.6%	75.1%
CIFAR100	ResNet	89%	73%	62.2%	62.2%	64.3%
CIFAR100	DenseNet	100%	82%	67.7%	67.7%	74.3%

Large capacity

High generalizability to test data

Low privacy (Significant leakage through parameters)

[Nasr, Shokri, Houmansadr] Comprehensive Privacy Analysis of Deep Learning: Passive and Active White-box Inference Attacks against Centralized and Federated Learning, SP'19

White-box Privacy Analysis

- Leakage through parameters (white-box) vs. predictions (black-box)

Most accurate pre-trained models				Mem inference attack accuracy		
Dataset	**Architecture**	**Train Accuracy**	**Test Accuracy**	**Black-box**	**White-box (Outputs)**	**White-box (Gradients)**
CIFAR100	Alexnet	99%	44%	74.2%	74.6%	75.1%
CIFAR100	ResNet	89%	73%	62.2%	62.2%	64.3%
CIFAR100	DenseNet	$\boxed{100\%}$	$\boxed{82\%}$	67.7%	67.7%	$\boxed{74.3\%}$

Large capacity

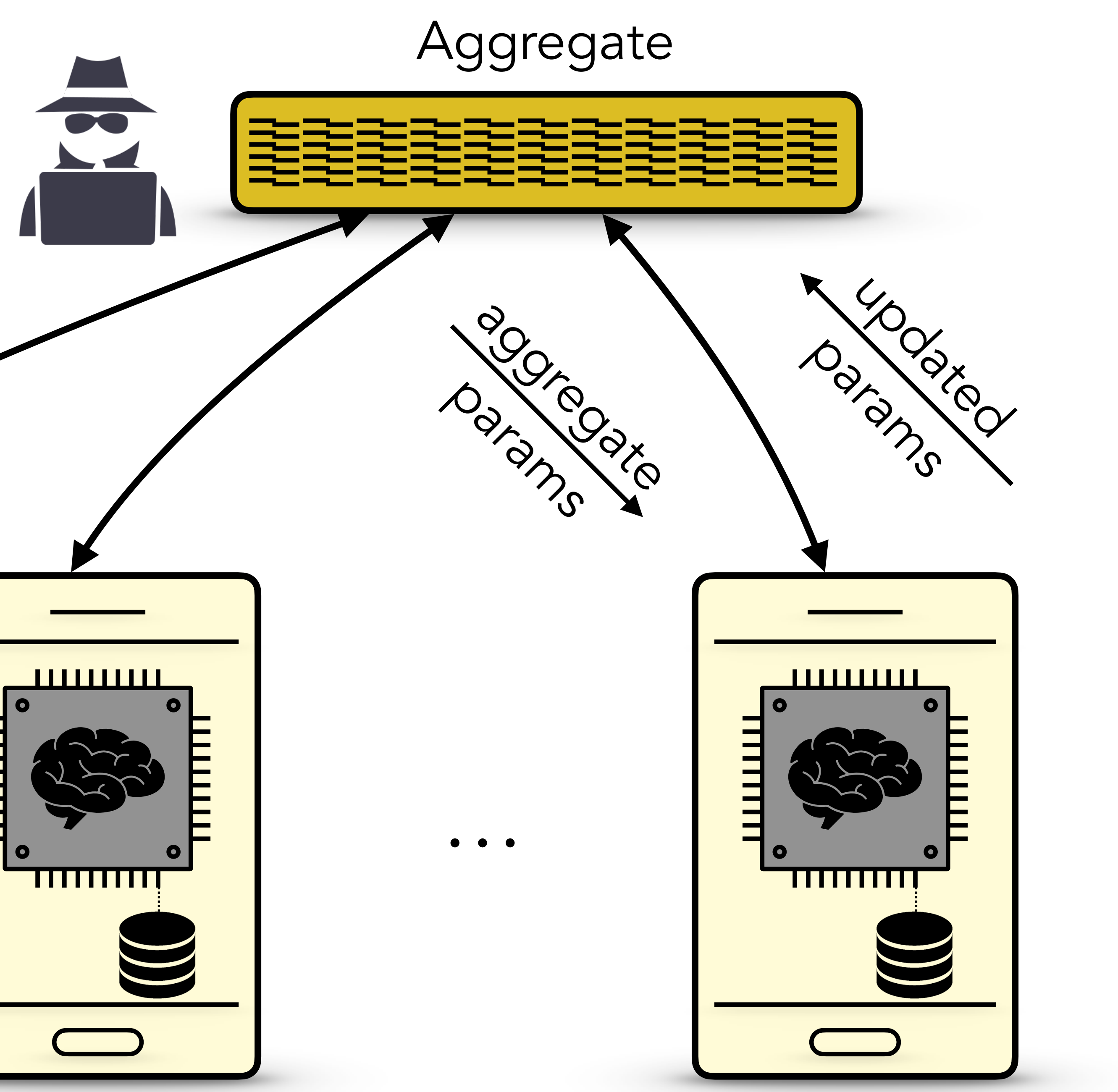
High generalizability to test data

Low privacy (Significant leakage through parameters)

[Nasr, Shokri, Houmansadr] Comprehensive Privacy Analysis of Deep Learning: Passive and Active White-box Inference Attacks against Centralized and Federated Learning, SP'19

[Feldman] Does Learning Require **Memorization**?A Short Tale about a Long Tail, STOC'20

Decentralized (Federated) Learning



[Shokri and Shmatikov] Privacy-Preserving Deep Learning, CCS'15

[Nasr, Shokri, Houmansadr] Comprehensive Privacy Analysis of Deep Learning: Passive and Active White-box Inference Attacks against Centralized and Federated Learning, SP'19

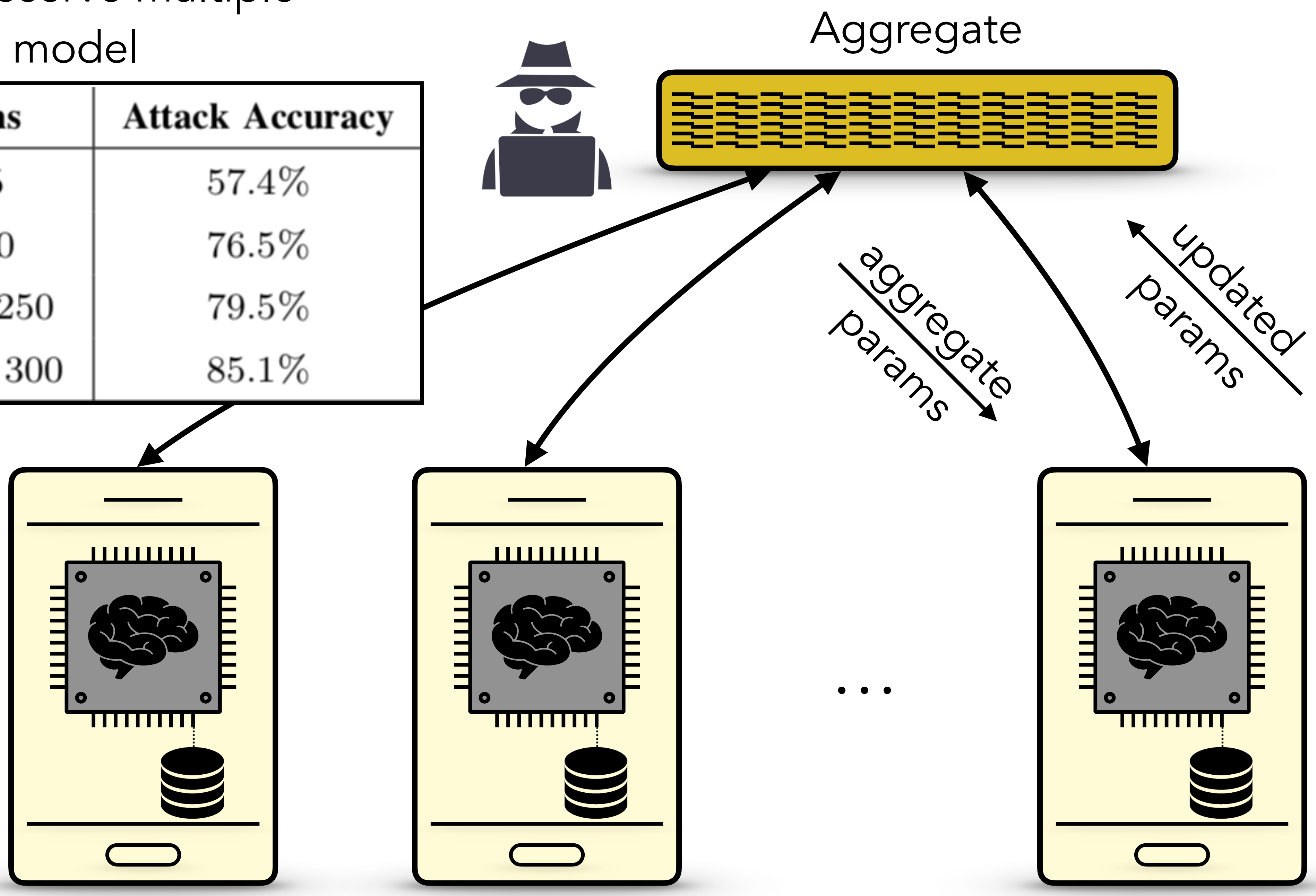
[Melis, Song, De Cristofaro, Shmatikov] Exploiting Unintended Feature Leakage in Collaborative Learning, SP'19

