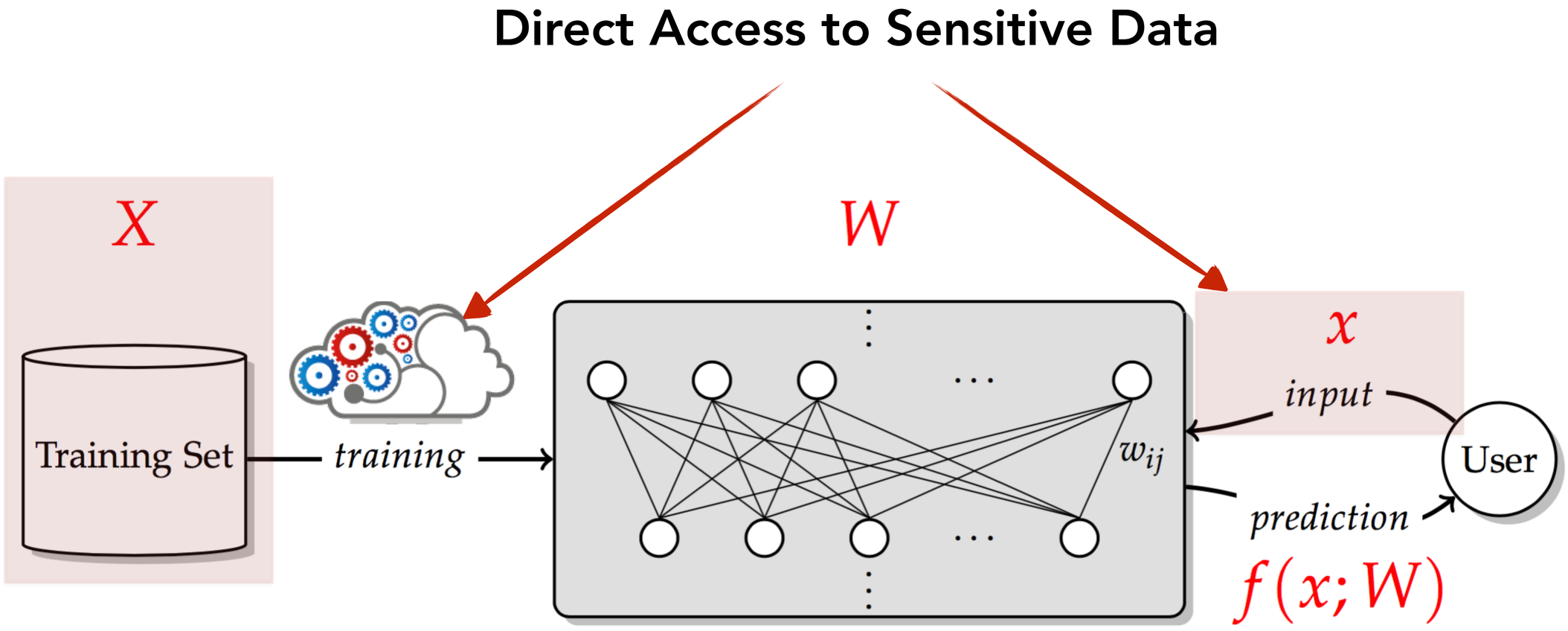
Privacy in Language Models

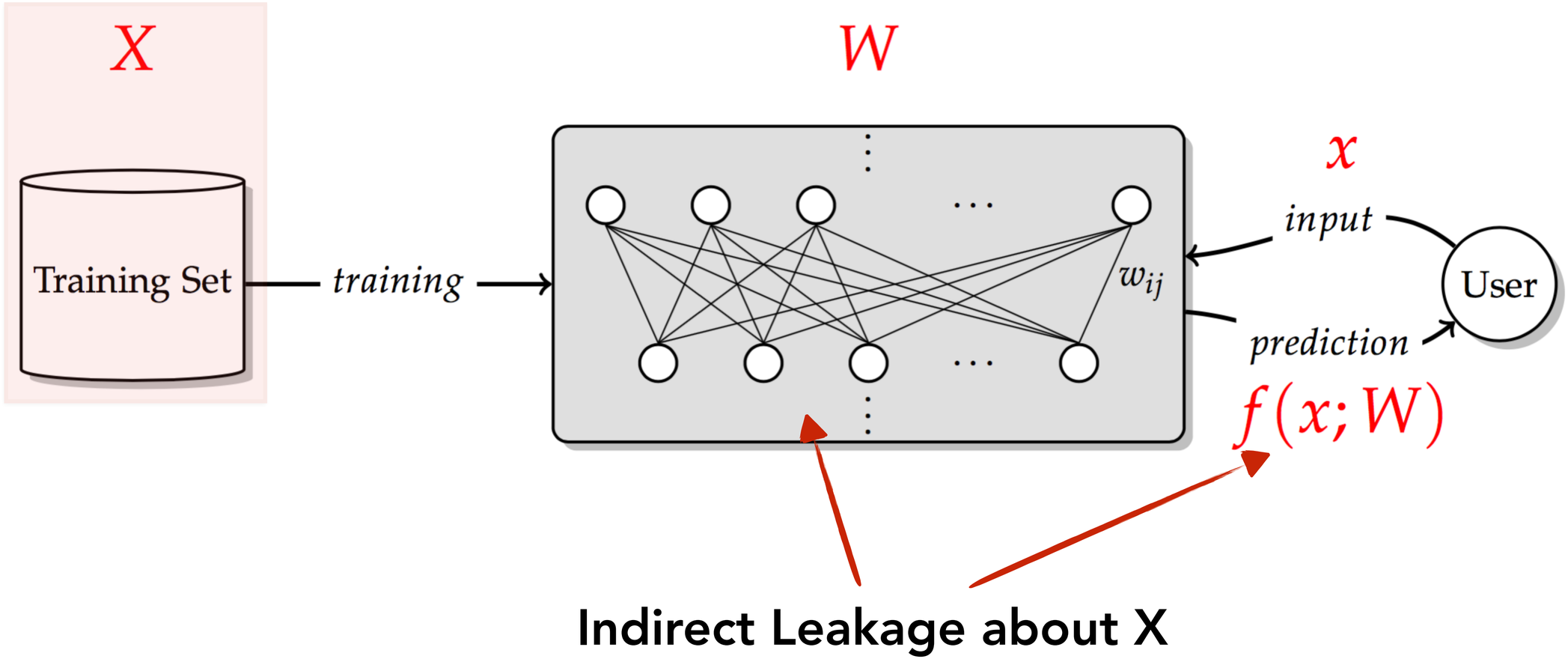
Reza Shokri

Data Privacy and Trustworthy ML Research Lab
National University of Singapore (NUS)

Direct Privacy Risks in Machine Learning



Indirect Privacy Risks in Machine Learning



Language Models

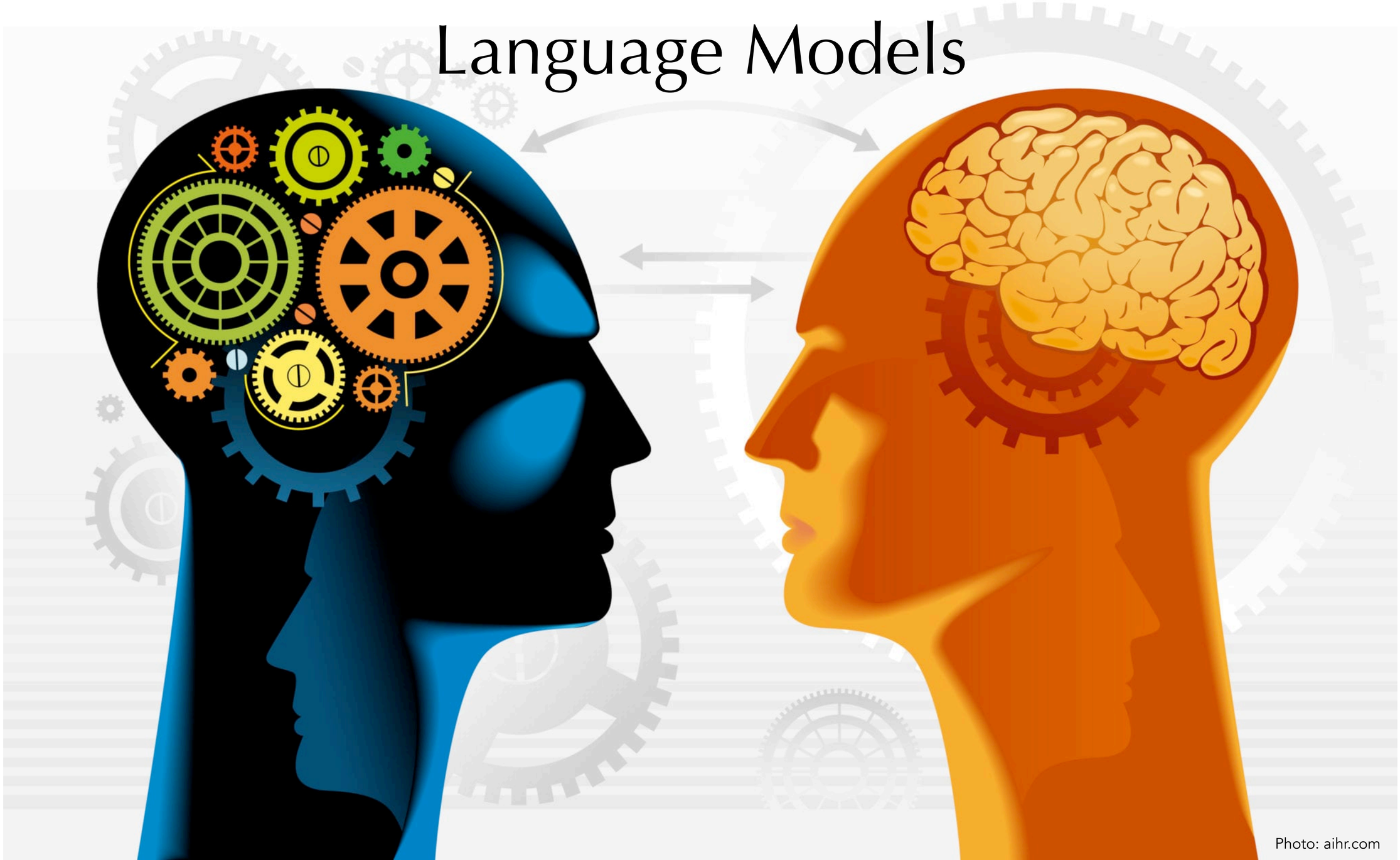


Photo: aihr.com

There is a cat on the



There is a cat on the table.

