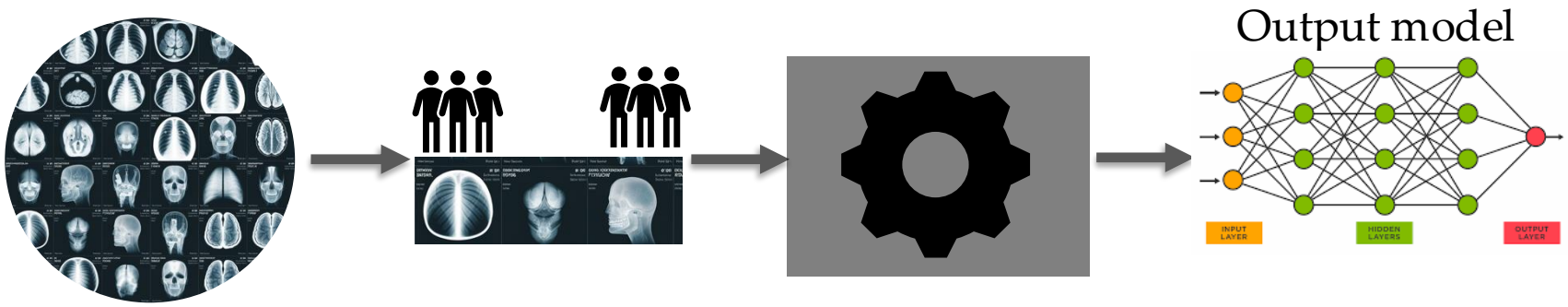
The Sample Complexity of Membership Inference Attacks & Privacy Attacks

Mahdi Haghifam
Research Assistant Professor at TTIC

Adam Smith (Boston University)
Jonathan Ullman (Northeastern University)

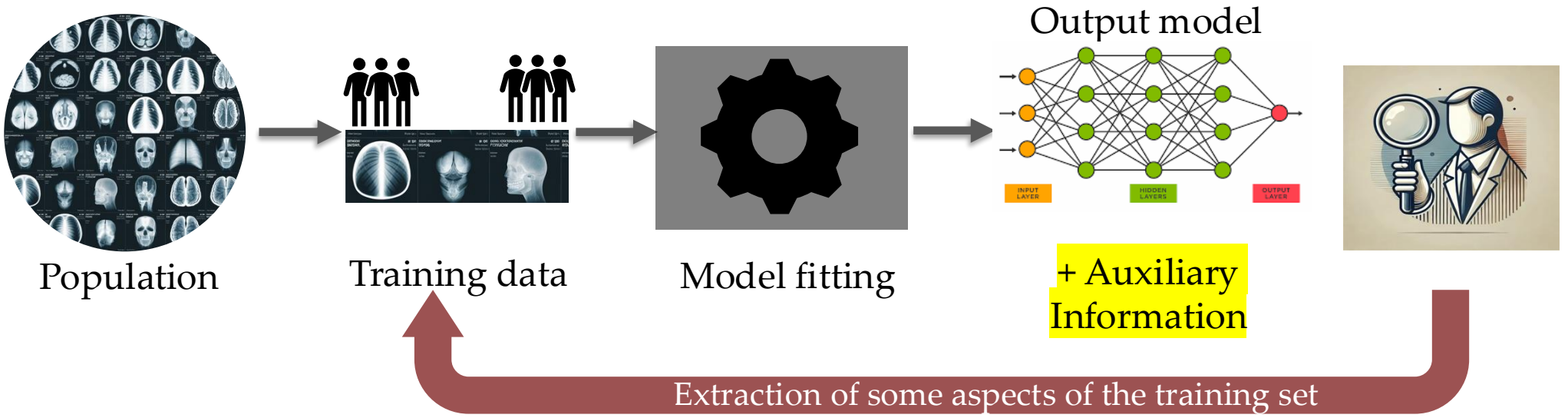
2 Minute Summary of the Talk: Background Knowledge of Privacy Attacks

2 Minute Summary of the Talk: Background Knowledge of Privacy Attacks

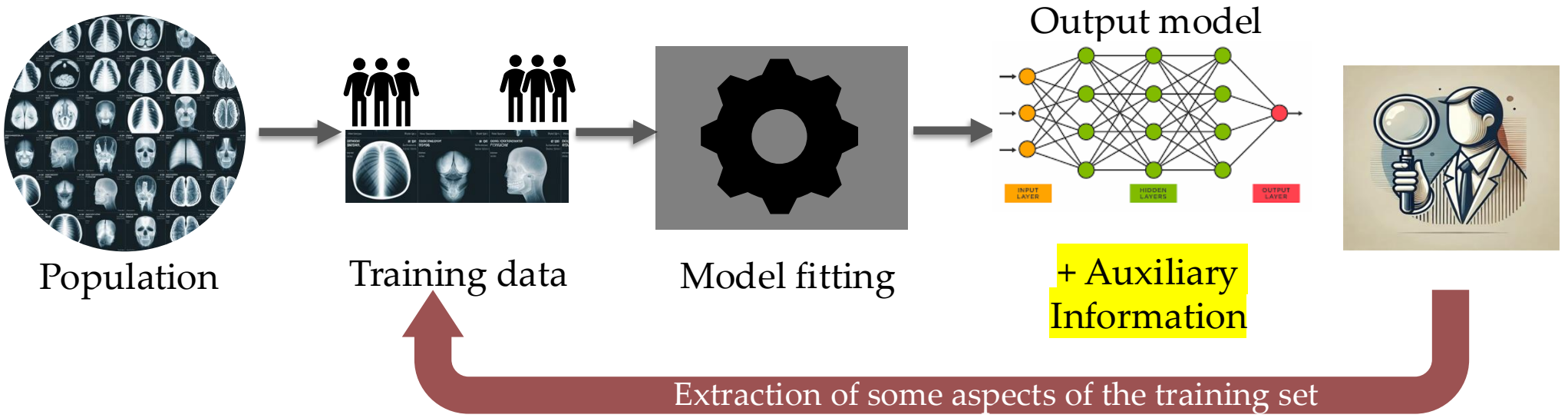


Population Training data Model fitting

2 Minute Summary of the Talk: Background Knowledge of Privacy Attacks



2 Minute Summary of the Talk: Background Knowledge of Privacy Attacks

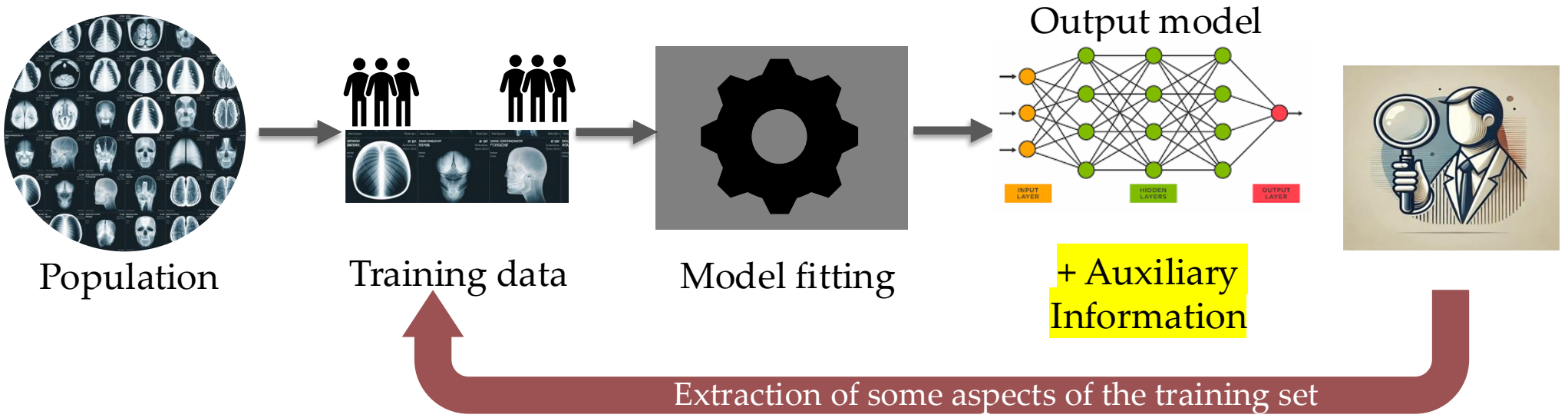


Setup of this talk

Privacy Attacks = Membership Inference Attack (MIA)

Auxiliary Information = Samples from the same population as the training data

2 Minute Summary of the Talk: Background Knowledge of Privacy Attacks



Setup of this talk

Privacy Attacks = Membership Inference Attack (MIA)

Auxiliary Information = Samples from the same population as the training data

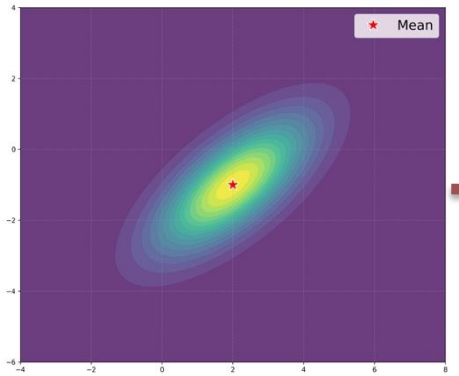
How much auxiliary information about population is necessary to carry-out a membership inference attack?

Main Result: Sample Complexity of MIAs

Main Result: Sample Complexity of MIAs

MIA on Gaussian Mean Estimation

Main Result: Sample Complexity of MIAs



MIA on Gaussian Mean Estimation

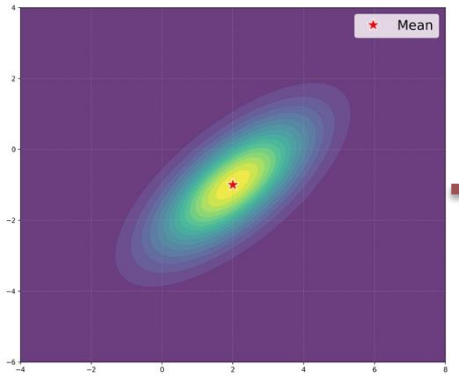
$S_n = (X_1, \ldots, X_n)$

Mean estimator

$\hat{\mu}$

+ Auxiliary Information

Main Result: Sample Complexity of MIAs



MIA on Gaussian Mean Estimation

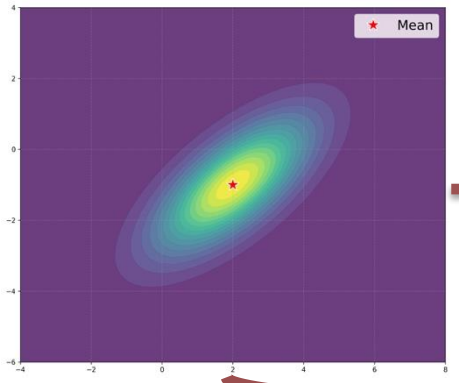
$S_n = (X_1, \ldots, X_n)$

Mean estimator

$\hat{\mu}$

+ Auxiliary Information

Main Result: Sample Complexity of MIAs



MIA on Gaussian Mean Estimation

$S_n = (X_1, \dots, X_n)$

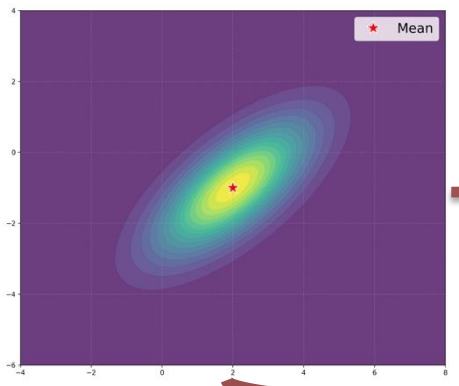
Mean estimator

$\hat{\mu}$

+ Auxiliary Information

$Y^m = (Y_1, \dots, Y_m)$

Main Result: Sample Complexity of MIAs



MIA on Gaussian Mean Estimation

$S_n = (X_1, \dots, X_n)$

Mean estimator

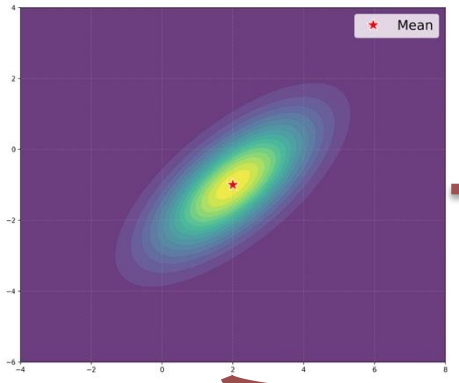
$\hat{\mu}$

+ Auxiliary Information

$Y^m = (Y_1, \dots, Y_m)$

Prior Work:
When the attacker has the full knowledge of the data distribution,
if the dimension is large and/or the mean estimator is accurate
an informed attacker can carry-out MIA.