Decentralized (Federated) Learning

Adversary can observe multiple snapshots of the model

Observed Epochs	Attack Accuracy
5, 10, 15, 20, 25	57.4%
10, 20, 30, 40, 50	76.5%
50, 100, 150, 200, 250	79.5%
100, 150, 200, 250, 300	85.1%

CIFAR100-Alexnet



[Shokri and Shmatikov] Privacy-Preserving Deep Learning, CCS'15

[Nasr, Shokri, Houmansadr] Comprehensive Privacy Analysis of Deep Learning: Passive and Active White-box Inference Attacks against Centralized and Federated Learning, SP'19

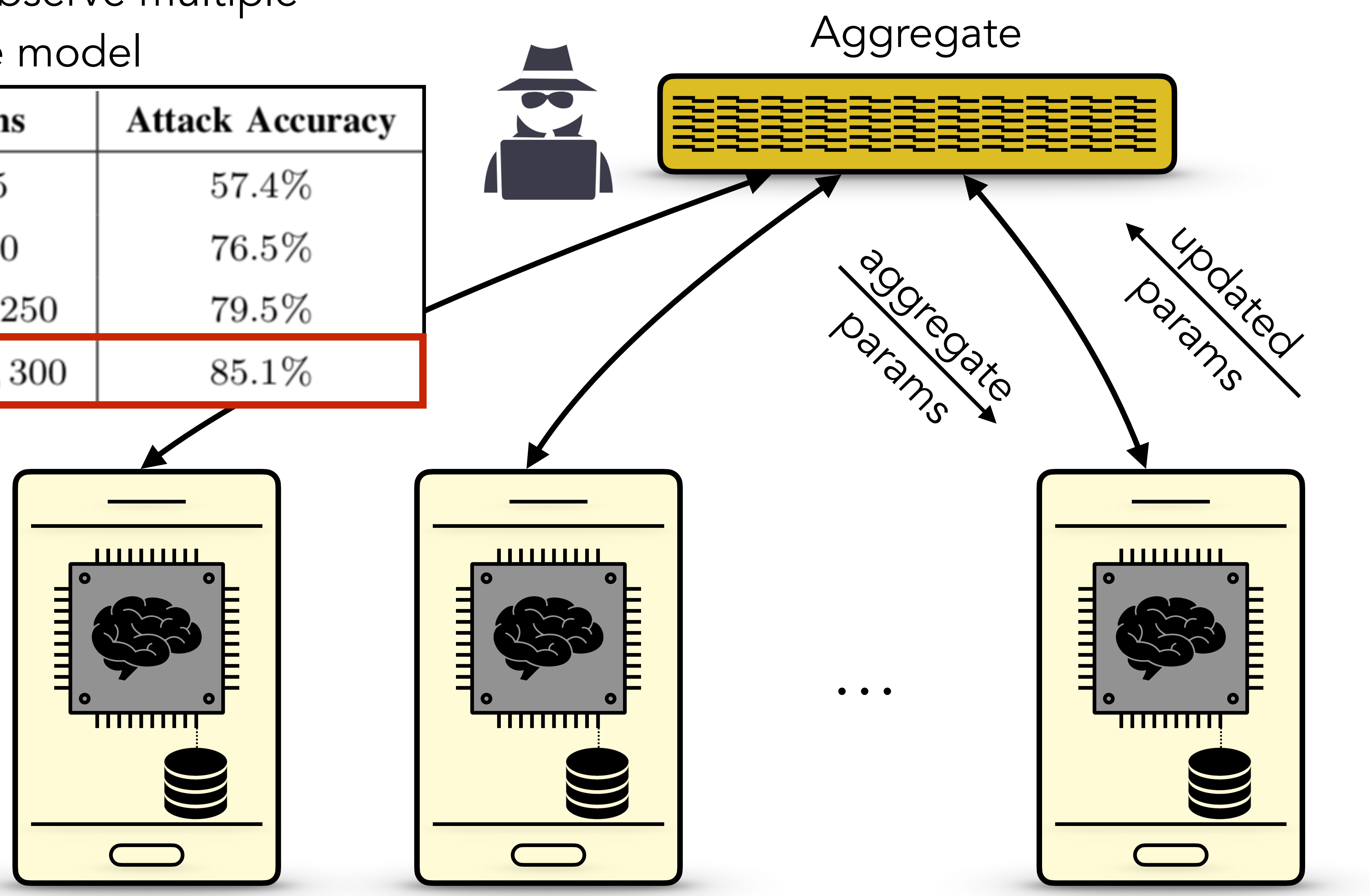
[Melis, Song, De Cristofaro, Shmatikov] Exploiting Unintended Feature Leakage in Collaborative Learning, SP'19

Decentralized (Federated) Learning

Adversary can observe multiple snapshots of the model

Observed Epochs	Attack Accuracy
5, 10, 15, 20, 25	57.4%
10, 20, 30, 40, 50	76.5%
50, 100, 150, 200, 250	79.5%
100, 150, 200, 250, 300	85.1%