There is a cat on the floor.

There is a cat on the chair.

There is a cat on the window sill.

There is a cat on the roof.

There is a cat on the counter.

There is a cat on the bed.

There is a cat on the porch.

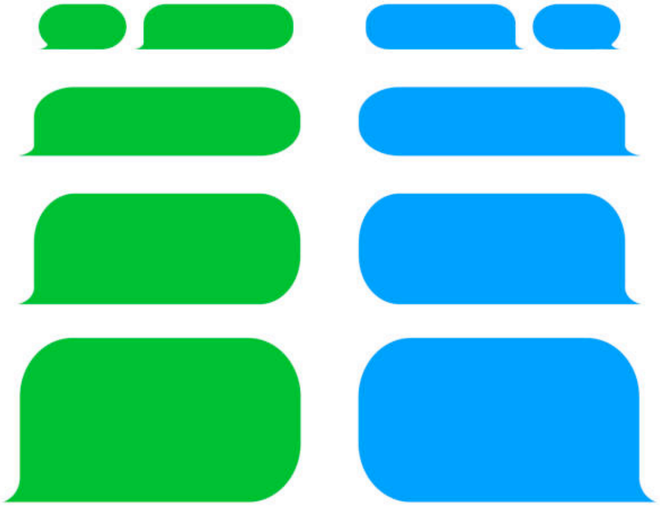
There is a cat on the fence.

There is a cat on the car.

Sources



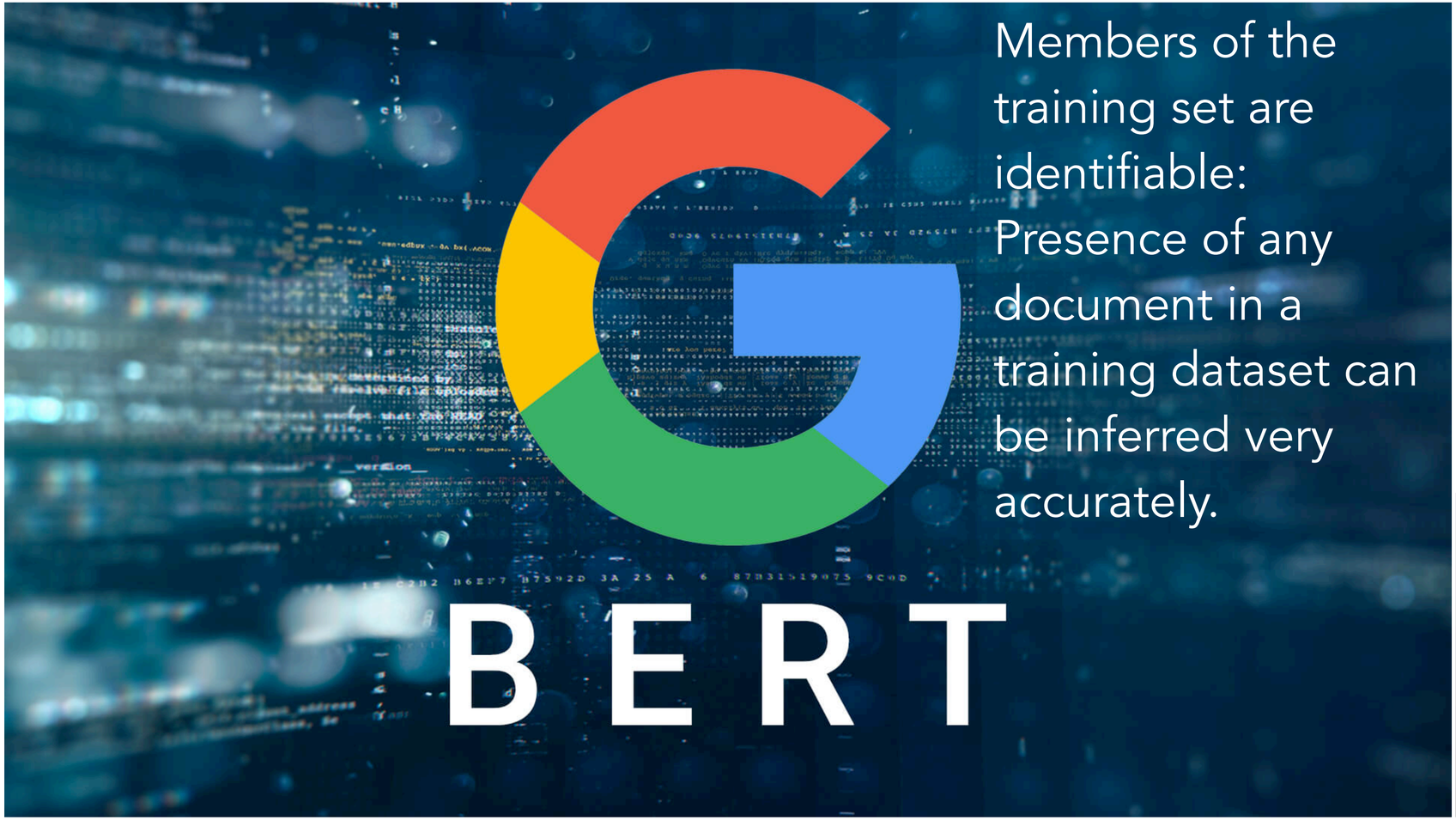






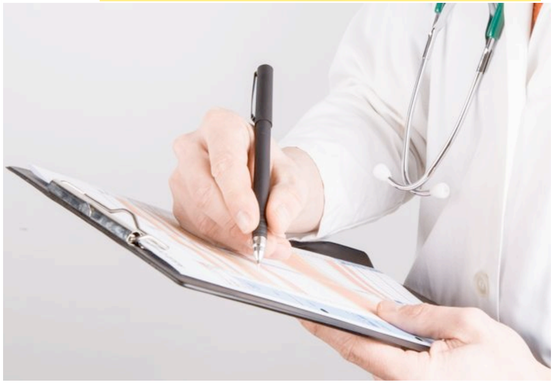
WHEN YOU TRAIN PREDICTIVE MODELS ON INPUT FROM YOUR USERS, IT CAN LEAK INFORMATION IN UNEXPECTED WAYS.

Real World Attacks against Large Language Models



Members of the training set are identifiable: Presence of any document in a training dataset can be inferred very accurately.