Main Result: Sample Complexity of MIAs



MIA on Gaussian Mean Estimation

$S_n = (X_1, \dots, X_n)$

Mean estimator

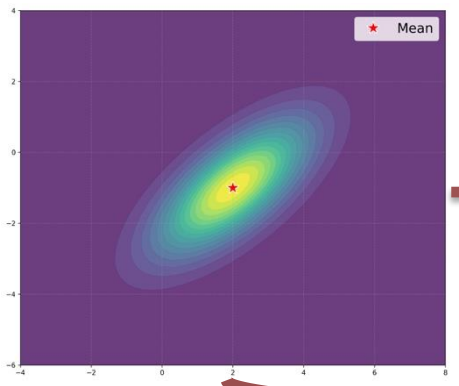
$\hat{\mu}$

+ Auxiliary Information

$Y^m = (Y_1, \dots, Y_m)$

Prior Work:
When the attacker has the full knowledge of the data distribution,
if the dimension is large and/or the mean estimator is accurate
an informed attacker can carry-out MIA.

Main Result: Sample Complexity of MIAs



MIA on Gaussian Mean Estimation

$S_n = (X_1, \dots, X_n)$

Mean estimator

$\hat{\mu}$

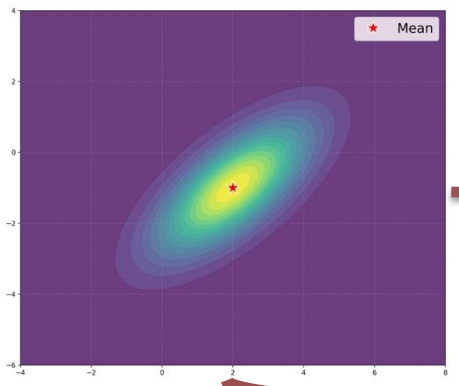
+ Auxiliary Information

$Y^m = (Y_1, \dots, Y_m)$

Prior Work:
When the attacker has the full knowledge of the data distribution,
if the dimension is large and/or the mean estimator is accurate
an informed attacker can carry-out MIA.

Question: What is the minimum m to design an MIA with a similar performance?

Main Result: Sample Complexity of MIAs



MIA on Gaussian Mean Estimation

$S_n = (X_1, \dots, X_n)$

Mean estimator

$\hat{\mu}$

+ Auxiliary Information

$Y^m = (Y_1, \dots, Y_m)$

Prior Work:
When the attacker has the full knowledge of the data distribution,
if the dimension is large and/or the mean estimator is accurate
an informed attacker can carry-out MIA.

Question: What is the minimum m to design an MIA with a similar performance?

There is a critical distinction based on the attacker's knowledge of the data's covariance structure.

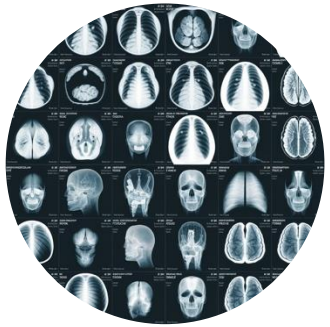
In certain setting the number of auxiliary samples any attacker needs is $\omega(n)$!

Outline of the Talk

- Membership Inference Attack and Auxiliary Information
- Membership Inference Attack on Gaussian Mean Estimators
- Proof Overview

Learning from Data

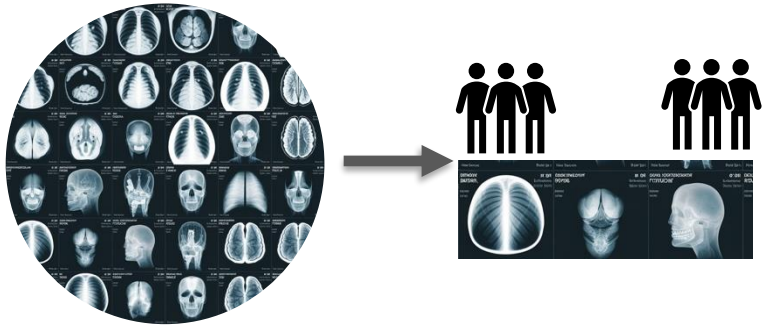
Learning from Data



Population

Motivating Example: Health effects of **stroke** on the survivors.

Learning from Data

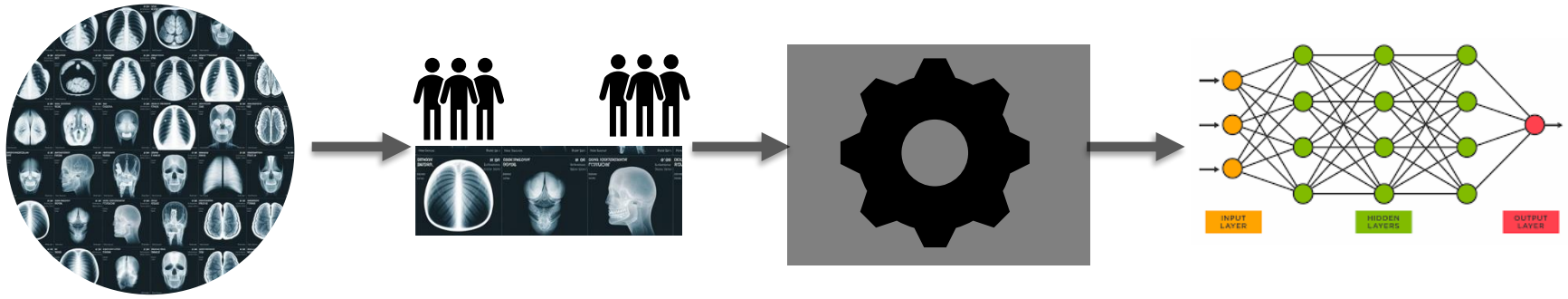


Population

Training data

Motivating Example: Health effects of **stroke** on the survivors.

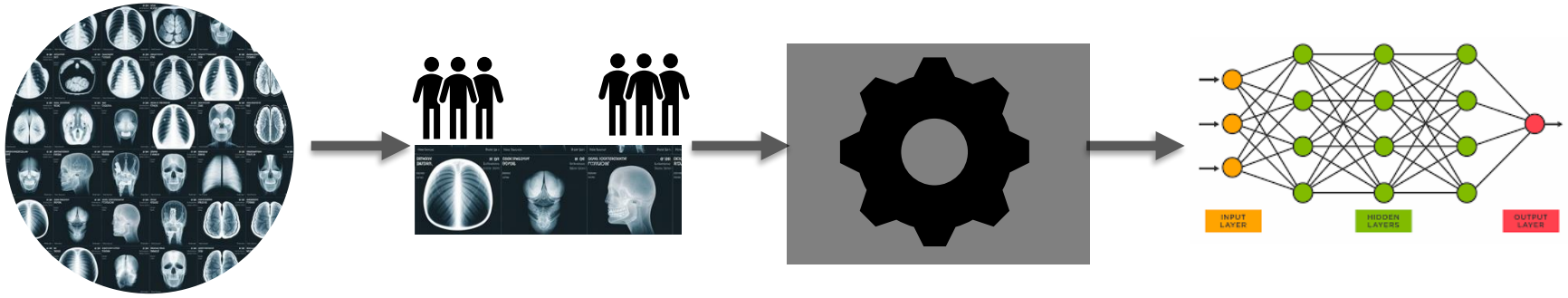
Learning from Data



Population → Training data → Model fitting → Output model

Motivating Example: Health effects of **stroke** on the survivors.

Learning from Data

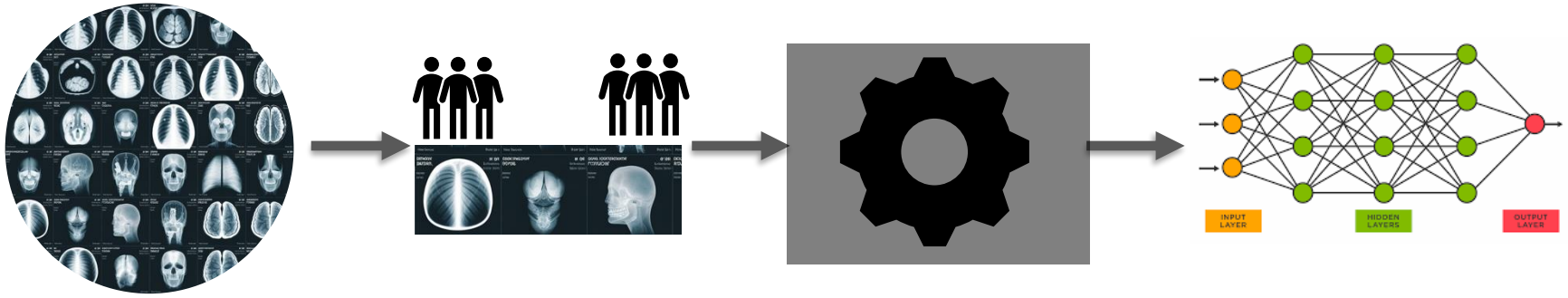


Population → Training data → Model fitting → Output model

Motivating Example: Health effects of **stroke** on the survivors.

Being part of the study
implies having stroke in the past.

Learning from Data



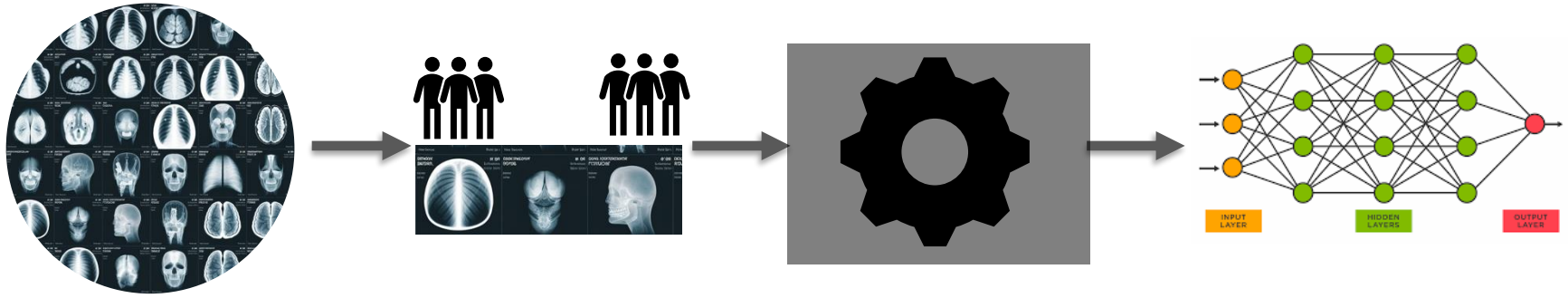
Population → Training data → Model fitting → Output model

Motivating Example: Health effects of **stroke** on the survivors.