CIFAR100-Alexnet

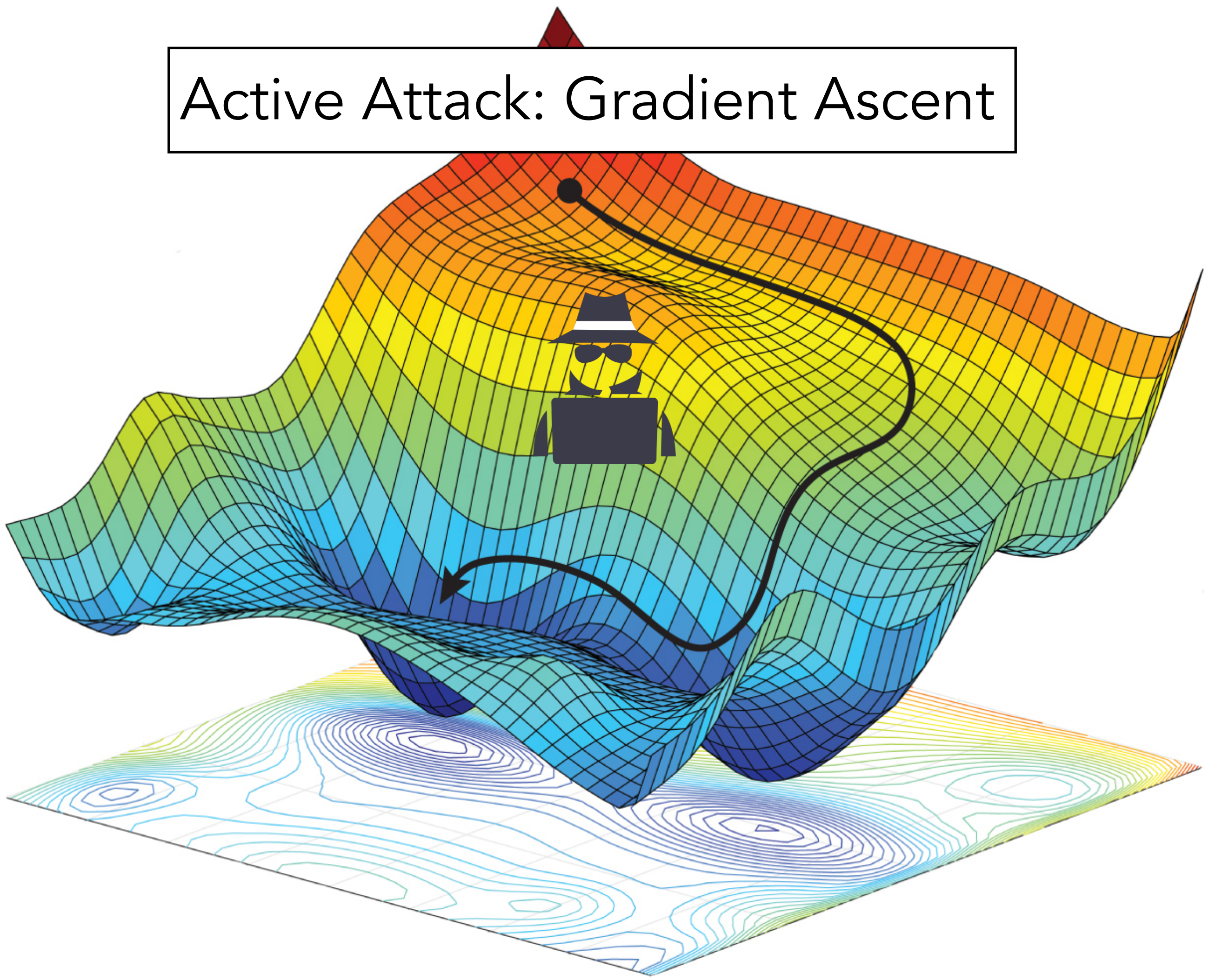


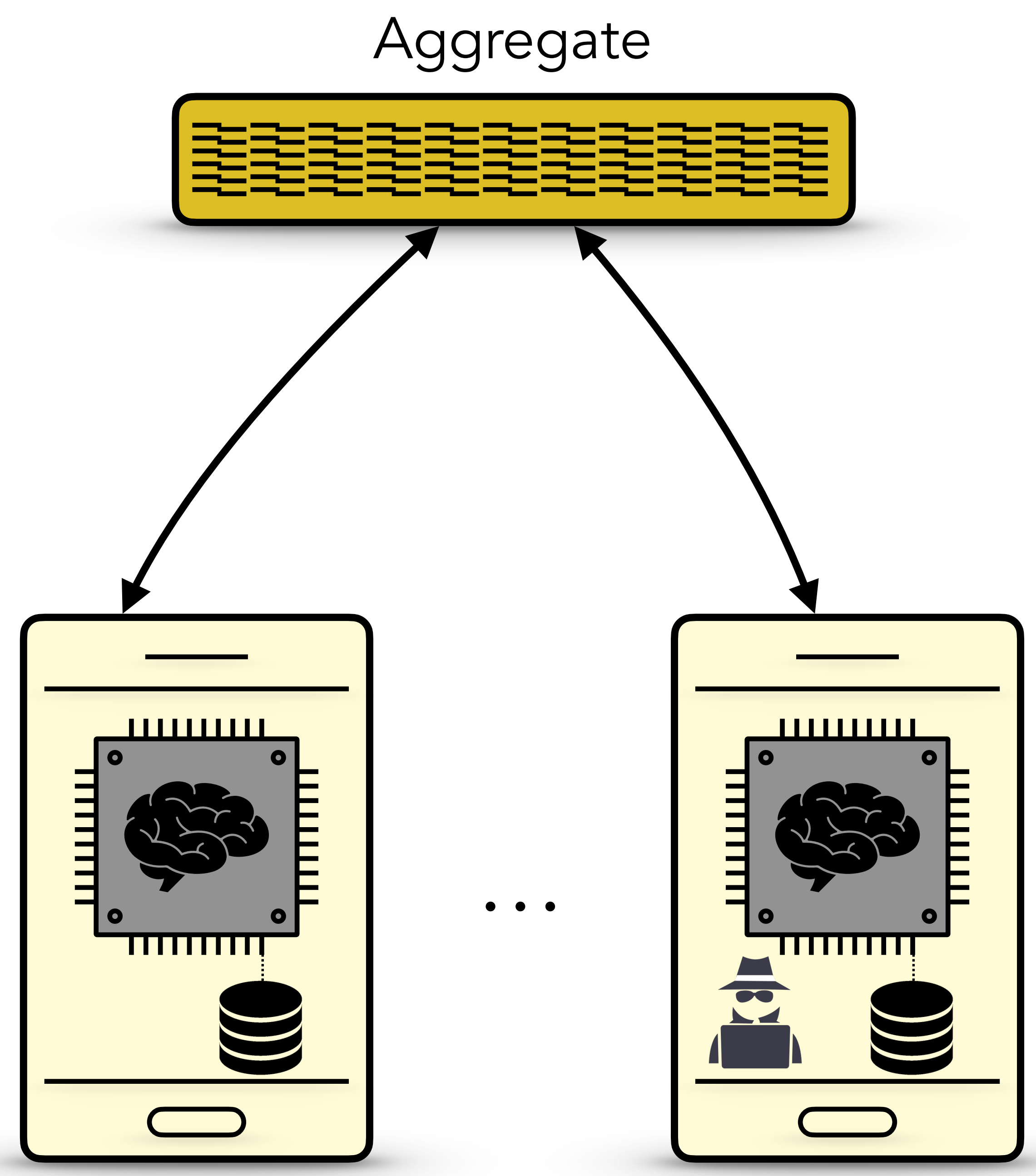
[Shokri and Shmatikov] Privacy-Preserving Deep Learning, CCS'15

[Nasr, Shokri, Houmansadr] Comprehensive Privacy Analysis of Deep Learning: Passive and Active White-box Inference Attacks against Centralized and Federated Learning, SP'19

[Melis, Song, De Cristofaro, Shmatikov] Exploiting Unintended Feature Leakage in Collaborative Learning, SP'19