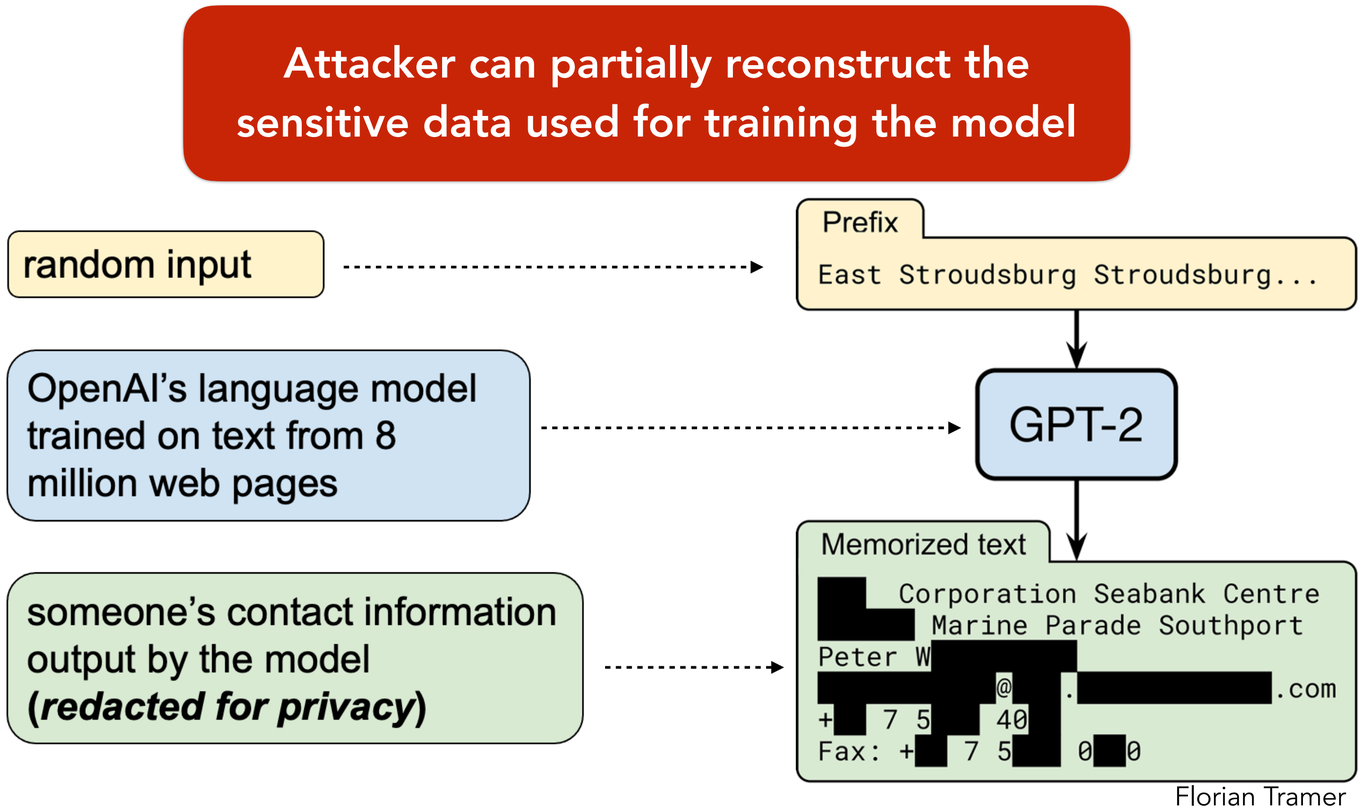




[Mireshghallah, Goyal, Uniyal, Berg-Kirkpatrick, Shokri] Quantifying Privacy Risks of Masked Language Models Using Membership Inference Attacks, EMNLP'22

Real World Attacks against Large Language Models

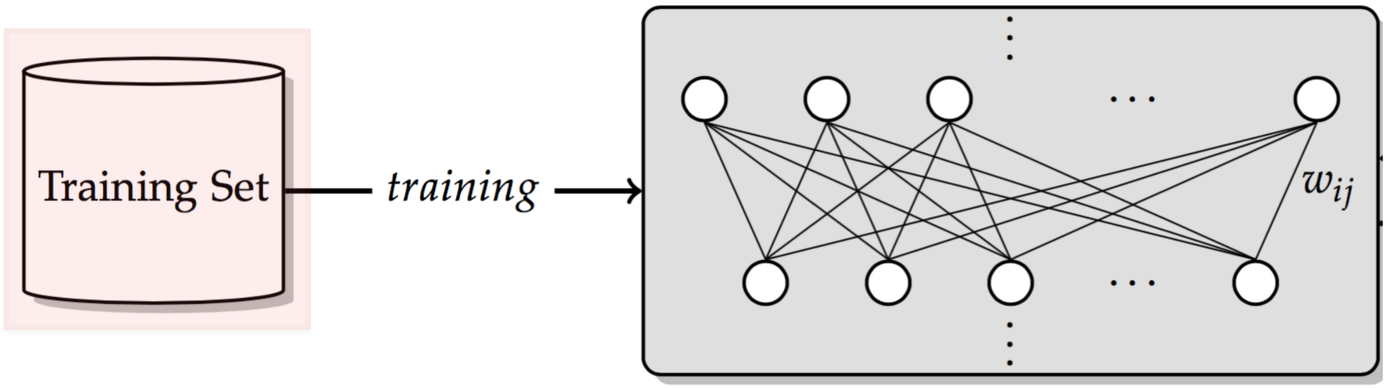
OpenAI

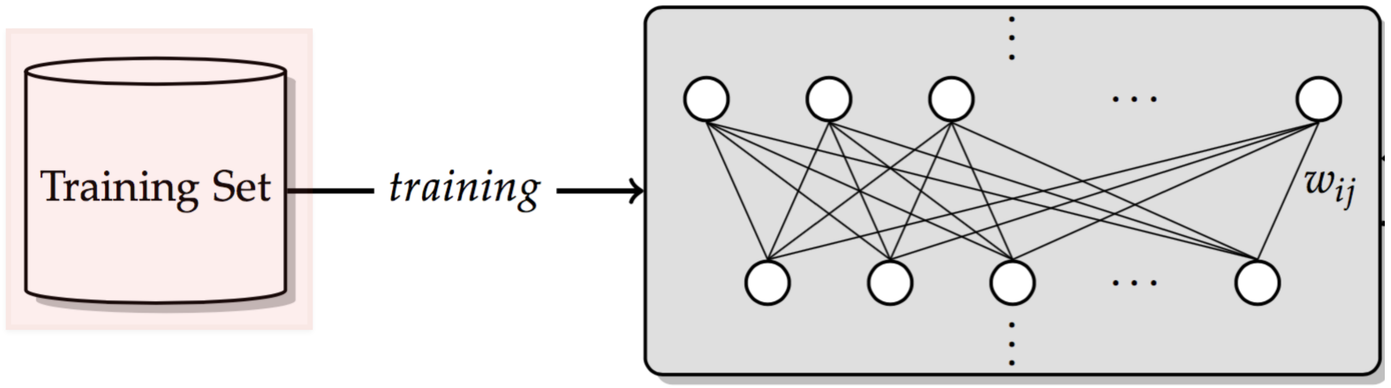


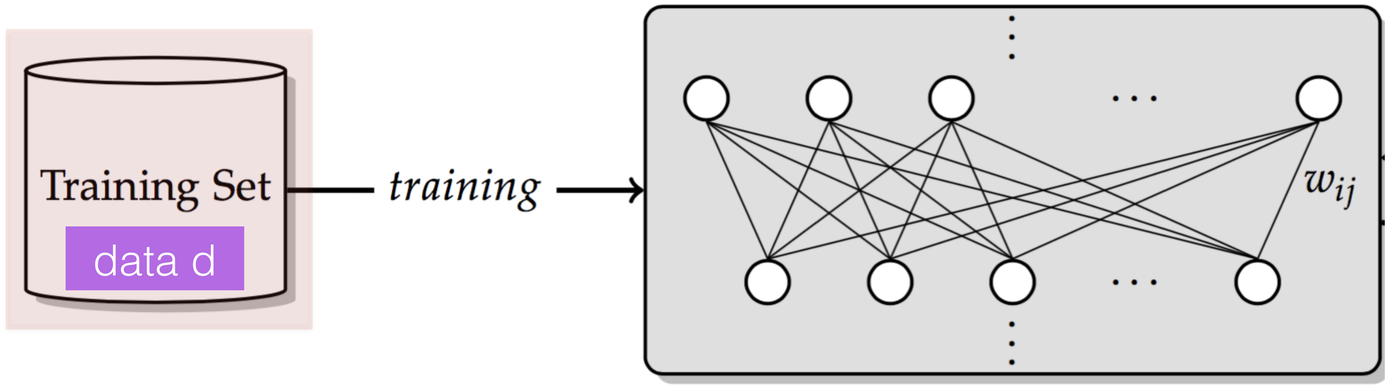
[Carlini, Tramer, et al.] Extracting Training Data from Large Language Models, Usenix security'21

Models are **personal data**

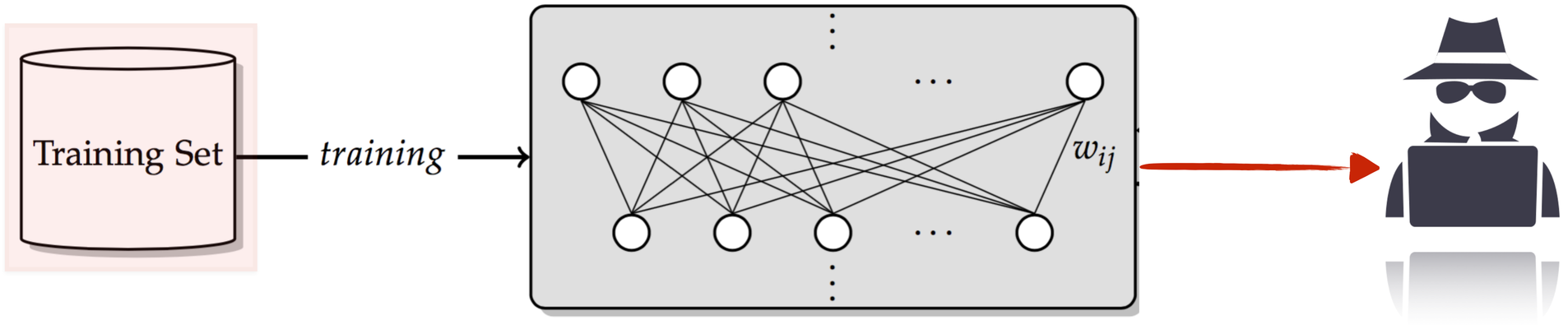
We need a standard method for auditing data privacy in machine learning systems