Being part of the study
implies having stroke in the past.

An ideal algorithm **extracts only the information from the training data relevant to the population-level task**, without learning irrelevant details about individual training points.

Learning from Data



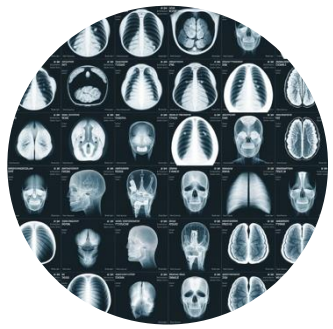
Population → Training data → Model fitting → Output model

Motivating Example: Health effects of **stroke** on the survivors.

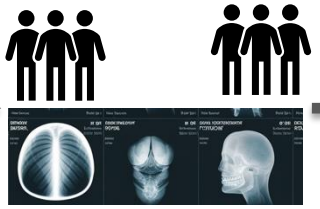
Being part of the study
implies having stroke in the past.

An ideal algorithm **extracts only the infor** **from the training data relevant to the population-level task**, without learning irr details about individual training points.

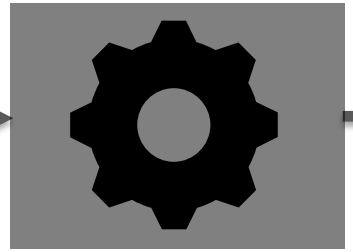
Privacy Attacks



Population

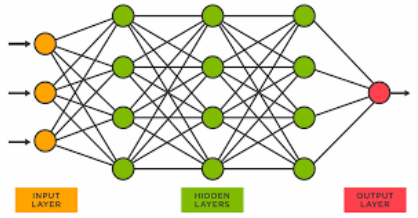


Training data

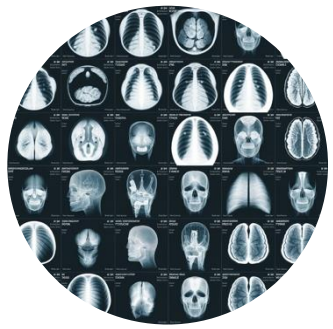


Model fitting

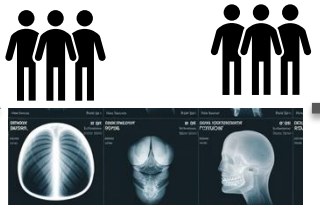
Output model



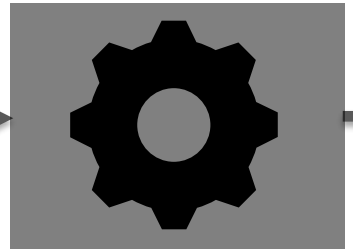
Privacy Attacks



Population

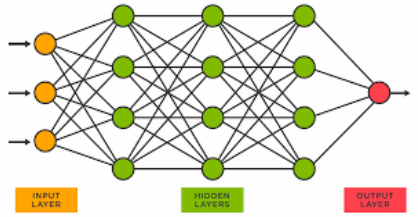


Training data



Model fitting

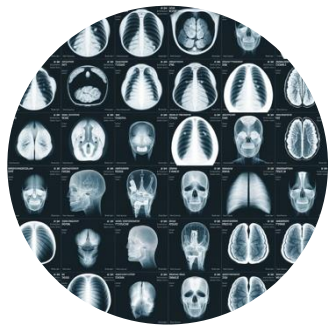
Output model



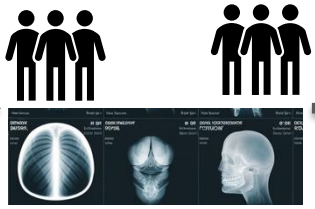
+ Auxiliary Information

Extraction of some aspects of the the training set

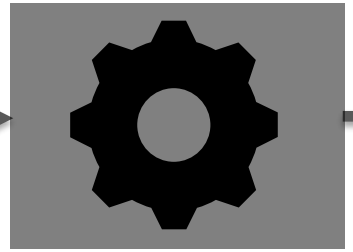
Privacy Attacks



Population

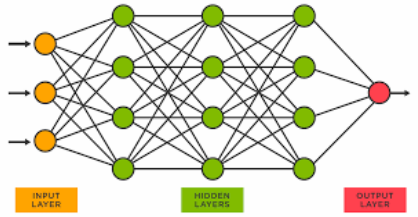


Training data



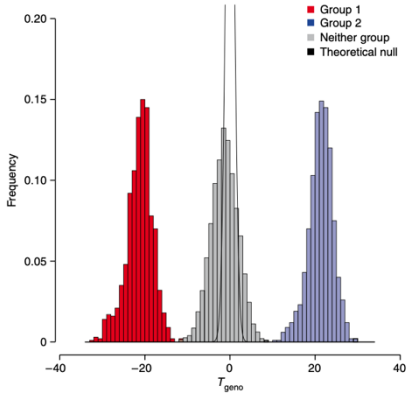
Model fitting

Output model

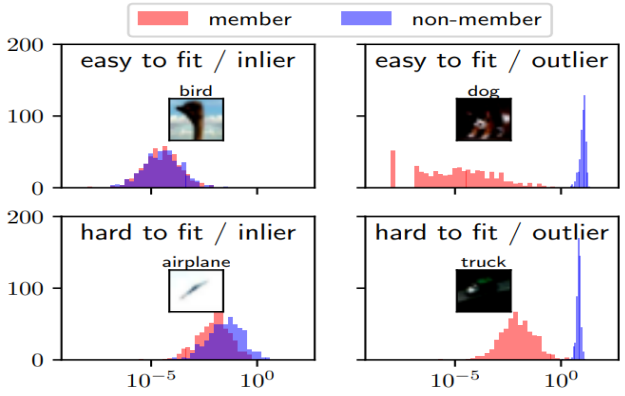


+ Auxiliary Information

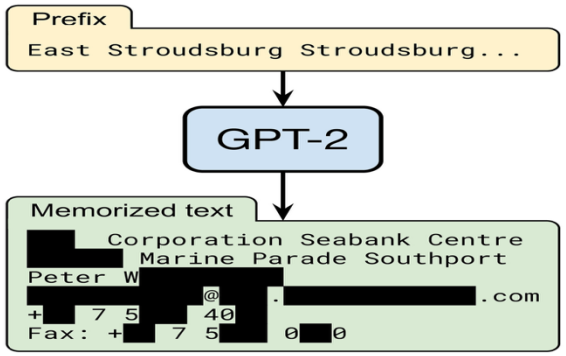
Extraction of some aspects of the the training set



Jacobs KB, Yeager M, Wacholder S, Craig D, Kraft P, Hunter DJ, Paschal J, Manolio TA, Tucker M, Hoover RN, Thomas GD. **A new statistic and its power to infer membership in a genome-wide association study using genotype frequencies.**



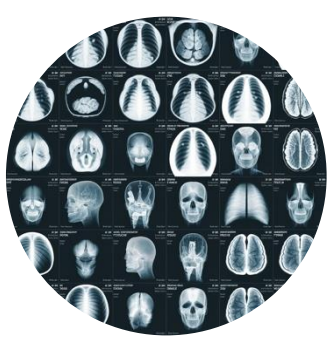
Carlini N, Chien S, Nasr M, Song S, Terzis A, Tramer F. **Membership inference attacks from first principles.**



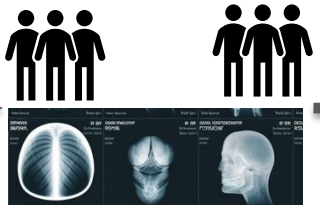
Carlini N, Tramer F, Wallace E, Jagielski M, Herbert-Voss A, Lee K, Roberts A, Brown T, Song D, Erlingsson U, Oprea A. **Extracting training data from large language models.**

Auxiliary Information of Privacy Attacks

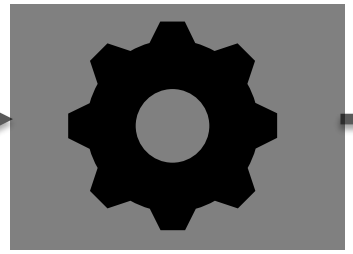
Auxiliary Information of Privacy Attacks