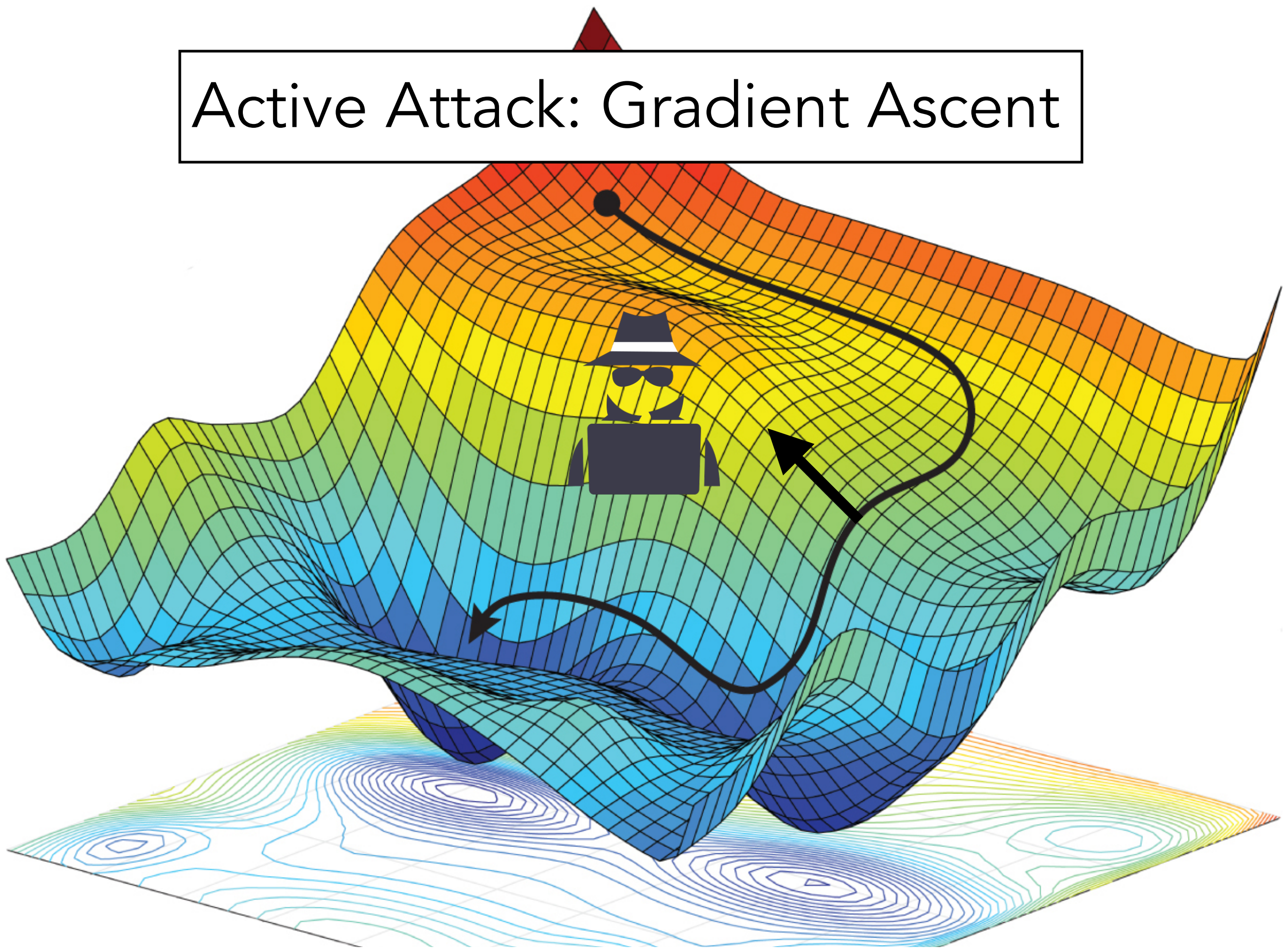
Decentralized (Federated) Learning



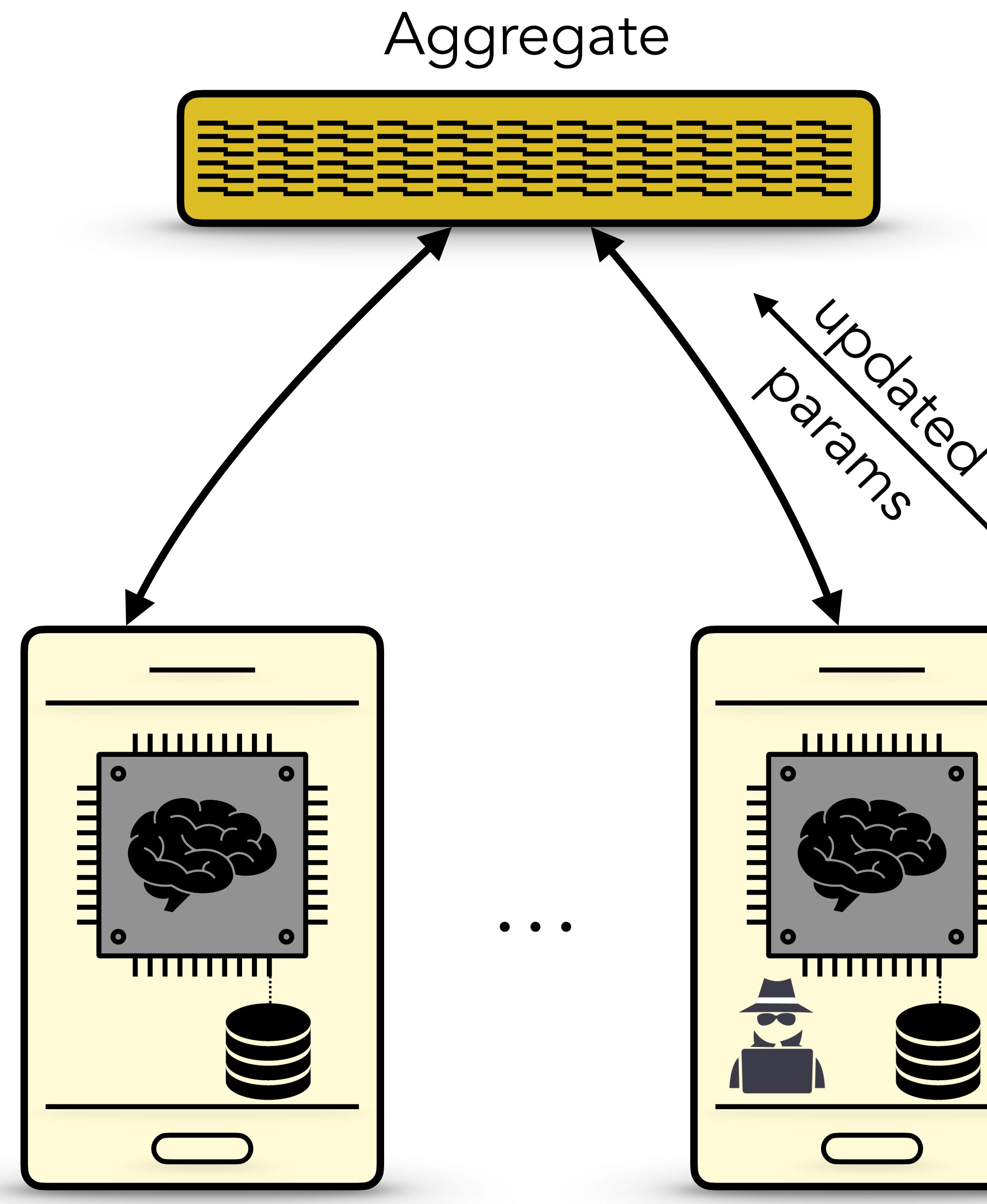


[Nasr, Shokri, Houmansadr] Comprehensive Privacy Analysis of Deep Learning: Passive and Active White-box Inference Attacks against Centralized and Federated Learning, SP'19

Decentralized (Federated) Learning



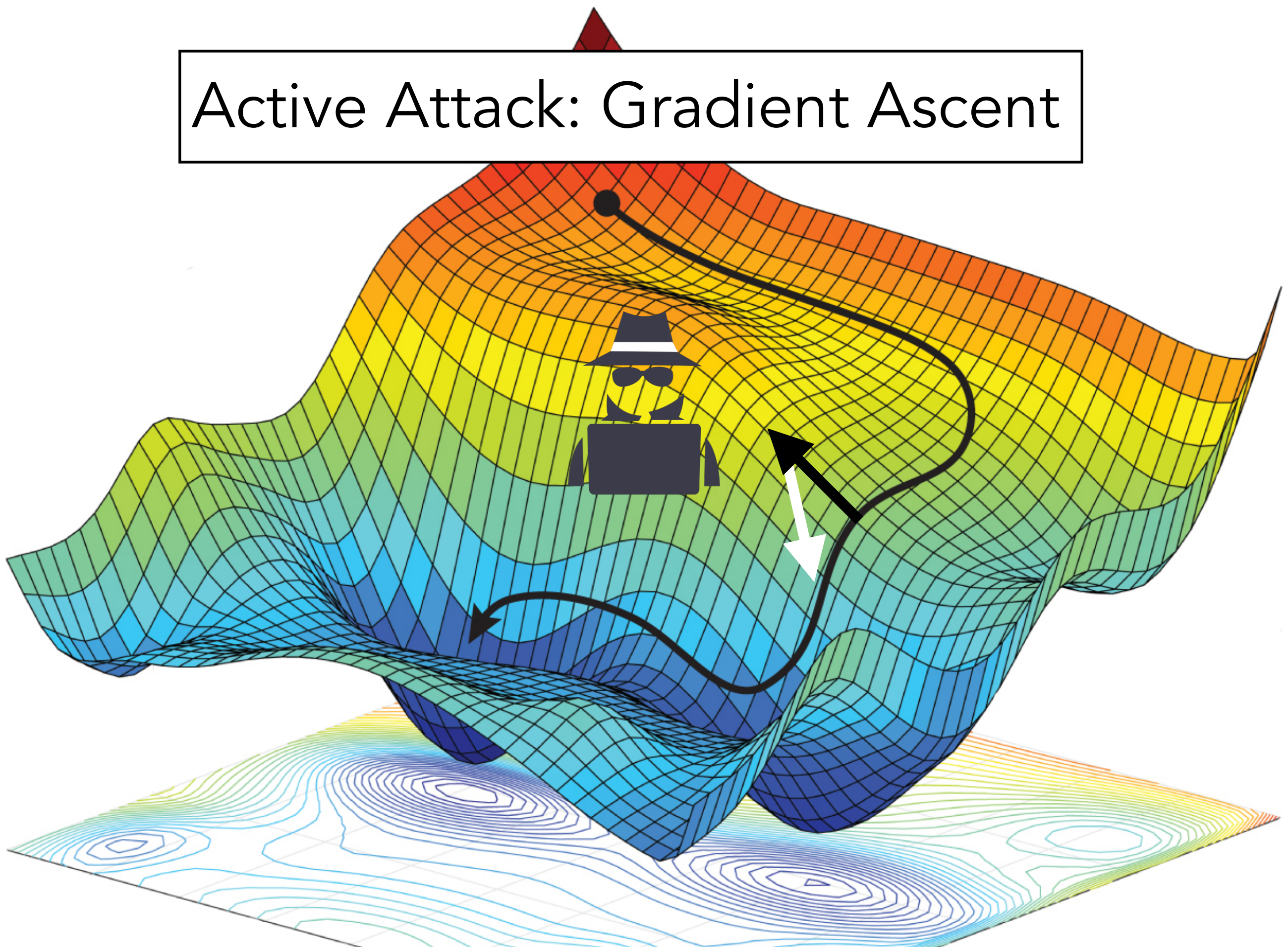
Increase loss on a particular data point x.



[Nasr, Shokri, Houmansadr] Comprehensive Privacy Analysis of Deep Learning: Passive and Active White-box Inference Attacks against Centralized and Federated Learning, SP'19