Training Set
training
w_{ij}

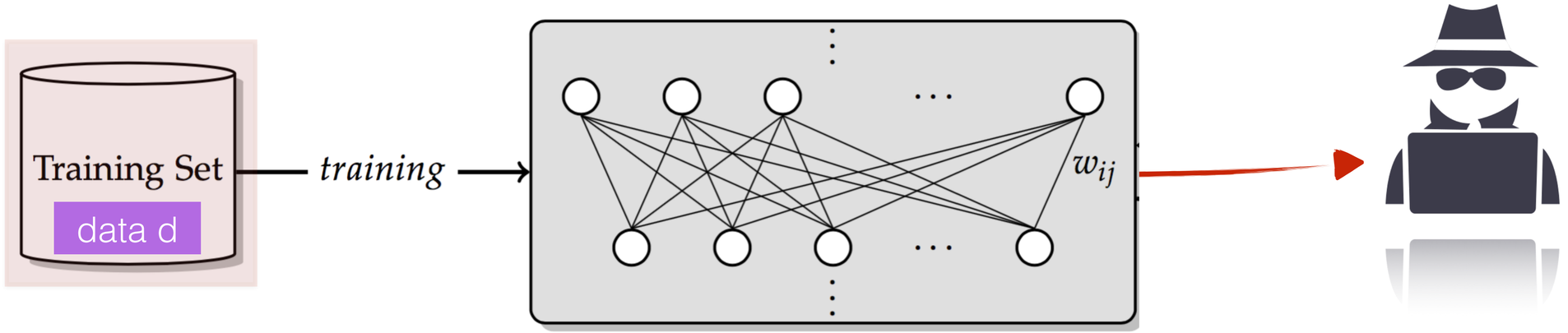

Training Set
training
w_{ij}


Training Set
data d
training
w_{ij}

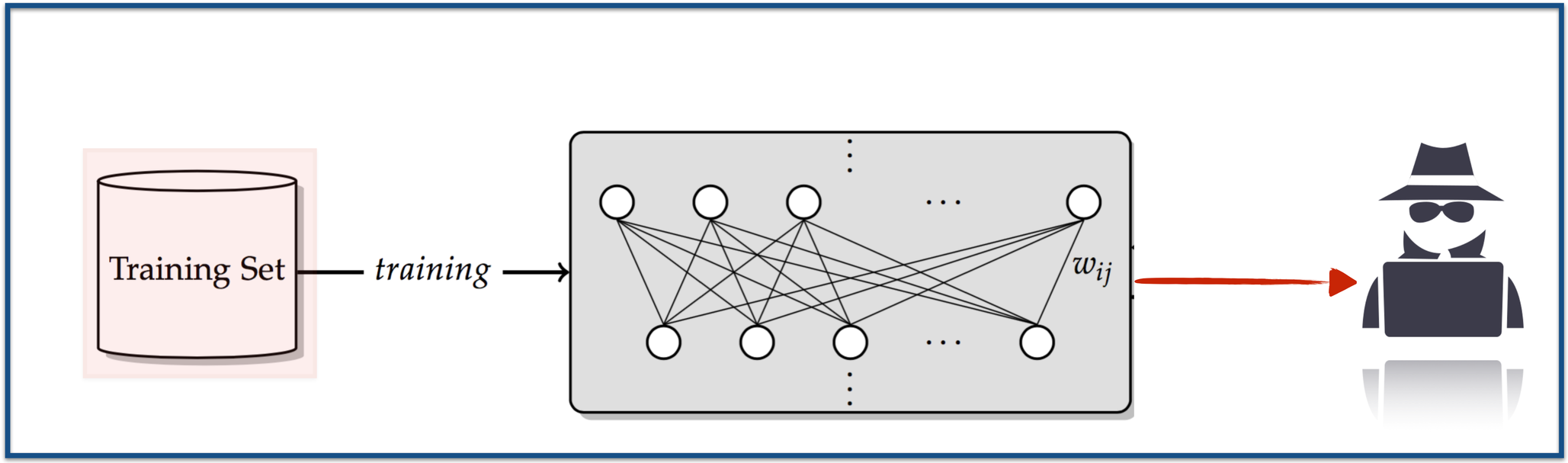
A world in which data d **is not** part of training set



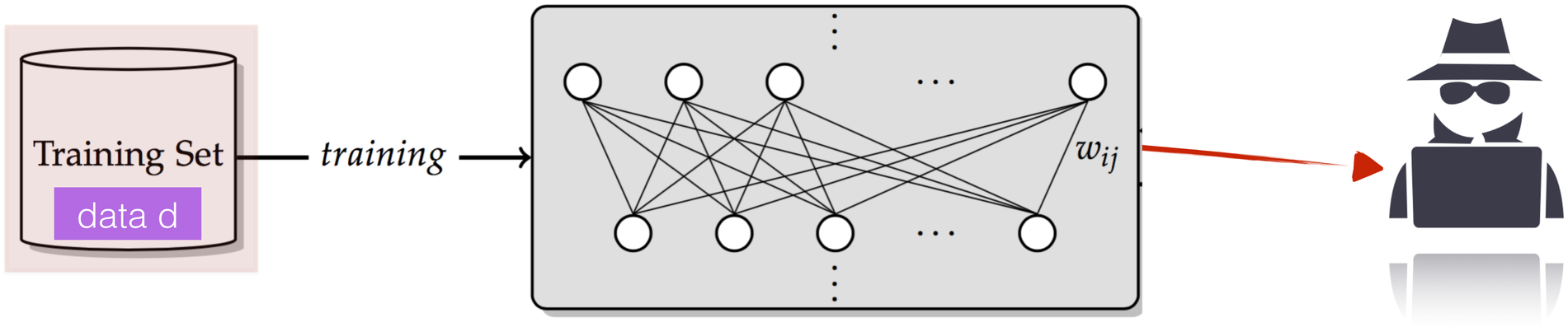
Alternative world where data d **is** part of training set



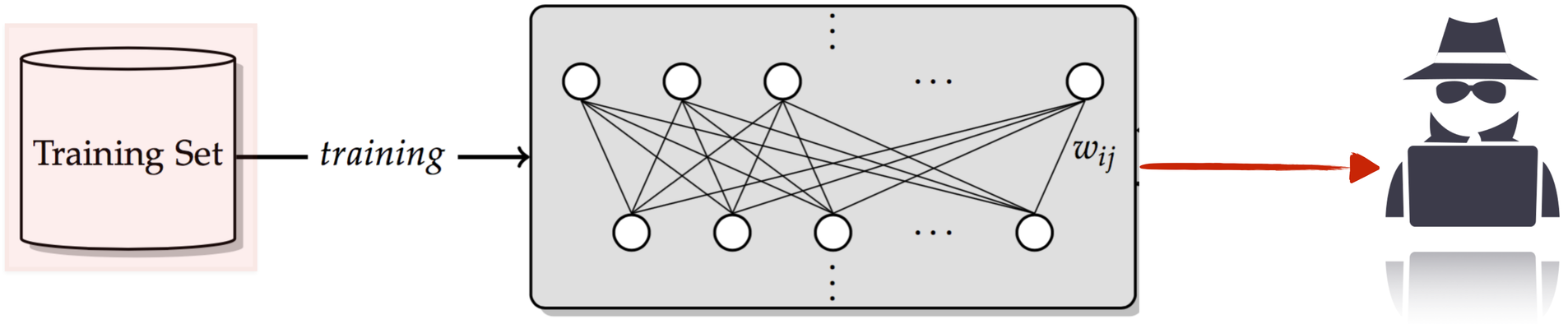
A world in which data d **is not** part of training set



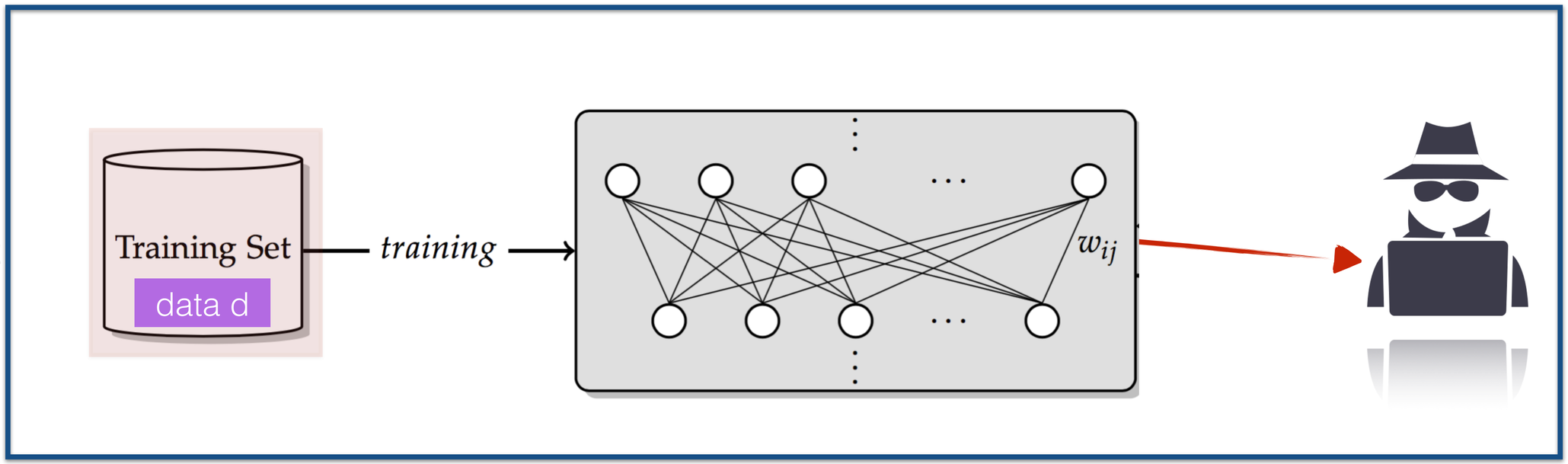
Alternative world where data d **is** part of training set