Population

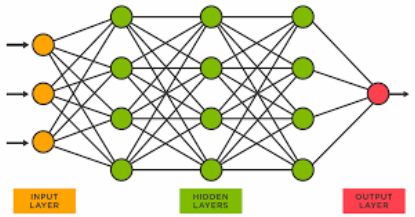


Training data



Model fitting

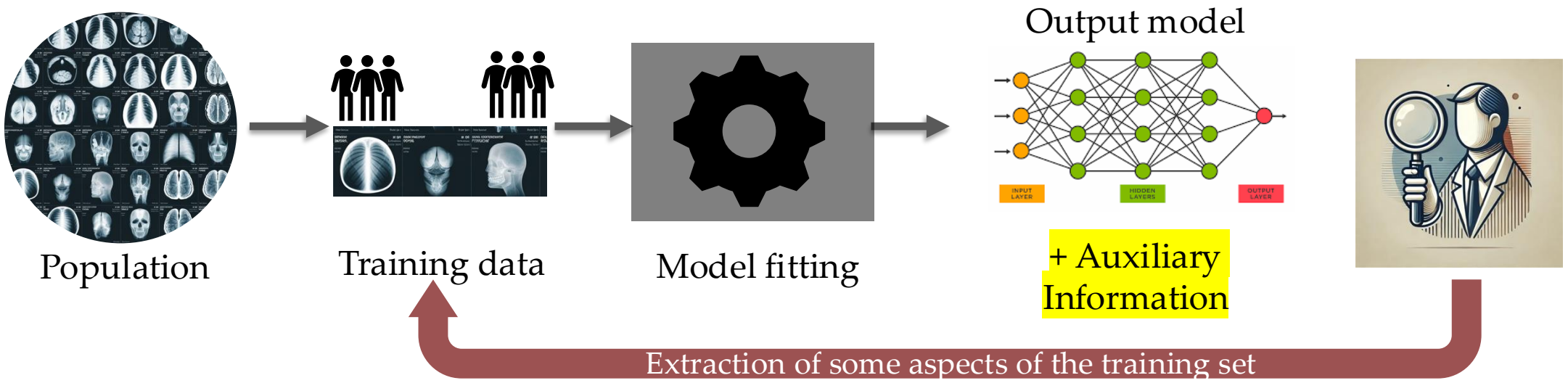
Output model



+ Auxiliary Information

Extraction of some aspects of the training set

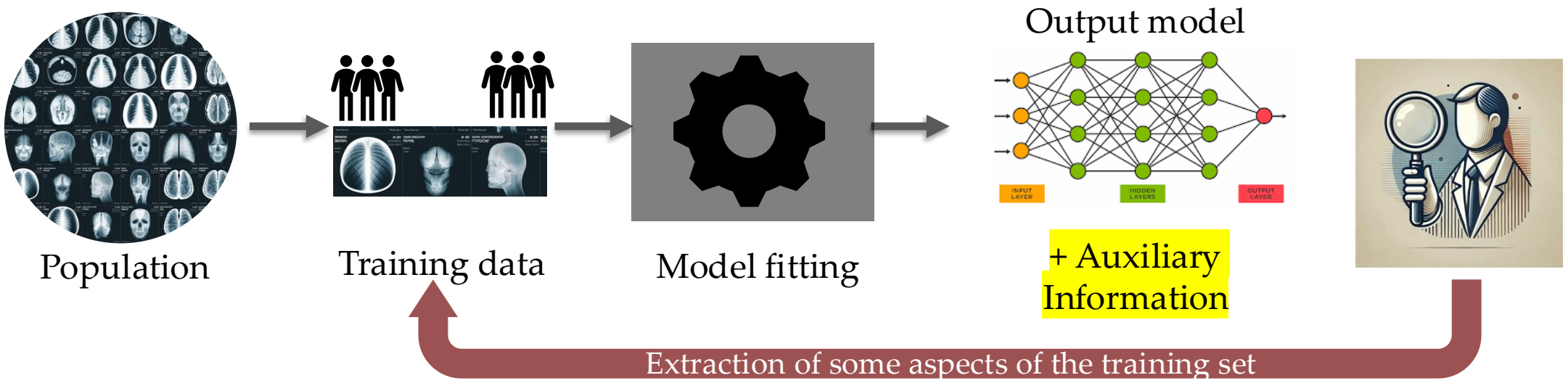
Auxiliary Information of Privacy Attacks



Auxiliary information:

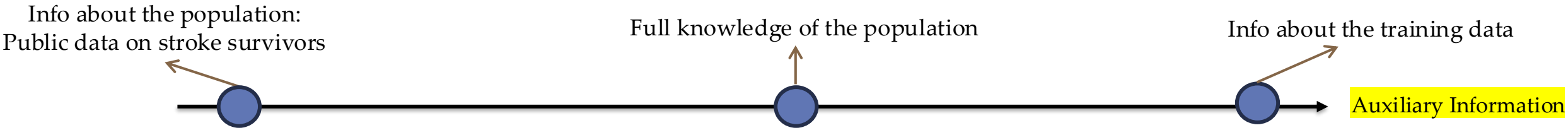
- Side information available to the attacker beyond the output model.
- The practicality of attack is closely related to how plausible the auxiliary information is.

Auxiliary Information of Privacy Attacks

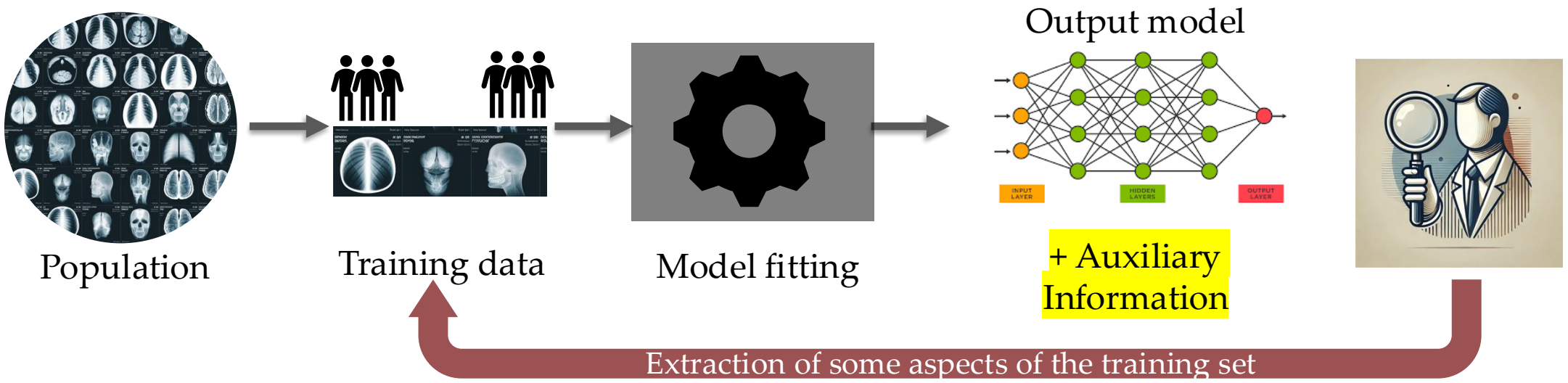


Auxiliary information:

- Side information available to the attacker beyond the output model.
- The practicality of attack is closely related to how plausible the auxiliary information is.

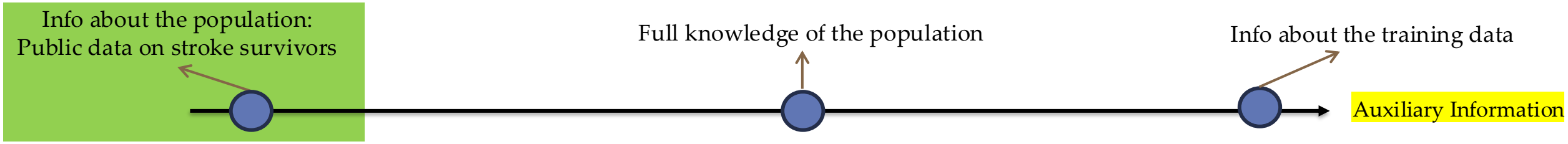


Auxiliary Information of Privacy Attacks

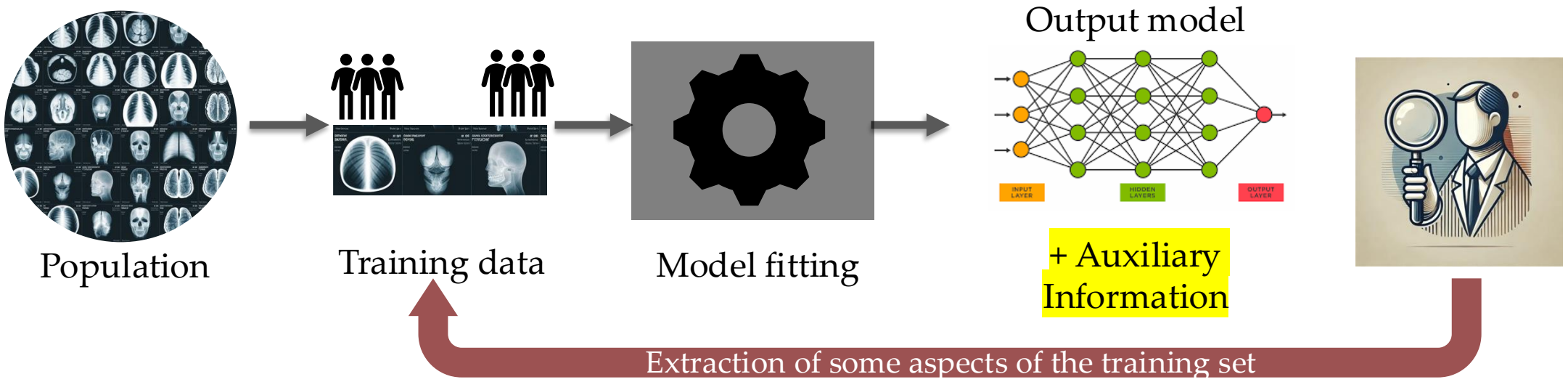


Auxiliary information:

- Side information available to the attacker beyond the output model.
- The practicality of attack is closely related to how plausible the auxiliary information is.

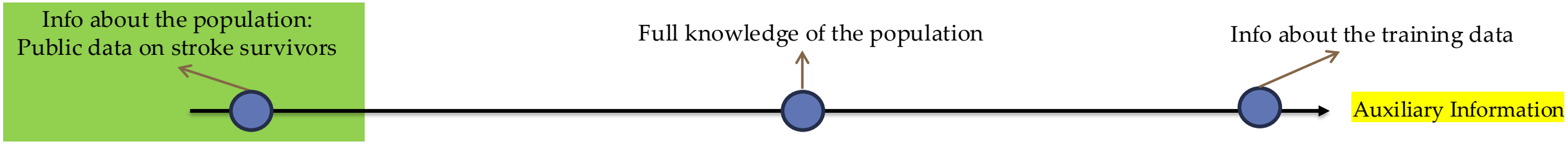


Auxiliary Information of Privacy Attacks



Auxiliary information:

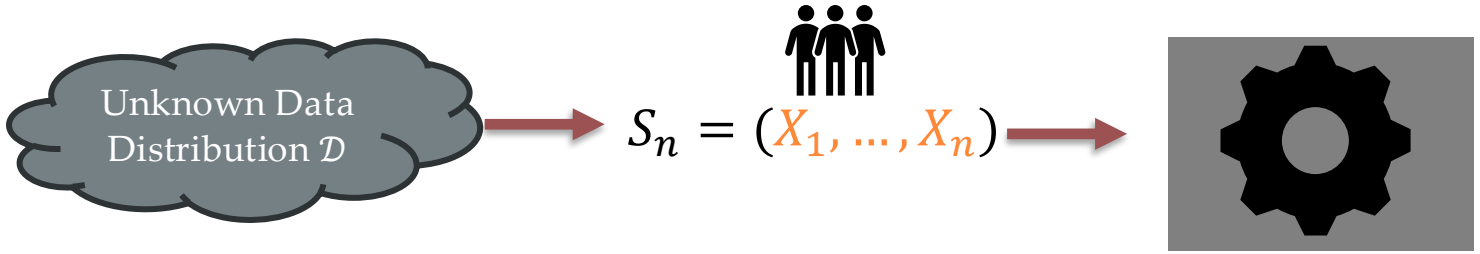
- Side information available to the attacker beyond the output model.
- The practicality of attack is closely related to how plausible the auxiliary information is.



How many auxiliary samples about population is necessary to carry-out a privacy attack?