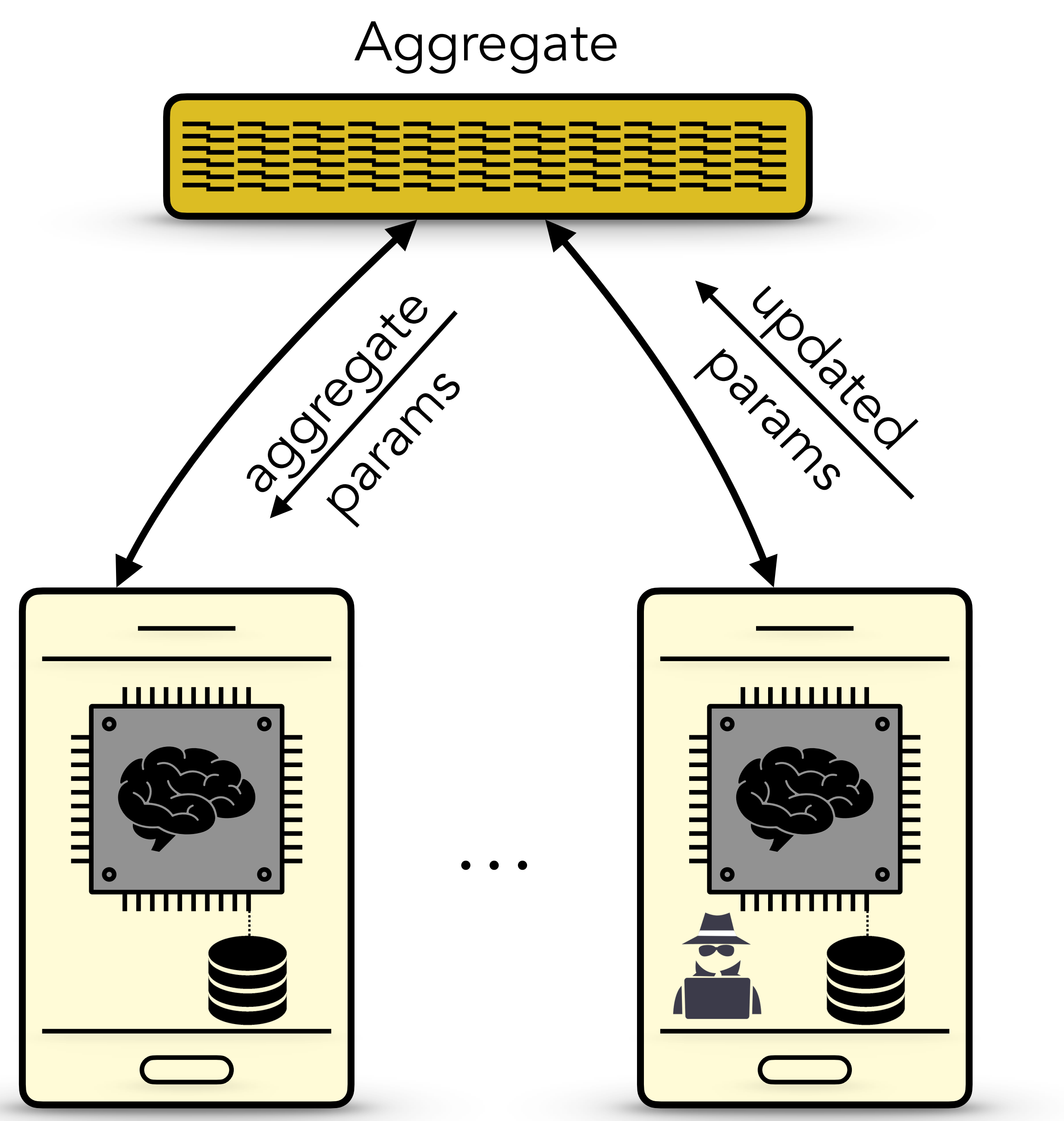
Decentralized (Federated) Learning

Active Attack: Gradient Ascent



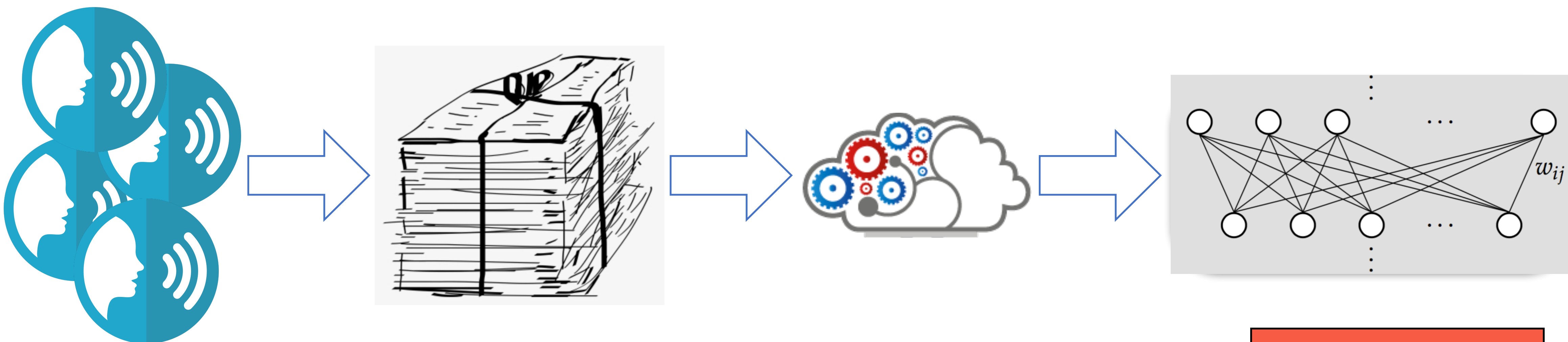
Increase loss on a particular data point x.

A participant correct it back (by running gradient descent locally) only if x is part of its training set. => **membership leakage**



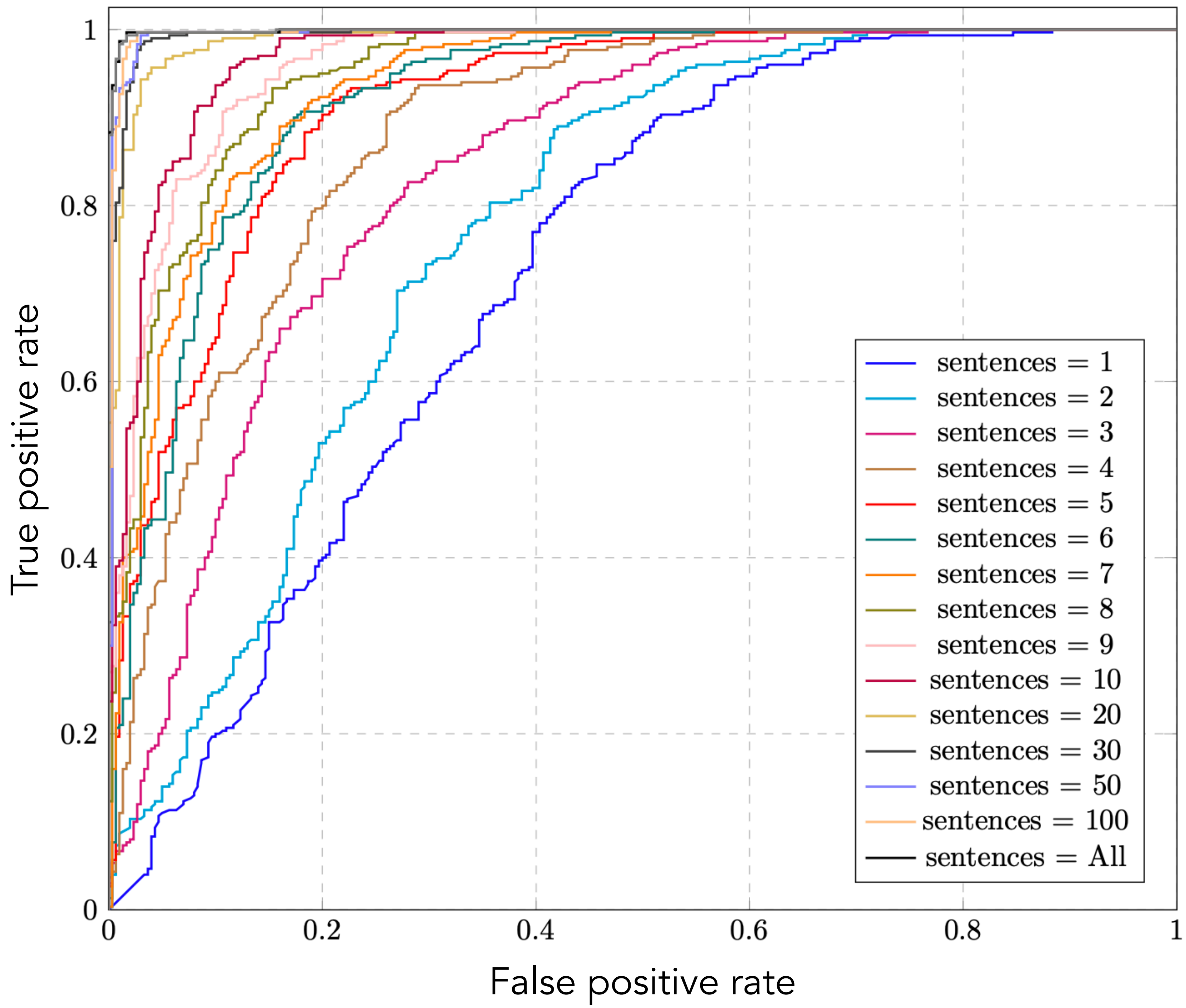
[Nasr, Shokri, Houmansadr] Comprehensive Privacy Analysis of Deep Learning: Passive and Active White-box Inference Attacks against Centralized and Federated Learning, SP'19

NLP Models



- How much does the model leak about the sentences of a particular author/speaker? What about the membership of the author in the training set (based on known samples)?
- Which samples are leaked?

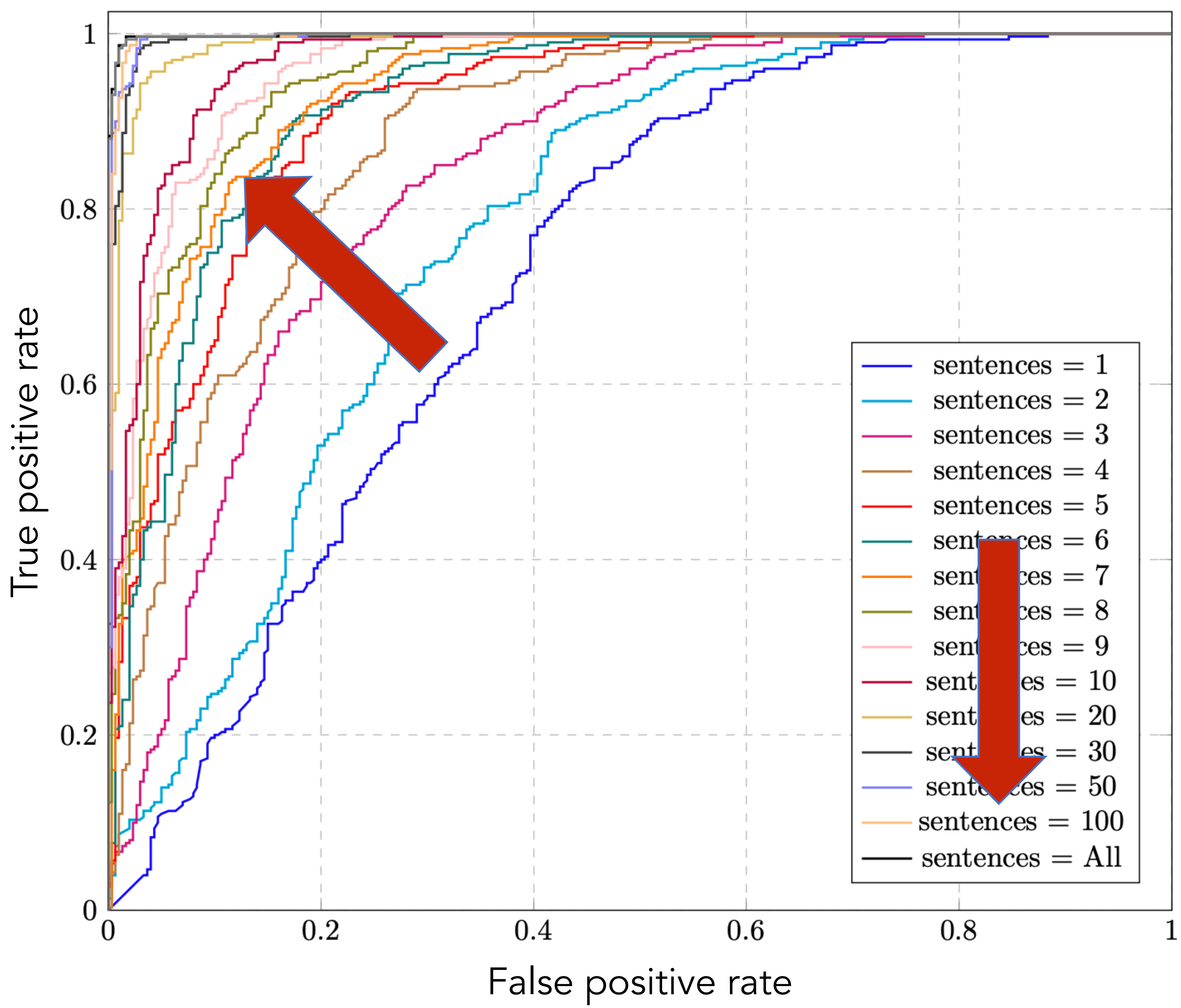
Membership Inference



SATED (Speaker Annotated TED talks) dataset

[Maddi] https://github.com/privacytrustlab/ml_privacy_meter based on [Song, Shmatikov] Auditing Data Provenance in Text-Generation Models, KDD'19