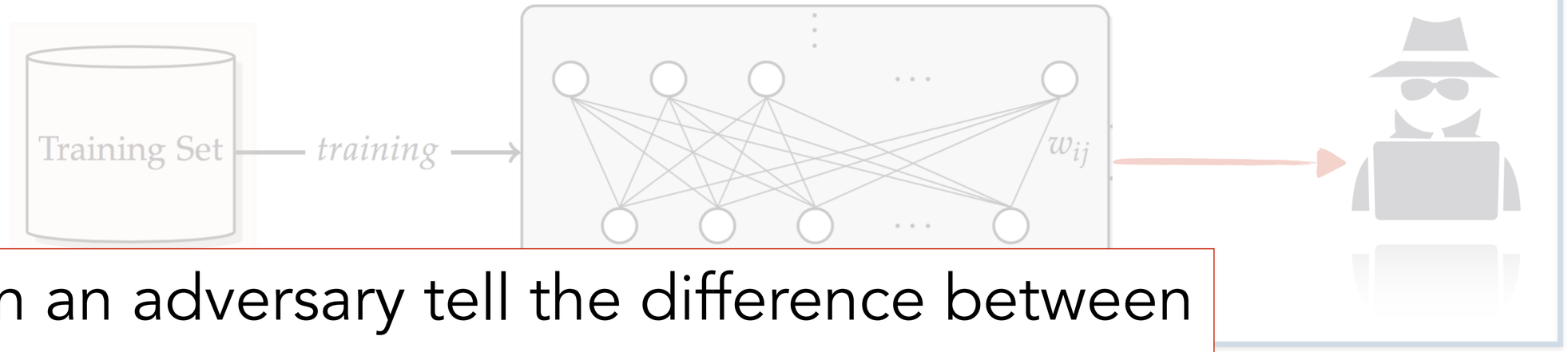
A world in which data d **is not** part of training set



Alternative world where data d **is** part of training set



A world in which data d **is not** part of training set

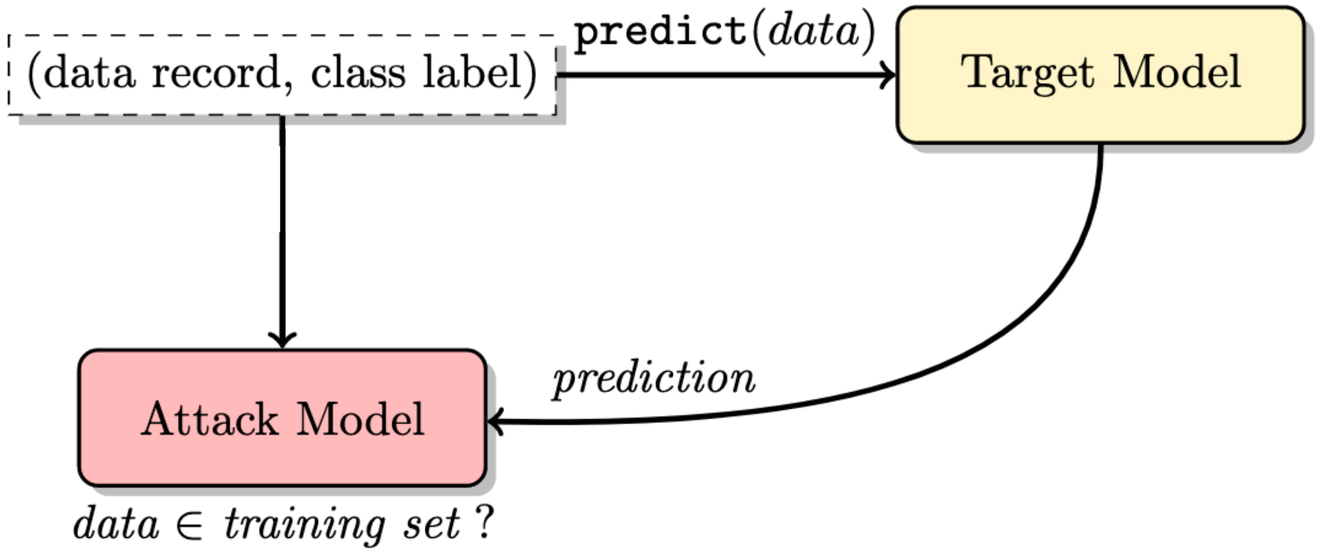


Can an adversary tell the difference between these two worlds?

Alternative world where data d **is** part of training set

Membership Inference

- Given a model, can an adversary infer whether a particular data point is part of its training set?
- Success of attacker is a metric for privacy loss

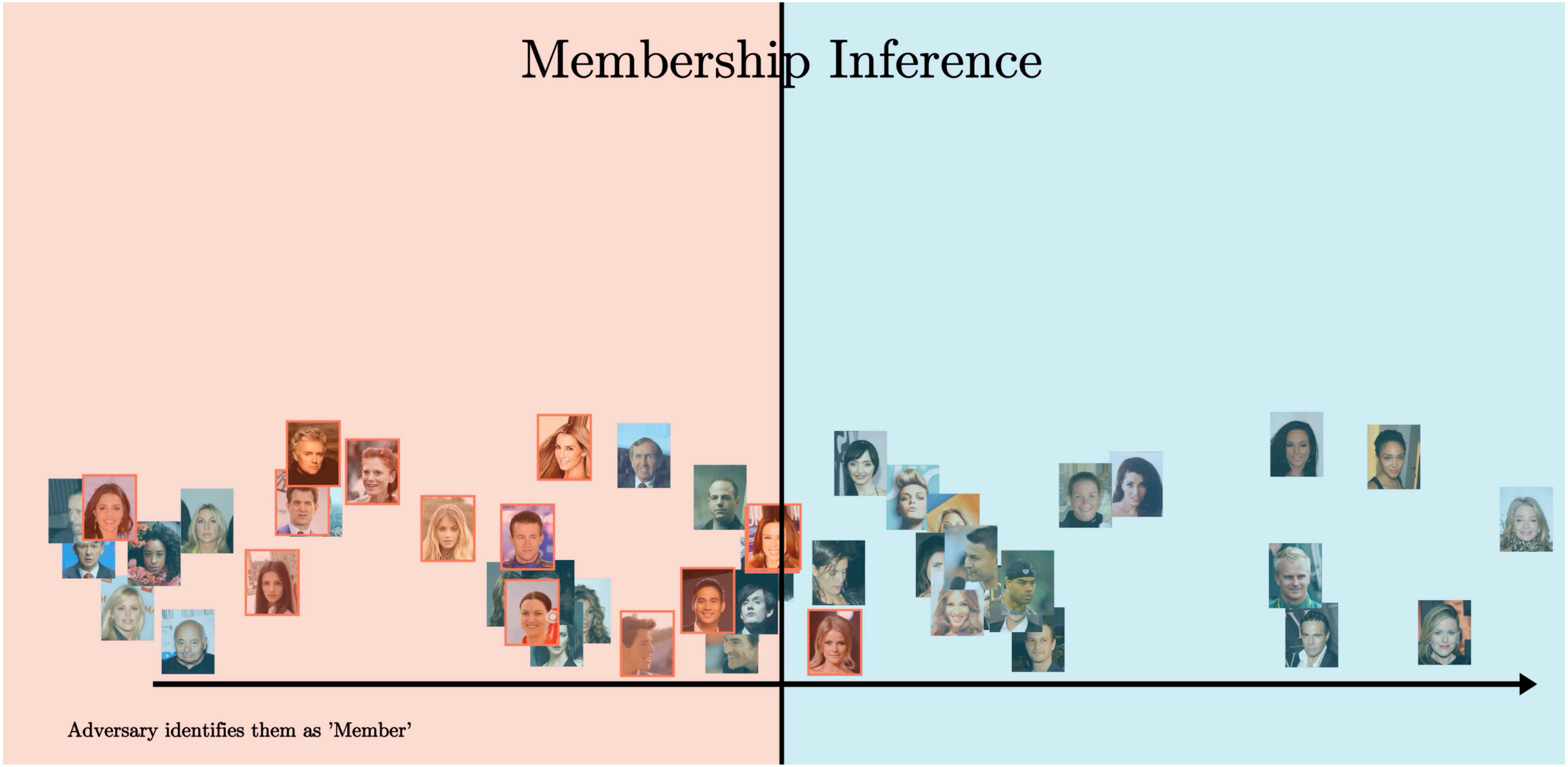


[Shokri, Stronati, Song, Shmatikov] Membership Inference Attacks against Machine Learning Models, SP'17