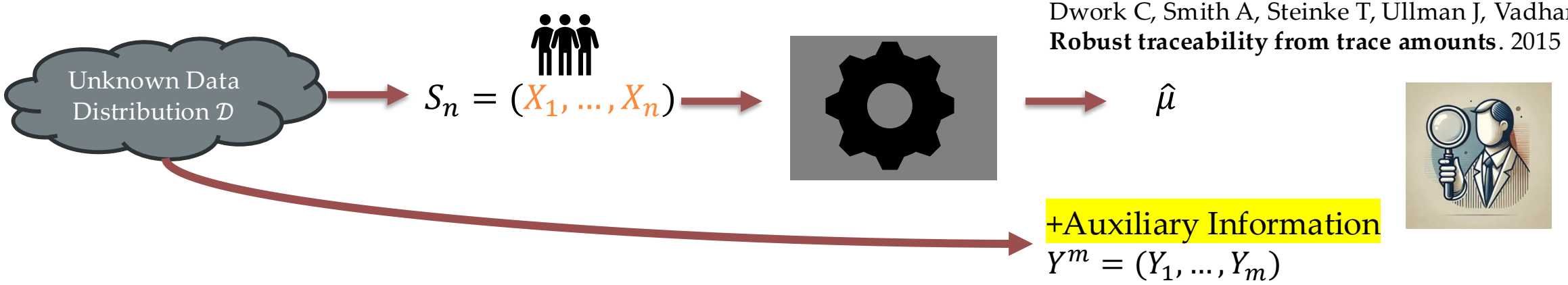
Membership Inference Attacks (MIA)

Membership Inference Attacks (MIA)



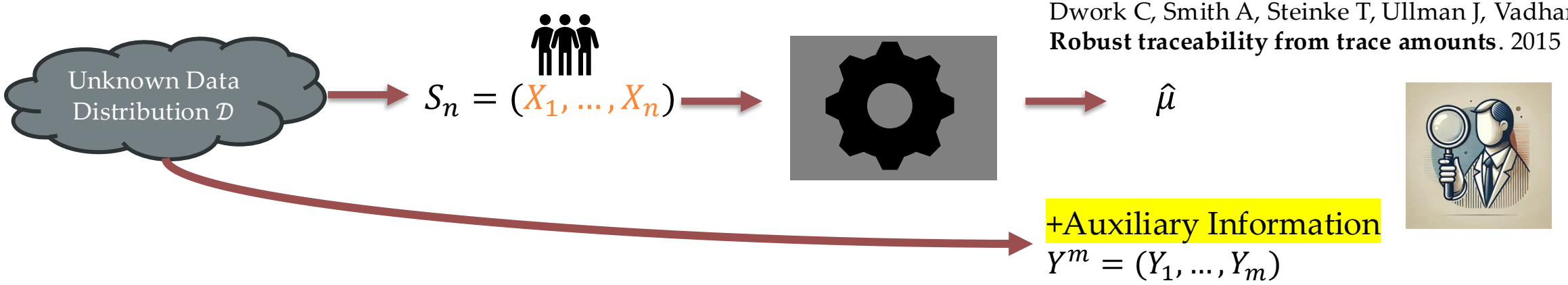
Dwork C, Smith A, Steinke T, Ullman J, Vadhan S. **Robust traceability from trace amounts**. 2015

Membership Inference Attacks (MIA)



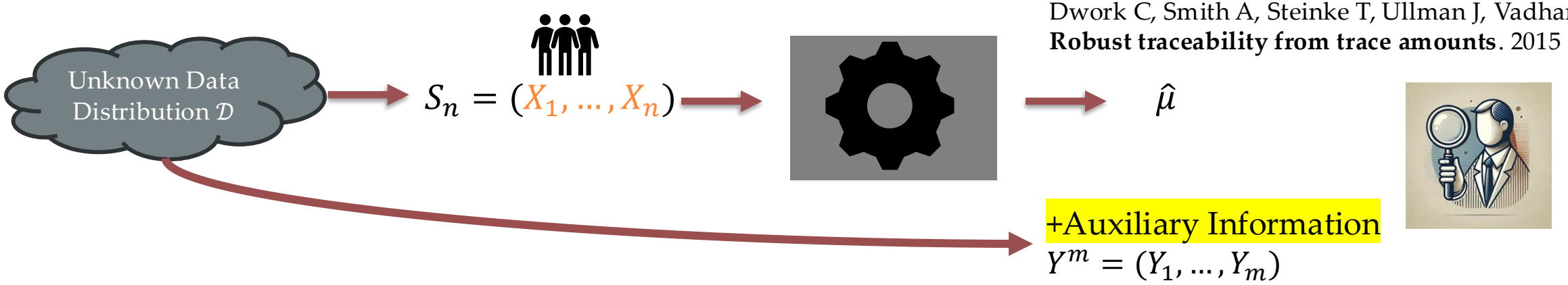
Dwork C, Smith A, Steinke T, Ullman J, Vadhan S. **Robust traceability from trace amounts**. 2015

Membership Inference Attacks (MIA)



Dwork C, Smith A, Steinke T, Ullman J, Vadhan S. **Robust traceability from trace amounts**. 2015

Membership Inference Attacks (MIA)

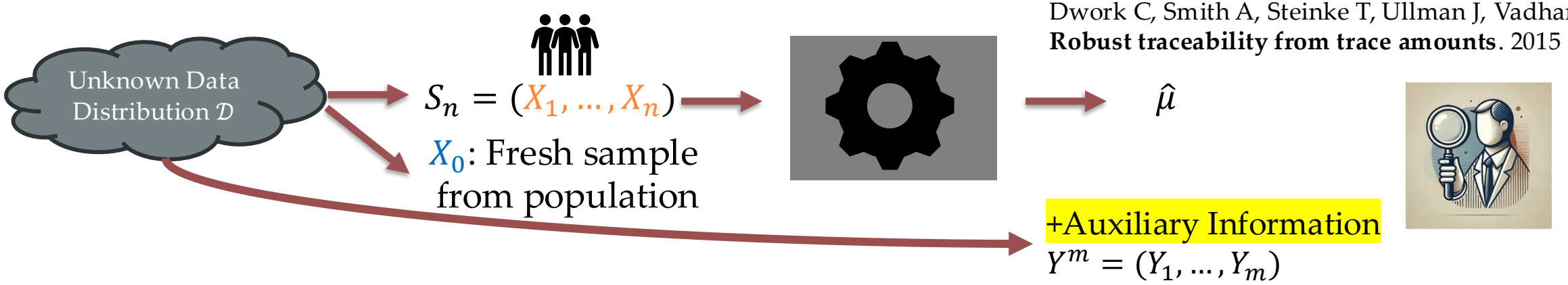


Dwork C, Smith A, Steinke T, Ullman J, Vadhan S. **Robust traceability from trace amounts**. 2015

Definition: **minimal m-sample based MIA**

ψ: auxillary information $\times$ output model $\times$ data point $\rightarrow \{IN, OUT\}$

Membership Inference Attacks (MIA)



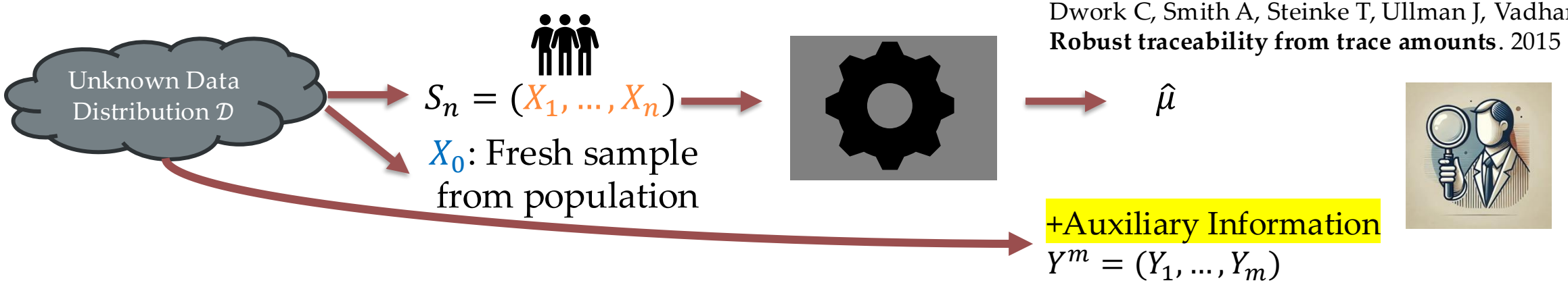
Dwork C, Smith A, Steinke T, Ullman J, Vadhan S. **Robust traceability from trace amounts**. 2015

Definition: **minimal m-sample based MIA**

ψ: auxillary information $\times$ output model $\times$ data point $\rightarrow \{IN, OUT\}$

$$\Pr(\psi(Y^m, \hat{\mu}, X_0) = OUT) \geq 0.51$$

Membership Inference Attacks (MIA)



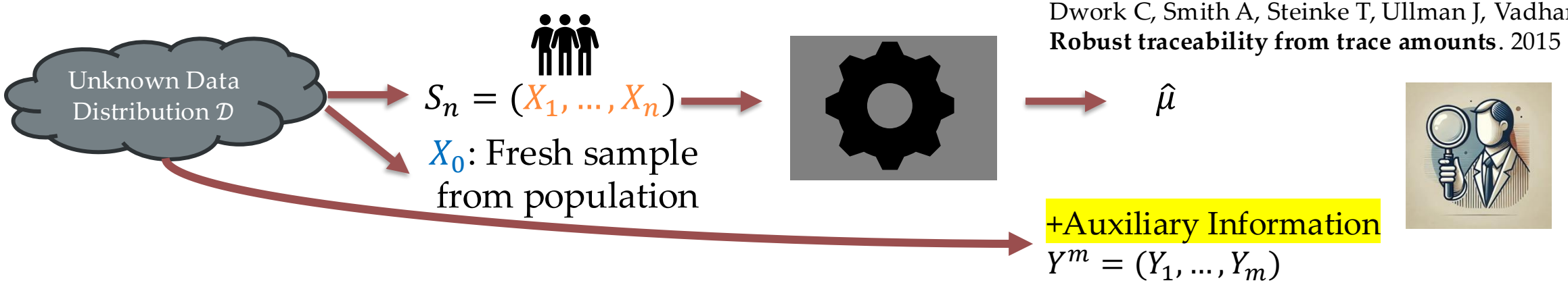
Dwork C, Smith A, Steinke T, Ullman J, Vadhan S. **Robust traceability from trace amounts**. 2015

Definition: **minimal m-sample based MIA**

ψ: auxillary information $\times$ output model $\times$ data point $\rightarrow \{IN, OUT\}$

$$\Pr(\psi(Y^m, \hat{\mu}, X_0) = OUT) \geq 0.51$$
$$\forall i \in [n]: \Pr(\psi(Y^m, \hat{\mu}, X_i) = IN) \geq 0.51$$

Membership Inference Attacks (MIA)



Dwork C, Smith A, Steinke T, Ullman J, Vadhan S. **Robust traceability from trace amounts**. 2015

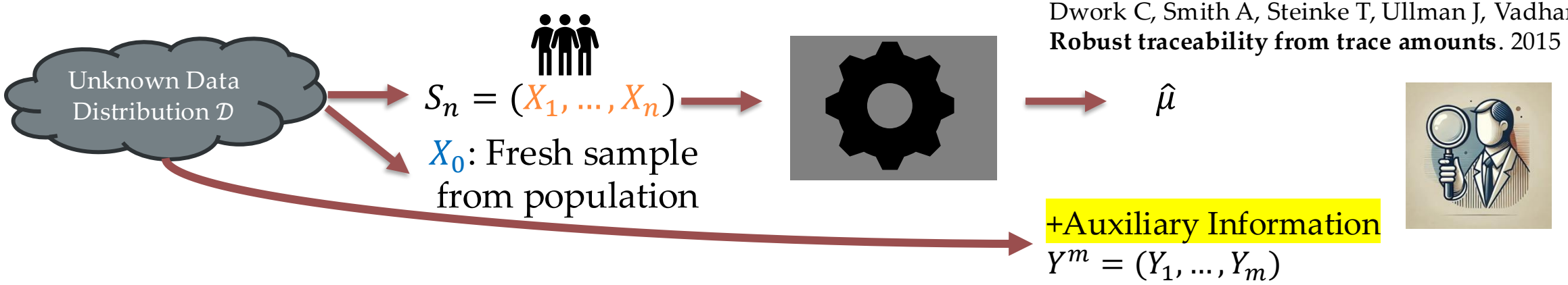
Definition: **minimal m-sample based MIA**

ψ: auxillary information $\times$ output model $\times$ data point $\rightarrow \{IN, OUT\}$

$$\Pr(\psi(Y^m, \hat{\mu}, X_0) = OUT) \geq 0.51$$
$$\forall i \in [n]: \Pr(\psi(Y^m, \hat{\mu}, X_i) = IN) \geq 0.51$$

How many auxiliary samples, i.e., m, required to design an attacker with a slightly better performance than random guessing?