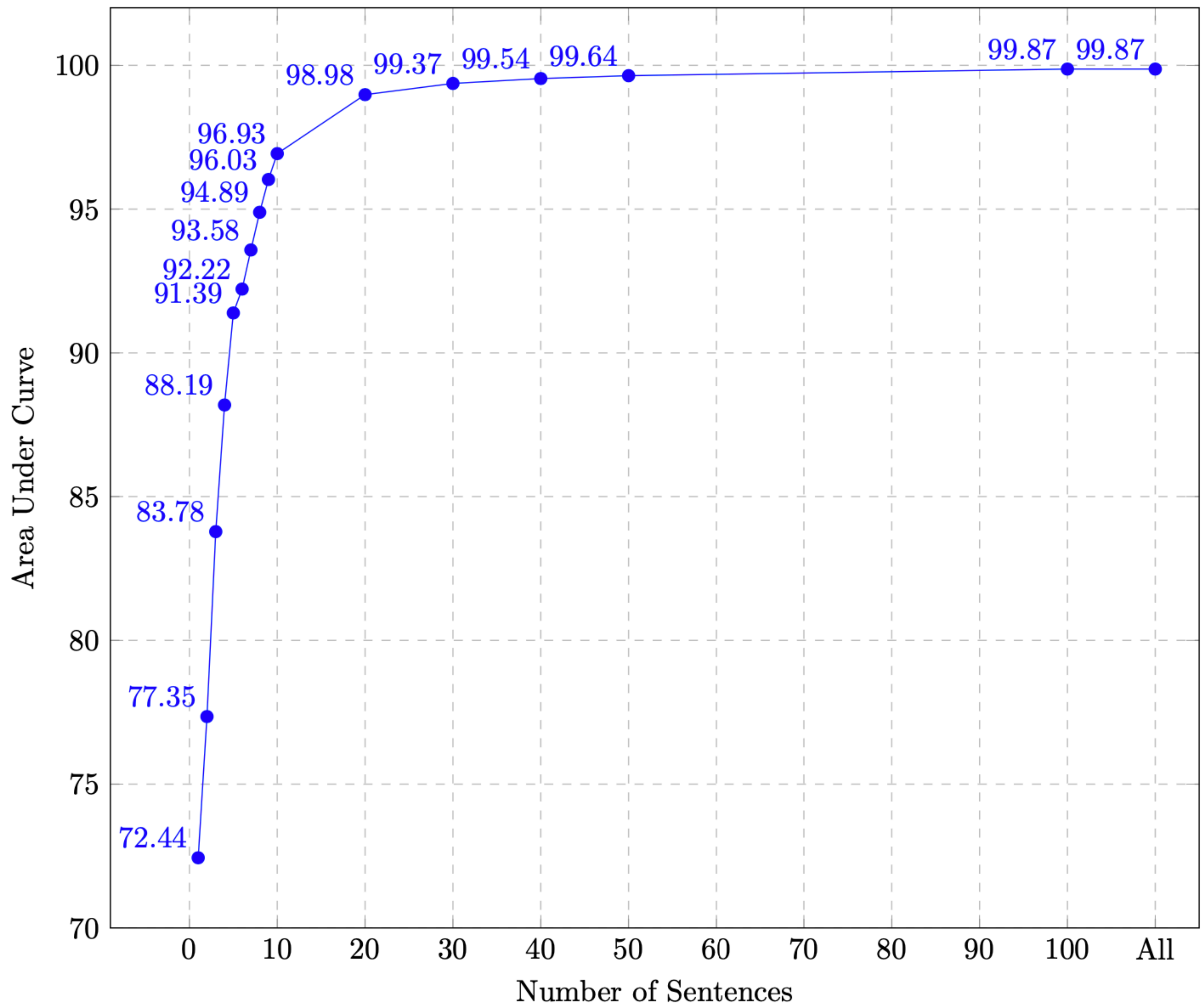
Membership Inference



SATED (Speaker Annotated TED talks) dataset

[Maddi] https://github.com/privacytrustlab/ml_privacy_meter based on [Song, Shmatikov] Auditing Data Provenance in Text-Generation Models, KDD'19

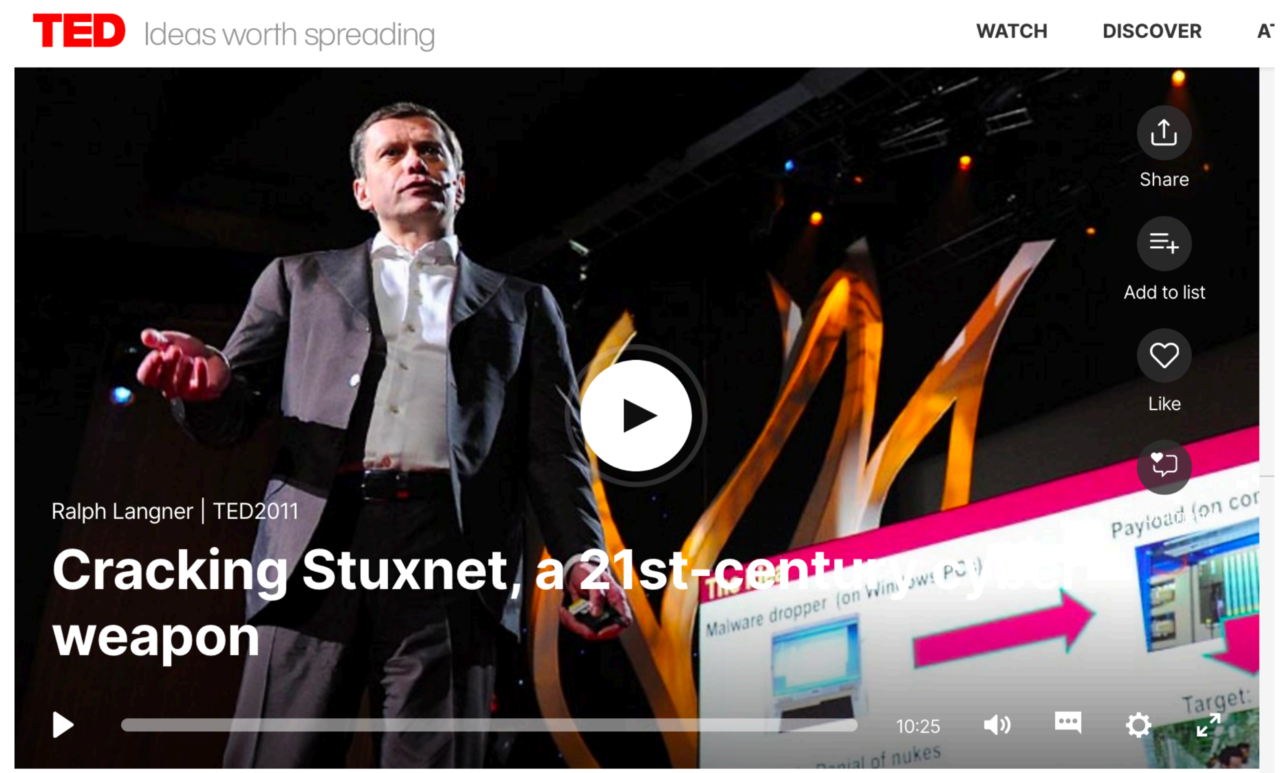
Membership Inference



SATED (Speaker Annotated TED talks) dataset

[Maddi] https://github.com/privacytrustlab/ml_privacy_meter based on [Song, Shmatikov] Auditing Data Provenance in Text-Generation Models, KDD'19

Examples of Vulnerable Training Data



But it gets worse. And this is very important, what I'm going to say. Think about this: this attack is generic. It doesn't have anything to do, in specifics, with centrifuges, with uranium enrichment. So it would work as well, for example, in a power plant or in an automobile factory. It is generic. And you don't have -- as an attacker -- you don't have to deliver this payload by a USB stick, as we saw it in the case of Stuxnet. You could also use conventional worm technology for spreading. Just spread it as