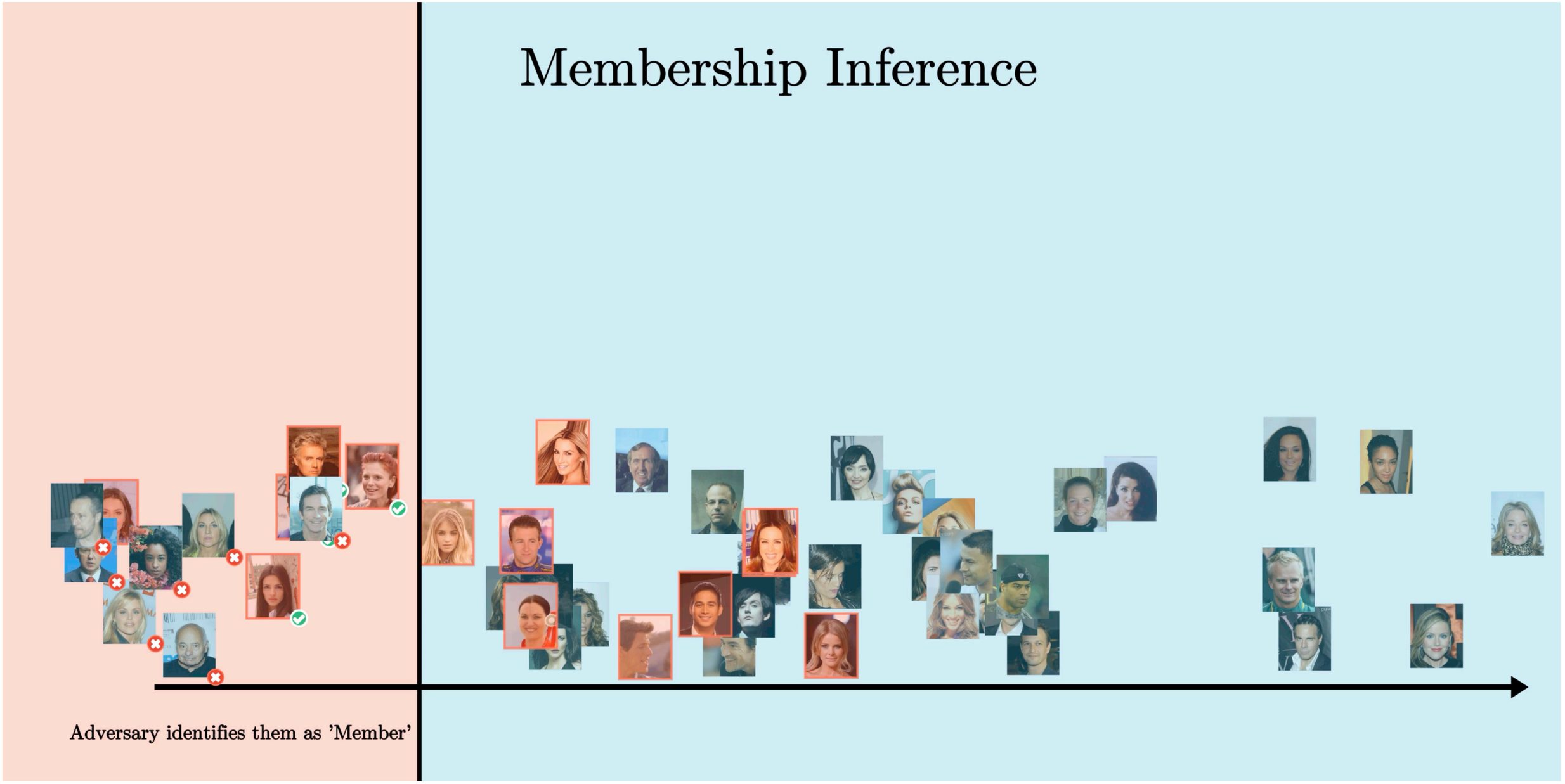
Success rate of adversary indicates information leakage of models about their training data



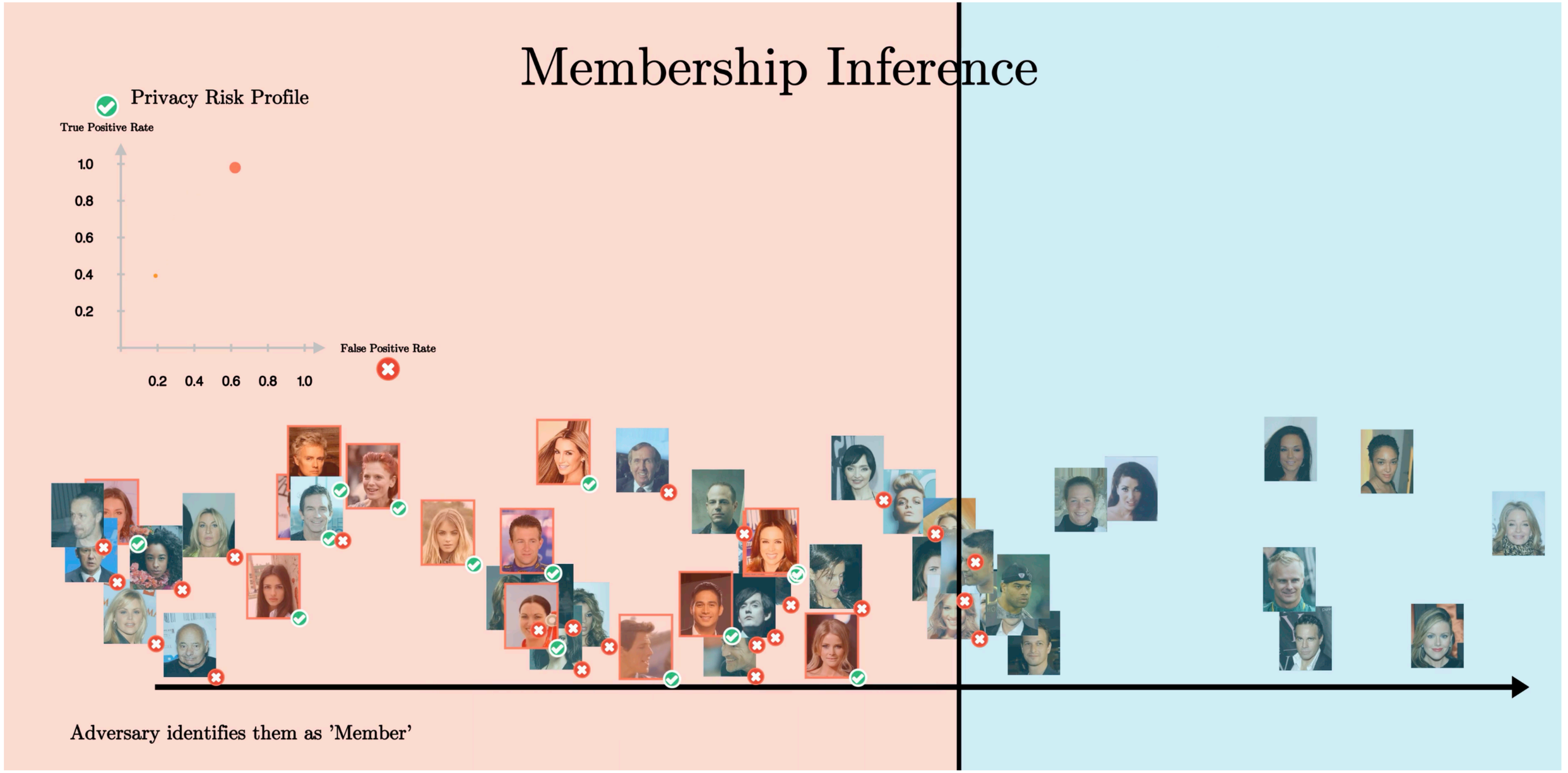
[Ye, Maddi, Murakonda, Bindschaedler, Shokri] Enhanced Membership Inference Attacks against Machine Learning Models, CCS'22



[Ye, Maddi, Murakonda, Bindschaedler, Shokri] Enhanced Membership Inference Attacks against Machine Learning Models, CCS'22



[Ye, Maddi, Murakonda, Bindschaedler, Shokri] Enhanced Membership Inference Attacks against Machine Learning Models, CCS'22