Membership Inference Attacks (MIA)



Dwork C, Smith A, Steinke T, Ullman J, Vadhan S. **Robust traceability from trace amounts**. 2015

Definition: **minimal m-sample based MIA**

ψ: auxillary information $\times$ output model $\times$ data point $\rightarrow \{IN, OUT\}$

$$\Pr(\psi(Y^m, \hat{\mu}, X_0) = OUT) \geq 0.51$$
$$\forall i \in [n]: \Pr(\psi(Y^m, \hat{\mu}, X_i) = IN) \geq 0.51$$

How many auxiliary samples, i.e., m, required to design an attacker with a slightly better performance than random guessing?

How do practical MIAs use auxiliary info?

How do practical MIAs use auxiliary info?

Current MIAs use auxiliary information about the distribution in very limited ways

How do practical MIAs use auxiliary info?

Current MIAs use auxiliary information about the distribution in very limited ways

1. Compute a generic test statistic **distribution-independent**.

How do practical MIAs use auxiliary info?

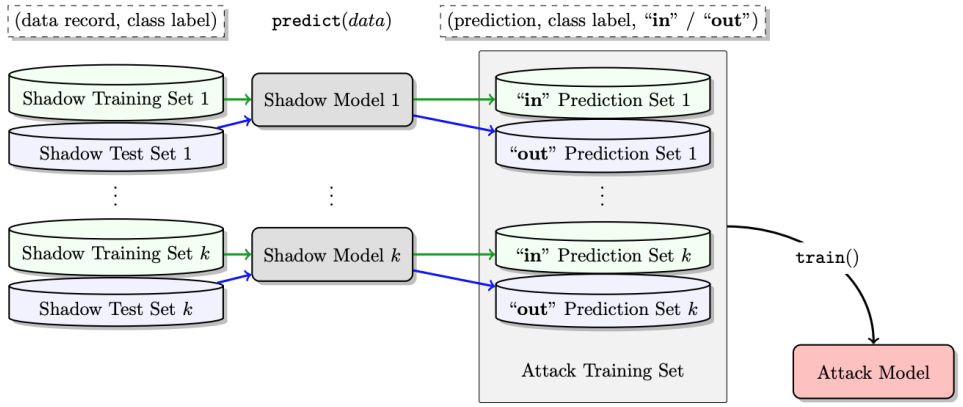
Current MIAs use auxiliary information about the distribution in very limited ways

1. Compute a generic test statistic **distribution-independent**.
2. Apply a **distribution-dependent** threshold
   - Find a good threshold by trying your test on some auxiliary data.

How do practical MIAs use auxiliary info?

Current MIAs use auxiliary information about the distribution in very limited ways

1. Compute a generic test statistic **distribution-independent**.
2. Apply a **distribution-dependent** threshold
   - Find a good threshold by trying your test on some auxiliary data.

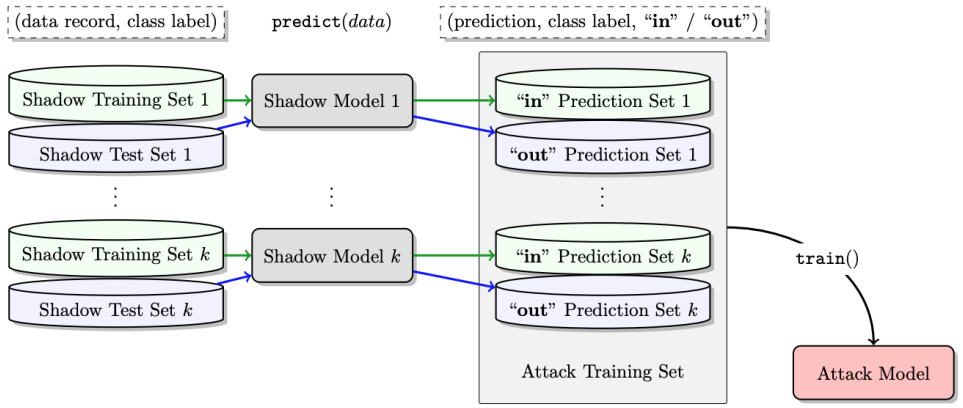


Shokri R, Stronati M, Song C, Shmatikov V. **Membership inference attacks against machine learning models.**

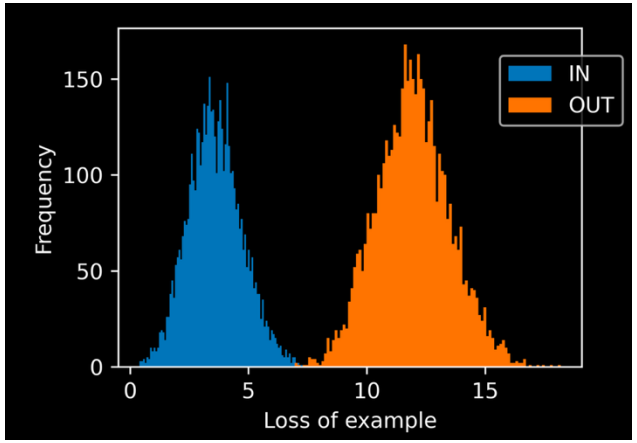
How do practical MIAs use auxiliary info?

Current MIAs use auxiliary information about the distribution in very limited ways

1. Compute a generic test statistic **distribution-independent**.
2. Apply a **distribution-dependent** threshold
   - Find a good threshold by trying your test on some auxiliary data.



Shokri R, Stronati M, Song C, Shmatikov V. **Membership inference attacks against machine learning models.**

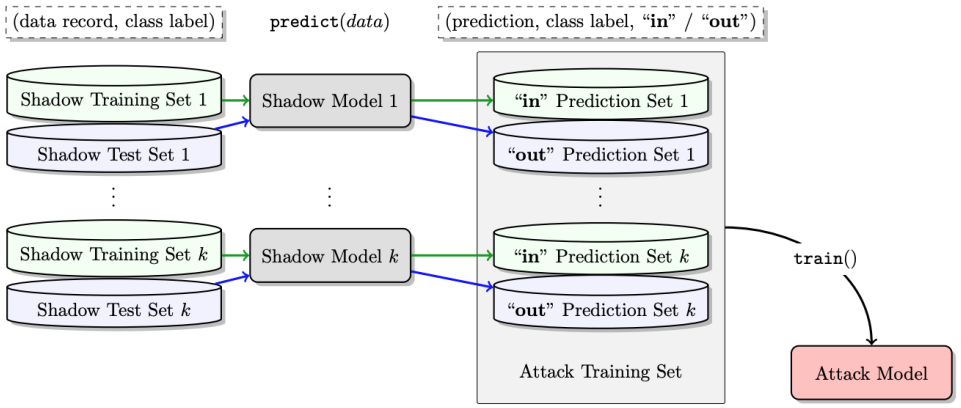


Carlini N, Chien S, Nasr M, Song S, Terzis A, Tramer F. **Membership inference attacks from first principles.**

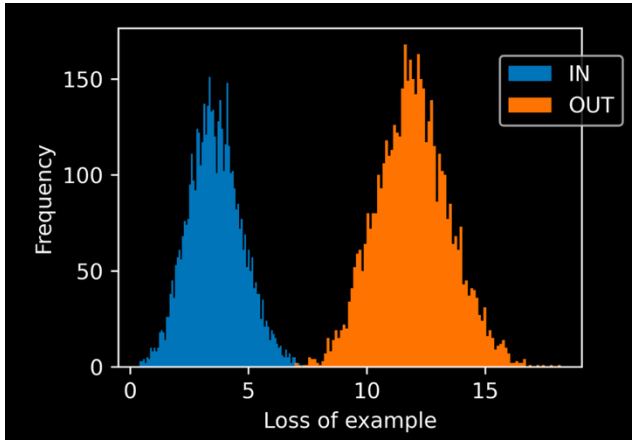
How do practical MIAs use auxiliary info?

Current MIAs use auxiliary information about the distribution in very limited ways

1. Compute a generic test statistic **distribution-independent**.
2. Apply a **distribution-dependent** threshold
   - Find a good threshold by trying your test on some auxiliary data.



Shokri R, Stronati M, Song C, Shmatikov V. **Membership inference attacks against machine learning models.**



Carlini N, Chien S, Nasr M, Song S, Terzis A, Tramer F. **Membership inference attacks from first principles.**

This approach requires at most $O(n)$ where n is the number of training samples.

Is this the best approach to use auxiliary information for building a MIA?

Outline of the Talk

- Membership Inference Attack and Auxiliary Information ✓

Membership Inference Attacks: identify the training samples.
Auxiliary Information: samples from the same population that the training set comes from.
Current paradigm: distribution-free score function and distribution-dependent threshold

- Membership Inference Attack on Gaussian Mean Estimators

- Proof Overview

Mean Estimation Setup and Main Question

Mean Estimation Setup and Main Question

- **Class of Gaussian distributions in $\mathbb{R}^d$:**

Mean Estimation Setup and Main Question

- **Class of Gaussian distributions in $\mathbb{R}^{d}$:**

$$\mathcal{P}_d = \{\mathcal{N}(\mu, \Sigma) : \mu \in \mathbb{R}^d \text{ and } \Sigma \in \mathbb{R}^{d \times d} \text{ is a PD matrix}\}$$

Mean Estimation Setup and Main Question

- **Class of Gaussian distributions in $\mathbb{R}^{d}$**:

$$\mathcal{P}_d = \{\mathcal{N}(\mu, \Sigma): \mu \in \mathbb{R}^d \text{ and } \Sigma \in \mathbb{R}^{d \times d} \text{ is a PD matrix}\}$$

- **(n, ρ)-accurate symmetric mean estimator $\mathcal{A}$**: For every $\mathcal{D} \in \mathcal{P}_d$, given $(X_1, \dots, X_n) \sim \mathcal{D}^n$ and $\hat{\mu} \sim \mathcal{A}(X_1, \cdots, X_n) = \bar{A}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right)$, with a high probability

Mean Estimation Setup and Main Question

- **Class of Gaussian distributions in $\mathbb{R}^{d}$:**

$$\mathcal{P}_d = \{\mathcal{N}(\mu, \Sigma): \mu \in \mathbb{R}^d \text{ and } \Sigma \in \mathbb{R}^{d\times d} \text{ is a PD matrix}\}$$

- **(n, ρ)-accurate symmetric mean estimator $\mathcal{A}$**: For every $\mathcal{D} \in \mathcal{P}_d$, given $(X_1, \dots, X_n) \sim \mathcal{D}^n$ and $\hat{\mu} \sim \mathcal{A}(X_1, \cdots, X_n) = \bar{A}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right)$, with a high probability

$$\|\hat{\mu} - \mu\|_{\Sigma}^2 = \left\|\Sigma^{-\frac{1}{2}}(\hat{\mu} - \mu)\right\|_2^2 \leq \left(\rho^2 + \frac{1}{n}\right) d \longrightarrow \frac{1}{\sqrt{n}} < \rho < 1$$

Mean Estimation Setup and Main Question

- **Class of Gaussian distributions in $\mathbb{R}^{d}$:**

$$\mathcal{P}_d = \{\mathcal{N}(\mu, \Sigma): \mu \in \mathbb{R}^d \text{ and } \Sigma \in \mathbb{R}^{d \times d} \text{ is a PD matrix}\}$$