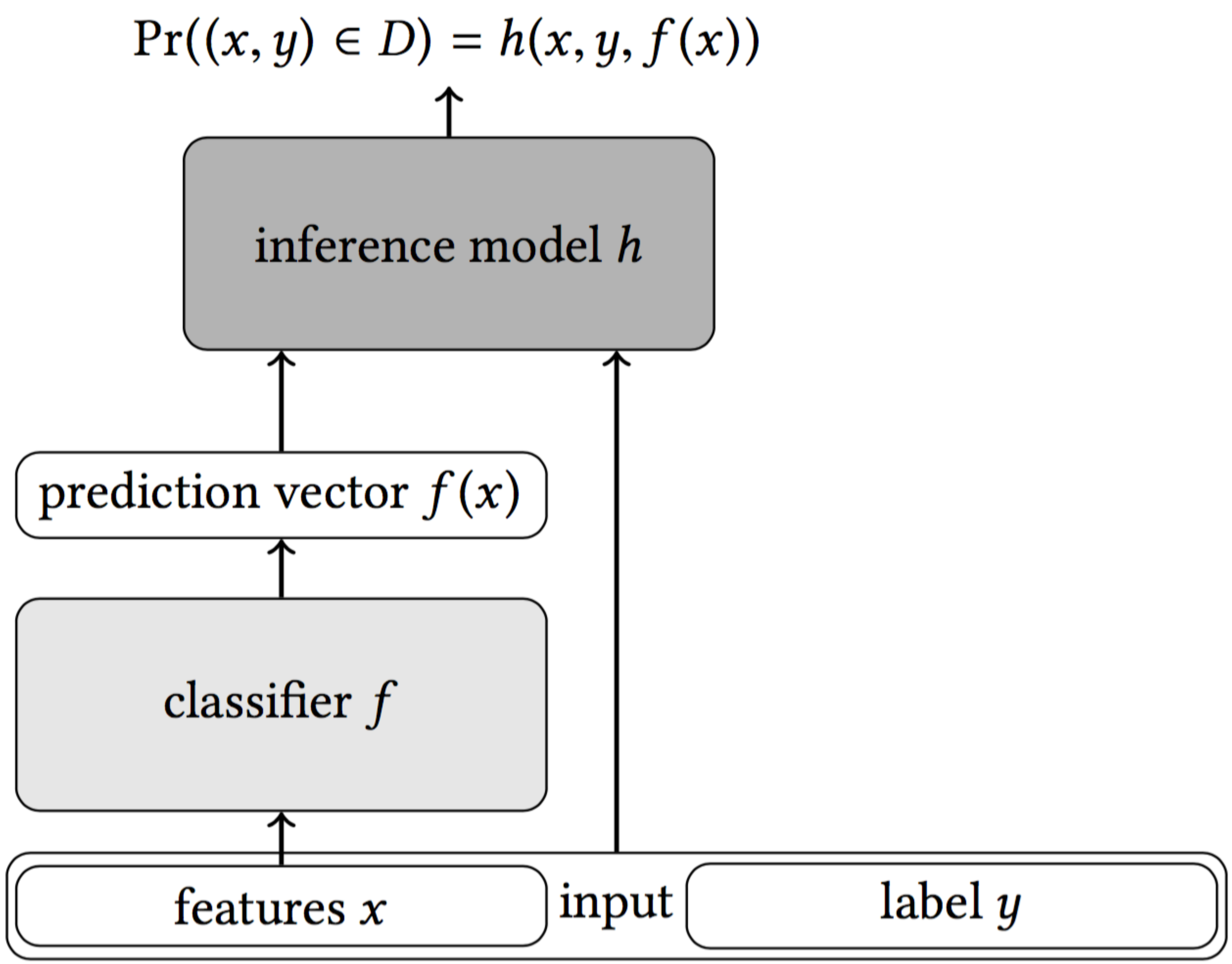
Chris Anderson: I've got a question. Ralph, it's been quite widely reported that people assume that Mossad is the main entity behind this. Is that your opinion?

Ralph Langner: Okay, you really want to hear that? Yeah. Okay. My opinion is that the Mossad is involved, but that the leading force is not Israel. So the leading force behind that is the cyber superpower. There is only one, and that's the United States -- fortunately, fortunately. Because otherwise, our problems would even be bigger.

Examples of Vulnerable Training Data

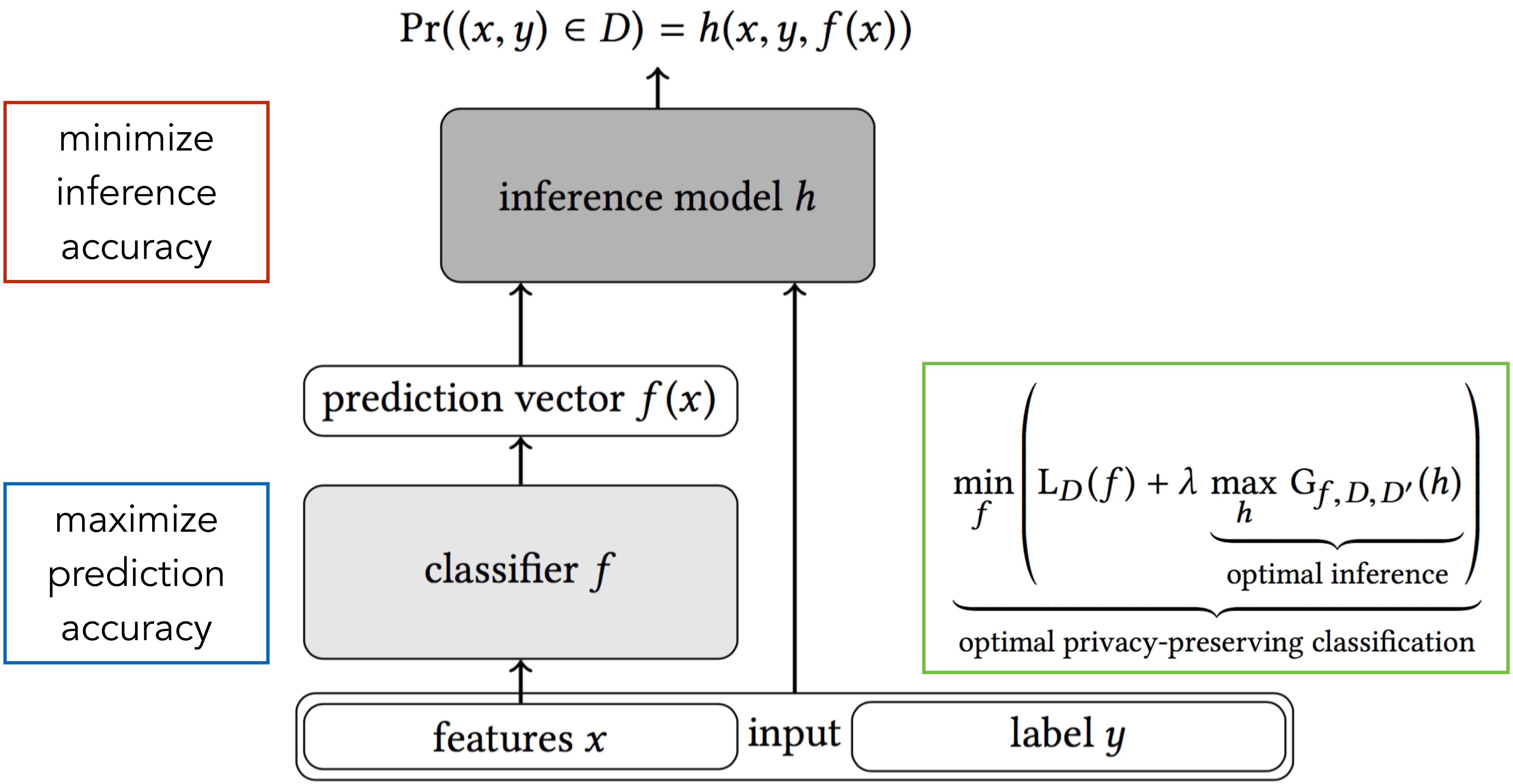
This year, Germany is celebrating the 25th anniversary of the peaceful revolution in East Germany. In 1989, the Communist regime was moved away, the Berlin Wall came down, and one year later, the German Democratic Republic, the GDR, in the East was unified with the Federal Republic of Germany in the West to found today's Germany. Among many other things, Germany inherited the archives of the East German secret police, known as the Stasi. Only two years after its dissolution, its documents were opened to the public, and historians such as me started to study these documents to learn more about how the GDR surveillance state functioned.

ML Privacy Meter as a Privacy Regularizer



[Nasr, Shokri, Houmansadr] Machine Learning with Membership Privacy using Adversarial Regularization, CCS'18