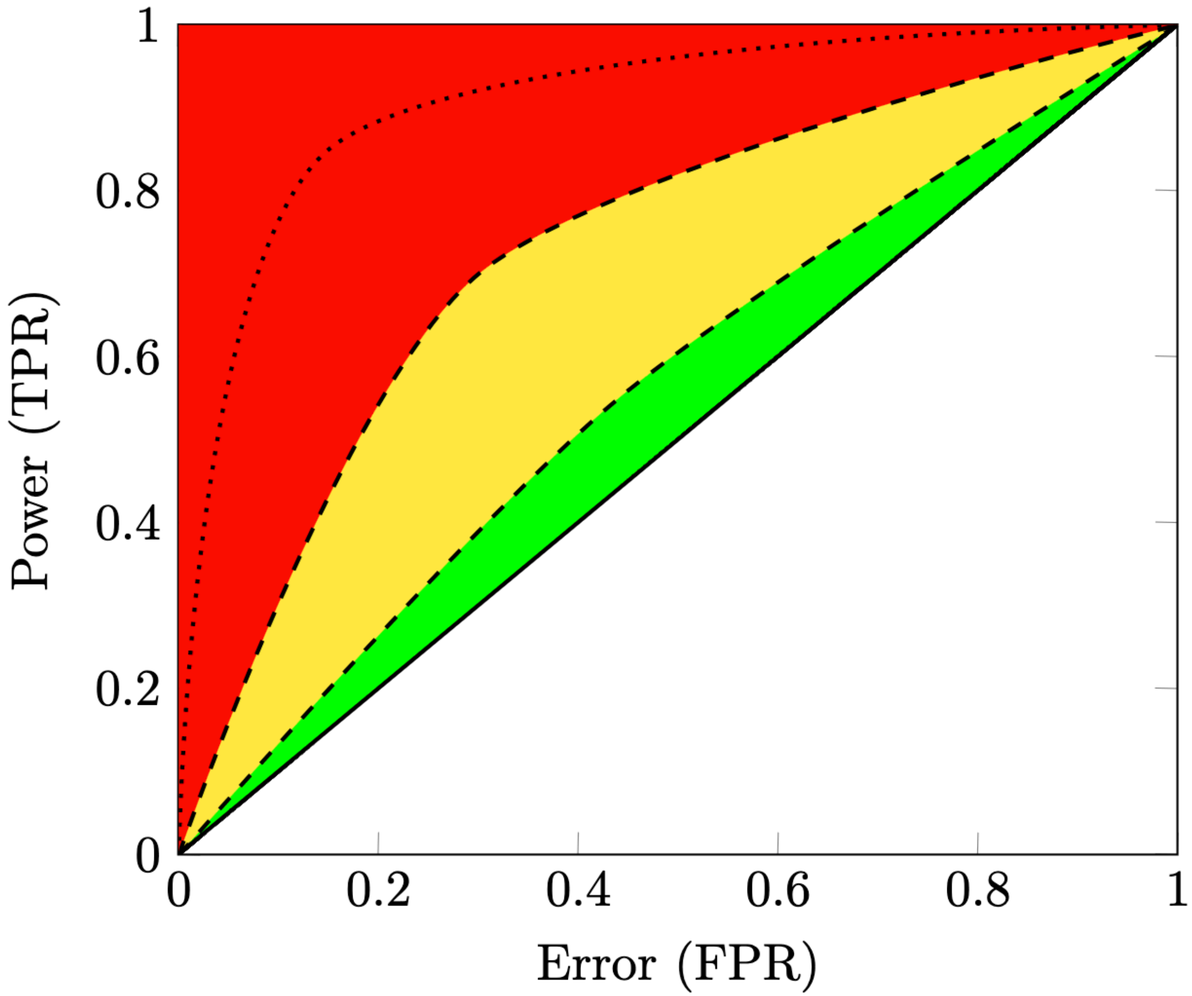


[Ye, Maddi, Murakonda, Bindschaedler, Shokri] Enhanced Membership Inference Attacks against Machine Learning Models, CCS'22

Actionable Assessment of Privacy Risks

Towards left: risk of being singled out ← → Towards right: more anonymity

The fraction of private dataset that are at risk



High risk — Do not deploy

Medium risk — Requires further investigations

Low risk — Safe to deploy

AI Regulations and Guidelines

A Taxonomy and Terminology of Adversarial Machine Learning

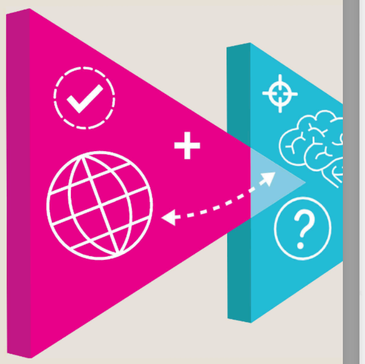
Elham Tabassi
Kevin J. Burns
Michael Hadjimichael
Andres D. Molina-Markham
Julian T. Sexton

This publication is available free of charge from:
https://doi.org/10.6028/NIST.IR.8269-draft

NIST
National Institute of Standards and Technology
U.S. Department of Commerce

Guidance on the AI auditing framework

Draft guidance for consultati



ico.
Information Commissioner's Office

European Commission

Home > Publications > White Paper on Artificial Intelligence: a European approach to excellence and trust

WHITE PAPER

White Paper on Artificial Intelligence: a European approach to excellence and trust

EXECUTIVE OFFICE OF THE PRESIDENT
OFFICE OF MANAGEMENT AND BUDGET
WASHINGTON, D.C. 20503

November 17, 2020

THE DIRECTOR

M-21-06

MEMORANDUM FOR THE HEADS OF EXECUTIVE DEPARTMENTS AND AG

FROM: Russell T. Vought
Director

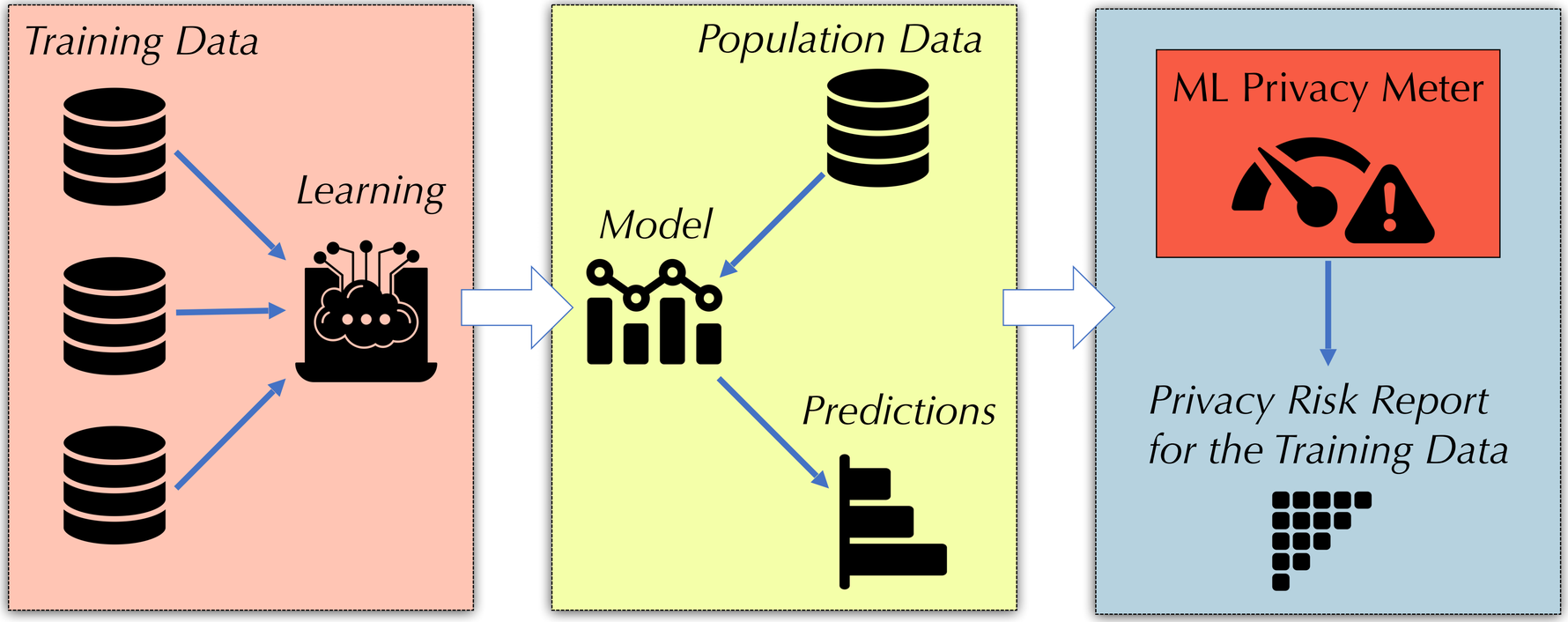
SUBJECT: Guidance for Regulation of Artificial Intelligence Applications

AI Regulations and Guidelines

- "… membership inferences show that AI models can inadvertently contain **personal data**"
- "Attacks that reveal confidential information about the data include membership inference …"
- **"… should consider the risks to data throughout the design, development, and operation of an AI system"**