- **(n, ρ)-accurate symmetric mean estimator $\mathcal{A}$**: For every $\mathcal{D} \in \mathcal{P}_d$, given $(X_1, \dots, X_n) \sim \mathcal{D}^n$ and $\hat{\mu} \sim \mathcal{A}(X_1, \cdots, X_n) = \bar{A}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right)$, with a high probability

$$\|\hat{\mu} - \mu\|_{\Sigma}^2 = \left\|\Sigma^{-\frac{1}{2}}(\hat{\mu} - \mu)\right\|_2^2 \leq \left(\rho^2 + \frac{1}{n}\right) d \longrightarrow \frac{1}{\sqrt{n}} < \rho < 1$$

- **Minimal MIA for Gaussian mean estimator**

Mean Estimation Setup and Main Question

- **Class of Gaussian distributions in $\mathbb{R}^{d}$**:

$$\mathcal{P}_d = \{\mathcal{N}(\mu, \Sigma): \mu \in \mathbb{R}^d \text{ and } \Sigma \in \mathbb{R}^{d\times d} \text{ is a PD matrix}\}$$

- **(n, ρ)-accurate symmetric mean estimator $\mathcal{A}$**: For every $\mathcal{D} \in \mathcal{P}_d$, given $(X_1, \dots, X_n) \sim \mathcal{D}^n$ and $\hat{\mu} \sim \mathcal{A}(X_1, \cdots, X_n) = \bar{A}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right)$, with a high probability

$$\|\hat{\mu} - \mu\|_{\Sigma}^2 = \left\|\Sigma^{-\frac{1}{2}}(\hat{\mu} - \mu)\right\|_2^2 \le \left(\rho^2 + \frac{1}{n}\right) d \longrightarrow \frac{1}{\sqrt{n}} < \rho < 1$$

- **Minimal MIA for Gaussian mean estimator**

False Negative Rate $\Pr(\psi(Y^m, \hat{\mu}, X_0) = OUT) \ge 0.51$

True Positive Rate $\forall i \in [n]$: $\Pr(\psi(Y^m, \hat{\mu}, X_i) = IN) \ge 0.51$

Mean Estimation Setup and Main Question

- **Class of Gaussian distributions in $\mathbb{R}^d$:**

$$\mathcal{P}_d = \{\mathcal{N}(\mu, \Sigma): \mu \in \mathbb{R}^d \text{ and } \Sigma \in \mathbb{R}^{d \times d} \text{ is a PD matrix}\}$$

- **(n, ρ)-accurate symmetric mean estimator $\mathcal{A}$**: For every $\mathcal{D} \in \mathcal{P}_d$, given $(X_1, \ldots, X_n) \sim \mathcal{D}^n$ and $\hat{\mu} \sim \mathcal{A}(X_1, \cdots, X_n) = \bar{A}\left(\frac{1}{n}\sum_{i=1}^n X_i\right)$, with a high probability

$$\|\hat{\mu} - \mu\|_\Sigma^2 = \left\|\Sigma^{-\frac{1}{2}}(\hat{\mu} - \mu)\right\|_2^2 \leq \left(\rho^2 + \frac{1}{n}\right) d \longrightarrow \frac{1}{\sqrt{n}} < \rho < 1$$

- **Minimal MIA for Gaussian mean estimator**

False Negative Rate $\Pr(\psi(Y^m, \hat{\mu}, X_0) = OUT) \geq 0.51$

True Positive Rate $\forall i \in [n]$: $\Pr(\psi(Y^m, \hat{\mu}, X_i) = IN) \geq 0.51$

What is the minimum m as a function of n, d, ρ such that
for every ρ-accurate mean estimator with n samples in $\mathbb{R}^d$,
there exists a strategy for m-sample based MIA that works for every $\mathcal{P}_d$?

Overview of Our Results

Overview of Our Results

- Consider an informed attacker that knows the data distribution: for every (n, ρ)-accurate symmetric mean estimator given $d \gtrsim d^{\star}(n, \rho) \simeq n + n^2\rho^2$ an informed attacker can succeed.

Overview of Our Results

- Consider an informed attacker that knows the data distribution: for every (n, ρ)-accurate symmetric mean estimator given $d \gtrsim d^{\star}(n, \rho) \simeq n + n^2\rho^2$ an informed attacker can succeed.

Assume $d \simeq d^{\star}(n, \rho)$

Overview of Our Results

- Consider an informed attacker that knows the data distribution: for every (n, ρ)-accurate symmetric mean estimator given $d \gtrsim d^{\star}(n, \rho) \simeq n + n^2\rho^2$ an informed attacker can succeed.

Assume $d \simeq d^{\star}(n, \rho)$

- [DSSUV15] shows that if only the mean vector is unknown

Auxiliary samples$\simeq \min\left\{n, \frac{1}{\rho^2}\right\}$

Overview of Our Results

- Consider an informed attacker that knows the data distribution: for every (n, ρ)-accurate symmetric mean estimator given $d \gtrsim d^{\star}(n, \rho) \simeq n + n^2\rho^2$ an informed attacker can succeed.

Assume $d \simeq d^{\star}(n, \rho)$

- [DSSUV15] shows that if only the mean vector is unknown

Auxiliary samples$\simeq \min\left\{n, \frac{1}{\rho^2}\right\}$ → Bonus Result: We show it is optimal!

Overview of Our Results

- Consider an informed attacker that knows the data distribution: for every (n, ρ)-accurate symmetric mean estimator given $d \gtrsim d^\star(n,\rho) \simeq n + n^2\rho^2$ an informed attacker can succeed.

Assume $d \simeq d^\star(n,\rho)$

- [DSSUV15] shows that if only the mean vector is unknown

Auxiliary samples$\simeq \min\left\{n, \frac{1}{\rho^2}\right\}$ → Bonus Result: We show it is optimal!

Main Result: Unknown covariance

Auxiliary samples$\geq n + n^2\rho^2 \simeq d^\star(n,\rho)$ is necessary.

Overview of Our Results

- Consider an informed attacker that knows the data distribution: for every (n, ρ)-accurate symmetric mean estimator given $d \gtrsim d^\star(n,\rho) \simeq n + n^2\rho^2$ an informed attacker can succeed.

Assume $d \simeq d^\star(n,\rho)$

- [DSSUV15] shows that if only the mean vector is unknown

Auxiliary samples$\simeq \min\left\{n, \frac{1}{\rho^2}\right\}$ → Bonus Result: We show it is optimal!

Main Result: Unknown covariance

Auxiliary samples$\geq n + n^2\rho^2 \simeq d^\star(n,\rho)$ is necessary.

Overview of Our Results

- Consider an informed attacker that knows the data distribution: for every (n, ρ)-accurate symmetric mean estimator given $d \gtrsim d^\star(n,\rho) \simeq n + n^2\rho^2$ an informed attacker can succeed.

Assume $d \simeq d^\star(n,\rho)$

- [DSSUV15] shows that if only the mean vector is unknown

Auxiliary samples$\simeq \min\left\{n, \frac{1}{\rho^2}\right\}$ → Bonus Result: We show it is optimal!

Main Result: Unknown covariance

Auxiliary samples$\geq n + n^2\rho^2 \simeq d^\star(n,\rho)$ is necessary.

Implication for practice of MIA

Overview of Our Results

- Consider an informed attacker that knows the data distribution: for every (n, ρ)-accurate symmetric mean estimator given $d \gtrsim d^\star(n,\rho) \simeq n + n^2\rho^2$ an informed attacker can succeed.

Assume $d \simeq d^\star(n,\rho)$

- [DSSUV15] shows that if only the mean vector is unknown

Auxiliary samples$\simeq \min\left\{n, \frac{1}{\rho^2}\right\}$ → Bonus Result: We show it is optimal!

Main Result: Unknown covariance

Auxiliary samples$\geq n + n^2\rho^2 \simeq d^\star(n,\rho)$ is necessary.

Implication for practice of MIA

Choosing a distribution-independent test statistics and finding a distribution-dependent threshold may not be optimal.

Formal Statements

Formal Statements

Formal Statements

Theorem [HSU'25]
Fix n, ρ, d, and assume $d \gtrsim d^{\star}(n,\rho) \triangleq n + n^2\rho^2$, i.e., the dimension is large enough for a fully-informed attacker to succeed. Then, every sample-based MIA for the Gaussian mean estimation problem requires $m \gtrsim n + n^2\rho^2$ samples.

Formal Statements

Theorem [HSU'25]
Fix n, ρ, d, and assume $d \gtrsim d^{\star}(n,\rho) \triangleq n + n^2\rho^2$, i.e., the dimension is large enough for a fully-informed attacker to succeed. Then, every sample-based MIA for the Gaussian mean estimation problem requires $m \gtrsim n + n^2\rho^2$ samples.

Formal Statements

Theorem [HSU'25]
Fix n, ρ, d, and assume $d \gtrsim d^{\star}(n,\rho) \triangleq n + n^2\rho^2$, i.e., the dimension is large enough for a fully-informed attacker to succeed. Then, every sample-based MIA for the Gaussian mean estimation problem requires $m \gtrsim n + n^2\rho^2$ samples.

Theorem [HSU'25]
Fix n, ρ, d and assume that $d \gtrsim d^{\star}(n,\rho) \triangleq n + n^2\rho^2$. Then every sample-based MIA for Gaussian mean estimation with known covariance requires $m \gtrsim \min\{n, 1/\rho^2\}$ samples.

More generally, for n,m, d, ρ, there exists an m-sample-based MIA only if $d \gtrsim n + n^2\rho^2 + \frac{n^2}{m}$.

Formal Statements

Theorem [HSU'25]
Fix n, ρ, d, and assume $d \gtrsim d^\star(n,\rho) \triangleq n + n^2\rho^2$, i.e., the dimension is large enough for a fully-informed attacker to succeed. Then, every sample-based MIA for the Gaussian mean estimation problem requires $m \gtrsim n + n^2\rho^2$ samples.

Theorem [HSU'25]
Fix n, ρ, d and assume that $d \gtrsim d^\star(n,\rho) \triangleq n + n^2\rho^2$. Then every sample-based MIA for Gaussian mean estimation with known covariance requires $m \gtrsim \min\{n, 1/\rho^2\}$ samples.

More generally, for n,m, d, ρ, there exists an m-sample-based MIA only if $d \gtrsim n + n^2\rho^2 + \frac{n^2}{m}$.

Attackers Side Information about the data distribution	Conditions under which an attacker exists
Full Information	$d \gtrsim d^\star(n,\rho) \simeq n + n^2\rho^2$
Known Covariance, Unknown Mean	$d \gtrsim n + n^2\rho^2 + \frac{n^2}{m}$
Uknown Covariance, Uknown Mean	$d \gtrsim d^\star(n,\rho) \simeq n + n^2\rho^2$ $m \gtrsim n + n^2\rho^2$

Outline of the Talk

- Membership Inference Attack and Auxiliary Information ✓

Membership Inference Attacks: identify the training samples.
Auxiliary Information: samples from the same population that the training set comes from.
Current paradigm: distribution-free score function and distribution-dependent threshold

- Membership Inference Attack on Gaussian Mean Estimators ✓

MIA on Gaussian mean estimators requires $\Omega(n^2\rho^2)$ auxillary samples!
This result challenges the current practice of design of MIAs.