ML Privacy Meter as a Privacy Regularizer

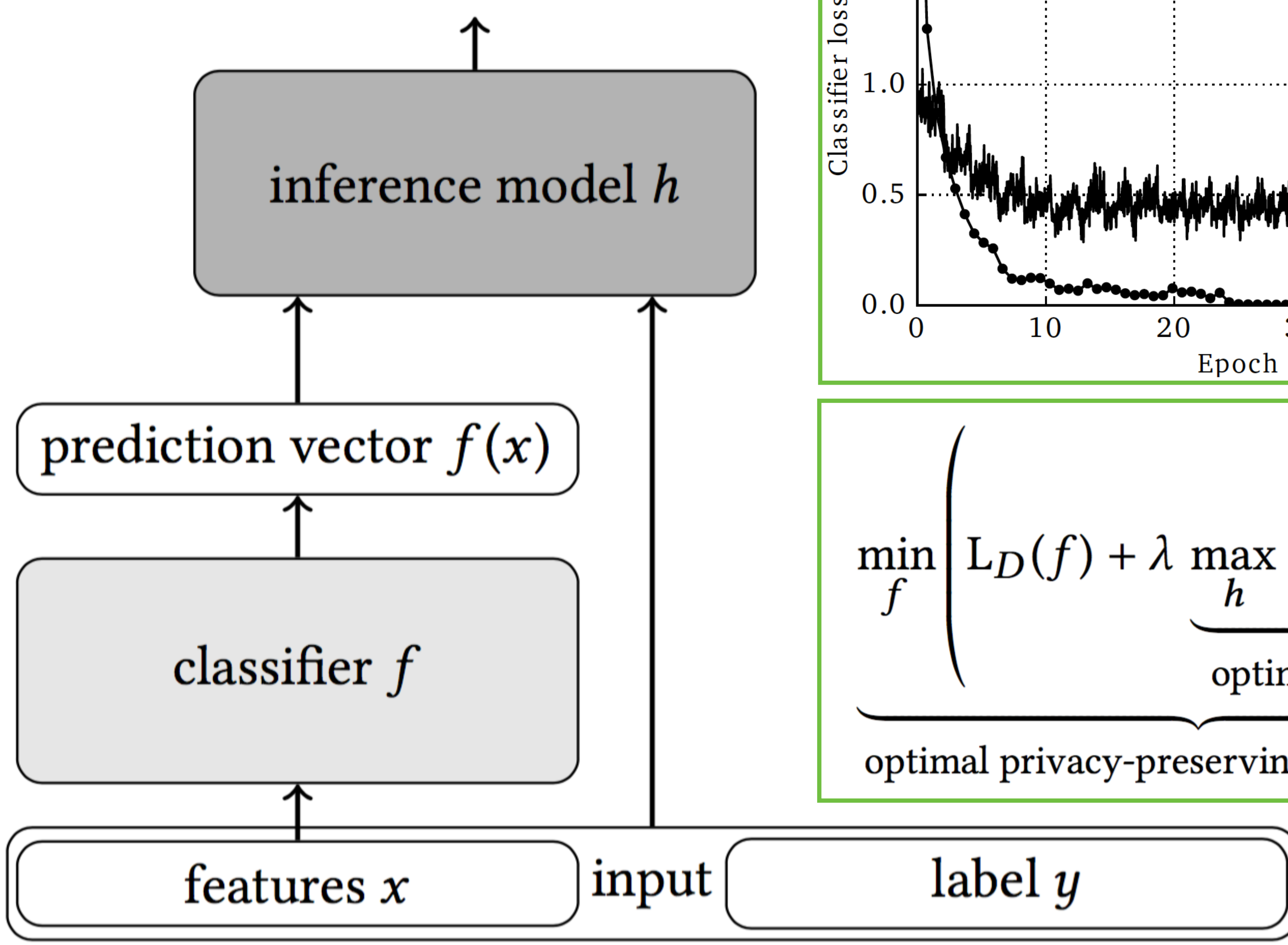


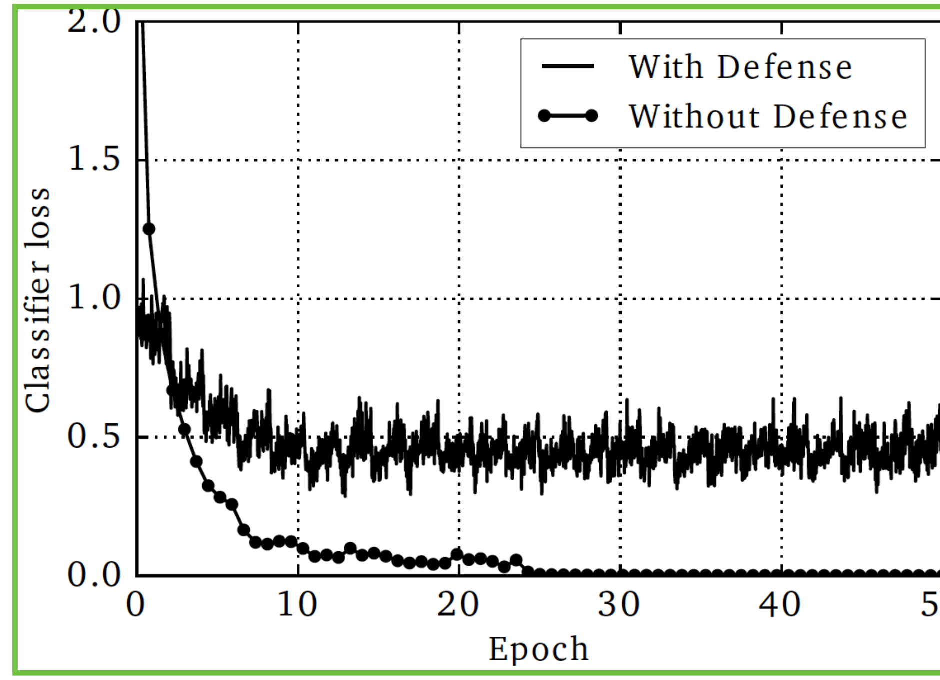
[Nasr, Shokri, Houmansadr] Machine Learning with Membership Privacy using Adversarial Regularization, CCS'18

ML Privacy Meter as a Privacy Regularizer

minimize inference accuracy

maximize prediction accuracy





$$\underbrace{\min_{f} \left(L_D(f) + \lambda \underbrace{\max_{h} G_{f,D,D'}(h)}_{\text{optimal inference}} \right)}_{\text{optimal privacy-preserving classification}}$$

[Nasr, Shokri, Houmansadr] Machine Learning with Membership Privacy using Adversarial Regularization, CCS’18

Privacy and Generalization

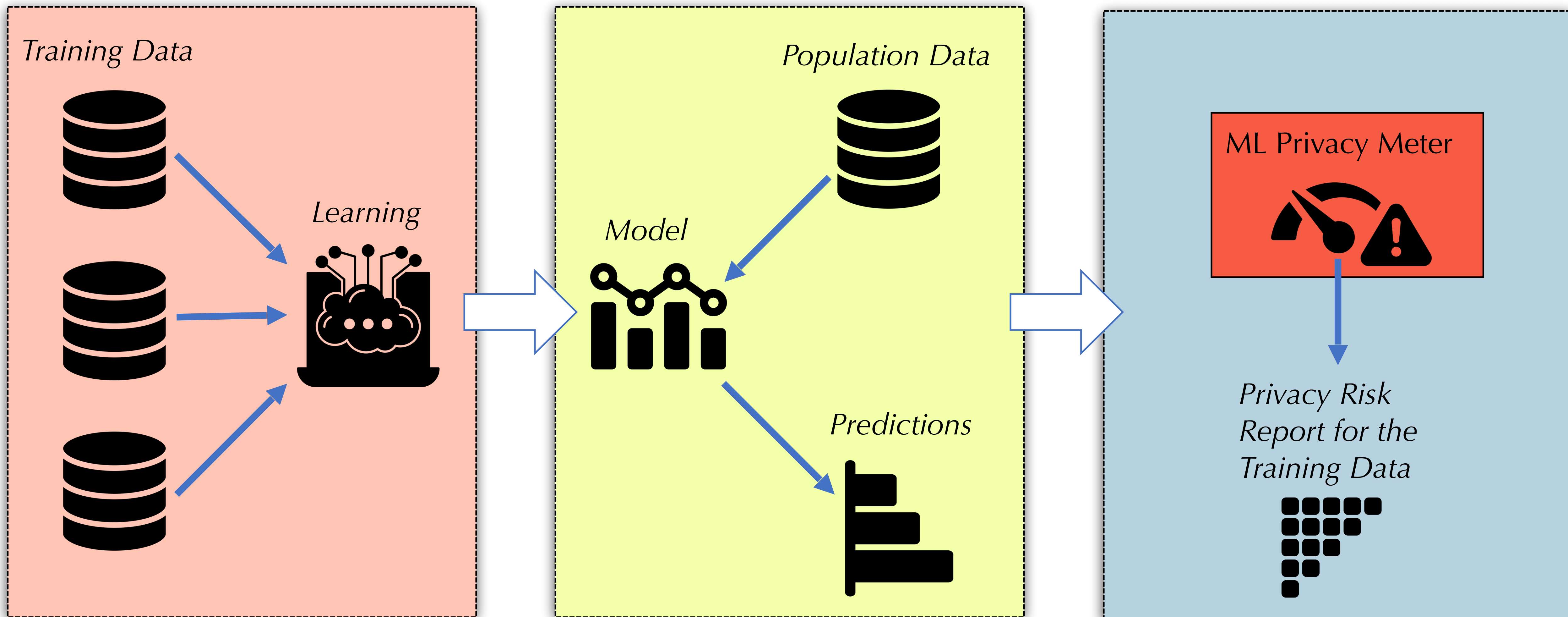
	Without defense			With defense		
Dataset	Training accuracy	Testing accuracy	Attack accuracy	Training accuracy	Testing accuracy	Attack accuracy
Purchase100	100%	80.1%	67.6%	92.2%	76.5%	51.6%
Texas100	81.6%	51.9%	63%	55%	47.5%	51.0%
CIFAR100- Alexnet	99%	44.7%	53.2%	66.3%	43.6%	50.7%
CIFAR100- DenseNET	100%	70.6%	54.5%	80.3%	67.6%	51.0%

Smaller gap

Random guess

[Nasr, Shokri, Houmansadr] Machine Learning with Membership Privacy using Adversarial Regularization, CCS'18

Tool: ML Privacy Meter



ML Privacy Meter is a Python library (`ml_privacy_meter`) that enables quantifying the privacy risks of machine learning models. https://github.com/privacytrustlab/ml_privacy_meter

[Murakonda, Shokri] ML Privacy Meter: Aiding Regulatory Compliance by Quantifying the Privacy Risks of Machine Learning, HotPETs 2020