ML Privacy Meter

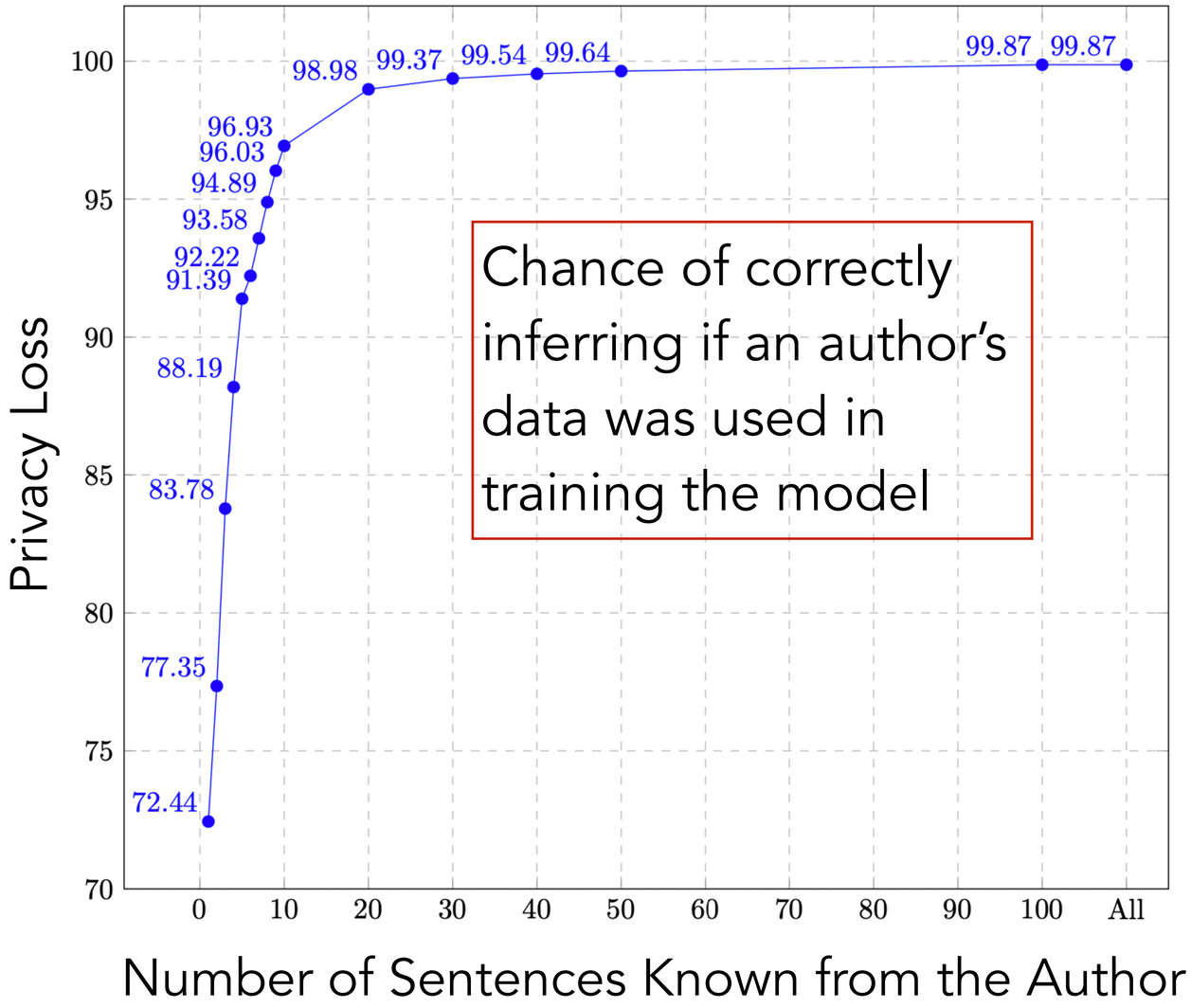
privacy-meter.com



Privacy Meter is an open source tool that enables quantifying the privacy risks of statistical and machine learning models.

python 3.6 | 3.7 | 3.8 downloads 2k pypi v1.5 slack @privacy meter license MIT cite citation Open in Colab
contributors 12 Forks 77 Stars 447 license MIT

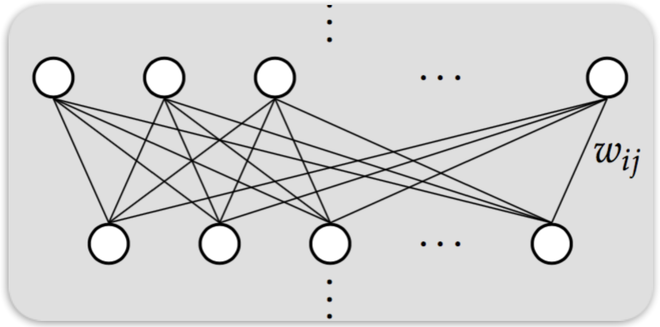
Example: Generative Model



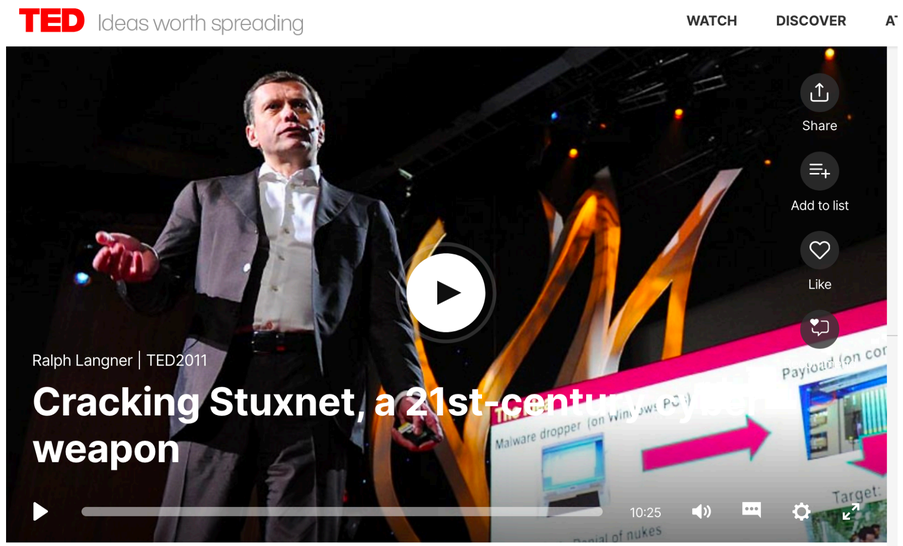




(Speaker Annotated TED talks)
SATED dataset



Examples of Vulnerable Training Data



But it gets worse. And this is very important, what I'
generic. It doesn't have anything to do, in specifics,
would work as well, for example, in a power plant or
don't have -- as an attacker -- you don't have to del
the case of Stuxnet. You could also use conventiona



This year, Germany is celebrating the 25th anniversar
1989, the Communist regime was moved away, the B
German Democratic Republic, the GDR, in the East wa
in the West to found today's Germany. Among many c
the East German secret police, known as the Stasi. Or
were opened to the public, and historians such as me
about how the GDR surveillance state functioned.

We need to “Preserve” Privacy in Language Models

[Brown, Lee, Mireshghallah, Shokri, Tramèr] What Does it Mean for a Language Model to Preserve Privacy?, FAccT 2022

What does it mean to “Preserve” Privacy in Language Models?

[Brown, Lee, Mireshghallah, Shokri, Tramèr] What Does it Mean for a Language Model to Preserve Privacy?, FAccT 2022

Privacy in the Realm of Language

- Language serves as a medium for communication
- Natural language reflects our private lives and identities
- Privacy concerns in natural language are as broad as those of real life

Challenges of Preserving Privacy in Language

- Privacy is **contextual**
  - We (as humans) choose whether to disclose or conceal information, based on the context.

- Detecting the context in which text is generated can be challenging (to algorithms) or may exist beyond the text itself.

Challenges of Preserving Privacy in Language

- Private information can take many different forms.
- Private information has no boundaries: information about Alice could be found in Bob's text data.
- Repeated information can still be private information.

Existing Privacy Preserving Methods

Existing Privacy Preserving Methods

- Scrubbing: Remove secrets from text (before training a model)

Existing Privacy Preserving Methods

- Scrubbing: Remove secrets from text (before training a model)

- Differential privacy: Use a training algorithm for which adding or removing a user to the training set results in almost indistinguishable models

Existing Privacy Preserving Methods

- Scrubbing: Remove secrets from text (before training a model)
- Differential privacy: Use a training algorithm for which adding or removing a user to the training set results in almost indistinguishable models
- Use public data: Crowl the internet to gather all available text

Existing Privacy Preserving Methods

- Scrubbing: Remove secrets from text (before training a model)

  Assumption: Every secret can be found through certain rules.

- Differential privacy: Use a training algorithm for which adding or removing a user to the training set results in almost indistinguishable models

- Use public data: Crowl the internet to gather all available text

Existing Privacy Preserving Methods

- Scrubbing: Remove secrets from text (before training a model)

  Assumption: Every secret can be found through certain rules.

- Differential privacy: Use a training algorithm for which adding or removing a user to the training set results in almost indistinguishable models

  Assumption: A secret is entirely in one user's data

- Use public data: Crowl the internet to gather all available text

Existing Privacy Preserving Methods

- Scrubbing: Remove secrets from text (before training a model)

  Assumption: Every secret can be found through certain rules.

- Differential privacy: Use a training algorithm for which adding or removing a user to the training set results in almost indistinguishable models

  Assumption: A secret is entirely in one user's data

- Use public data: Crowl the internet to gather all available text

  Assumption: Publicly available data is public-intended

Existing Privacy Preserving Methods

- Scrubbing: Remove secrets from text (before training a model)

~~Assumption: Every secret can be found through certain rules.~~

- Differential privacy: Use a training algorithm for which adding or removing a user to the training set results in almost indistinguishable models

~~Assumption: A secret is entirely in one user's data~~

- Use public data: Crowl the internet to gather all available text

~~Assumption: Publicly available data is public-intended~~

Paths Forward

- The problem is hard. Privacy concerns in natural language are as broad as those of real life, and the context in which some information is private is hard if not impossible to infer

Paths Forward

- The problem is hard. Privacy concerns in natural language are as broad as those of real life, and the context in which some information is private is hard if not impossible to infer
- Nevertheless, we can reduce privacy risks
  - Make use of publicly-intended data
  - Use private personalization
    - (Only X has access to model trained on X's data)
  - Use privacy-preserving algorithms
  - Audit privacy risks of algorithms