- Proof Overview

Overview of Proof: Informed-Attacker

$$\psi_{\mathrm{NP}}(X,\hat{\mu}) = \begin{cases} IN\colon\ \left\langle \Sigma^{-\frac{1}{2}}(X-\mu), \Sigma^{-\frac{1}{2}}(\hat{\mu}-\mu) \right\rangle > c\dfrac{d}{n} \\ OUT\colon\ \left\langle \Sigma^{-\frac{1}{2}}(X-\mu), \Sigma^{-\frac{1}{2}}(\hat{\mu}-\mu) \right\rangle < c\dfrac{d}{n} \end{cases}$$

Overview of Proof: Informed-Attacker

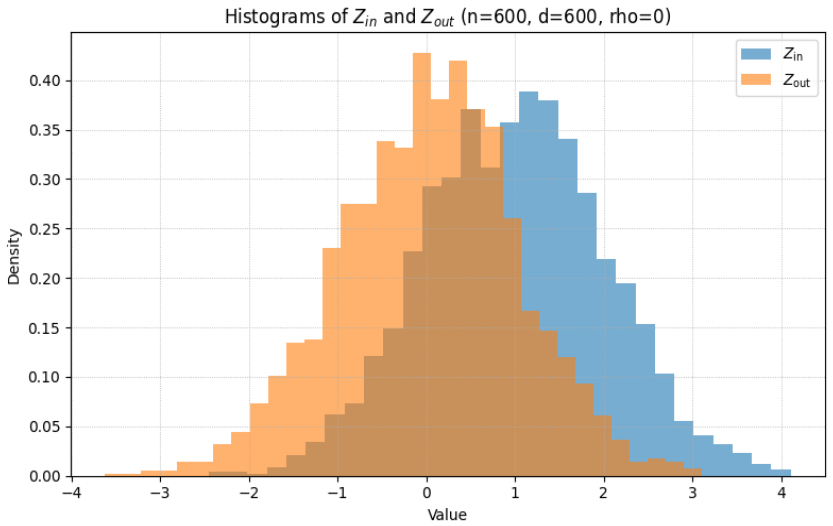
$$\psi_{\mathrm{NP}}(X,\hat{\mu}) = \begin{cases} IN\colon\ \left\langle \Sigma^{-\frac{1}{2}}(X-\mu), \Sigma^{-\frac{1}{2}}(\hat{\mu}-\mu) \right\rangle > c\dfrac{d}{n} \\ OUT\colon\ \left\langle \Sigma^{-\frac{1}{2}}(X-\mu), \Sigma^{-\frac{1}{2}}(\hat{\mu}-\mu) \right\rangle < c\dfrac{d}{n} \end{cases}$$

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i + \rho Z \text{ where } Z \sim N(0,\Sigma)$$

Overview of Proof: Informed-Attacker

$$\psi_{\mathrm{NP}}(X,\hat{\mu}) = \begin{cases} IN\colon\ \left\langle \Sigma^{-\frac{1}{2}}(X-\mu), \Sigma^{-\frac{1}{2}}(\hat{\mu}-\mu) \right\rangle > c\dfrac{d}{n} \\ OUT\colon\ \left\langle \Sigma^{-\frac{1}{2}}(X-\mu), \Sigma^{-\frac{1}{2}}(\hat{\mu}-\mu) \right\rangle < c\dfrac{d}{n} \end{cases}$$

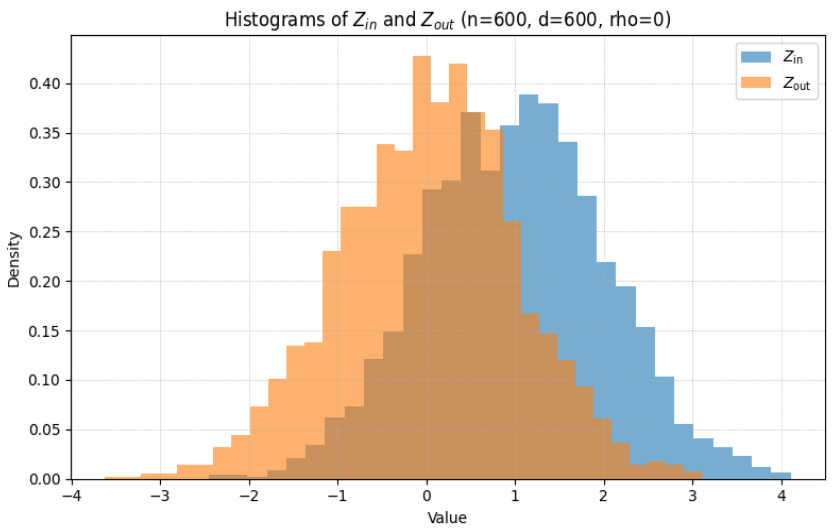
$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i + \rho Z \text{ where } Z \sim N(0,\Sigma)$$

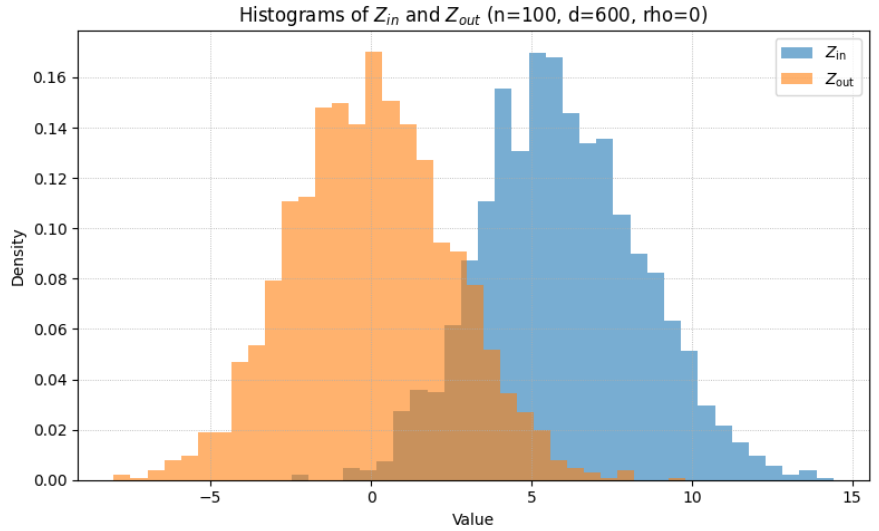


Overview of Proof: Informed-Attacker

$$\psi_{\mathrm{NP}}(X,\hat{\mu}) = \begin{cases} IN\colon \left\langle \Sigma^{-\frac{1}{2}}(X-\mu), \Sigma^{-\frac{1}{2}}(\hat{\mu}-\mu) \right\rangle > c\dfrac{d}{n} \\ OUT\colon \left\langle \Sigma^{-\frac{1}{2}}(X-\mu), \Sigma^{-\frac{1}{2}}(\hat{\mu}-\mu) \right\rangle < c\dfrac{d}{n} \end{cases}$$

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i + \rho Z \text{ where } Z \sim N(0,\Sigma)$$

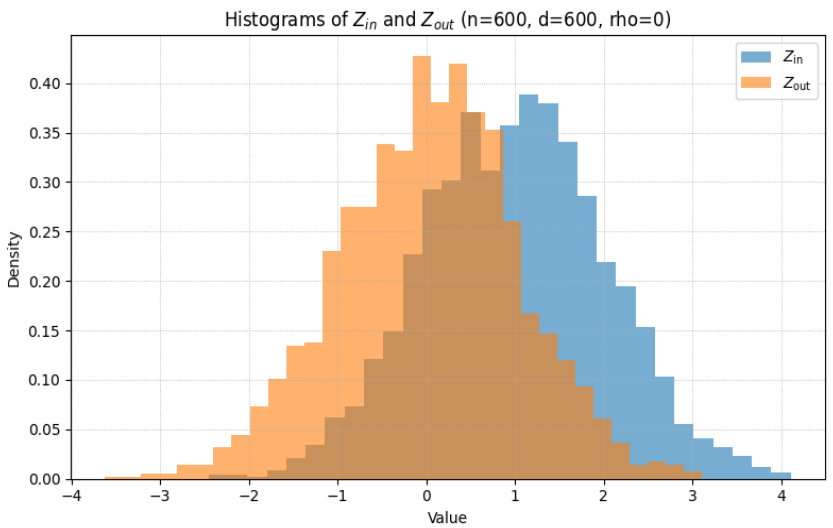


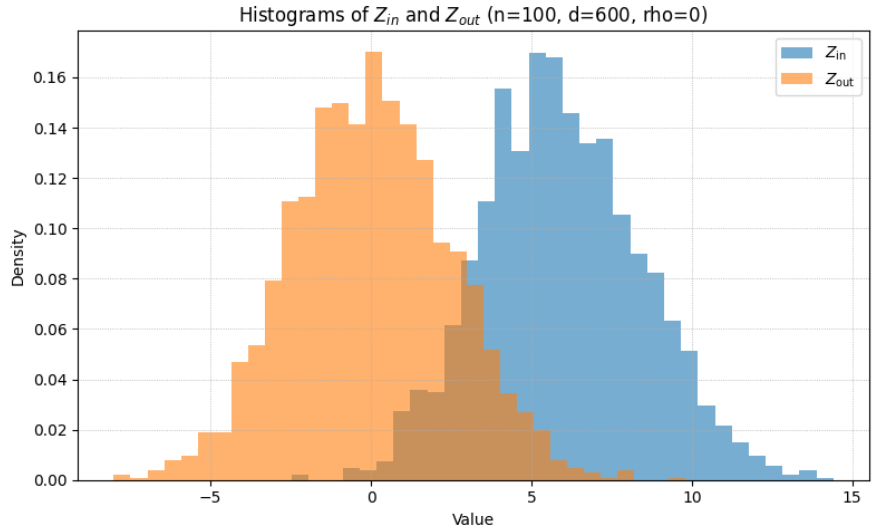


Overview of Proof: Informed-Attacker

$$\psi_{\mathrm{NP}}(X,\hat{\mu}) = \begin{cases} IN\colon \ \left\langle \Sigma^{-\frac{1}{2}}(X-\mu), \Sigma^{-\frac{1}{2}}(\hat{\mu}-\mu) \right\rangle > c\dfrac{d}{n} \\ OUT\colon \ \left\langle \Sigma^{-\frac{1}{2}}(X-\mu), \Sigma^{-\frac{1}{2}}(\hat{\mu}-\mu) \right\rangle < c\dfrac{d}{n} \end{cases}$$

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i + \rho Z \text{ where } Z \sim N(0,\Sigma)$$





$d \gtrsim d^{\star}(n,\rho) \simeq n + n^2\rho^2$ an informed attacker can succeed

Lower bound: analysis of this specific algorithm

Upper bound: $d \gtrsim d^{\star}(n,\rho)$ is enough for every symmetric mean estimators [DSSUV'15]

Overview of Proof: Attempts for MIA

Overview of Proof: Attempts for MIA

- Strategy of an **informed attacker**

Overview of Proof: Attempts for MIA

- Strategy of an **informed attacker**

$$\psi_{\mathrm{NP}}(X,\hat{\mu}) = \begin{cases} IN\colon \left\langle \Sigma^{-\frac{1}{2}}(X-\mu), \Sigma^{-\frac{1}{2}}(\hat{\mu}-\mu) \right\rangle > c\dfrac{d}{n} \\ OUT\colon \left\langle \Sigma^{-\frac{1}{2}}(X-\mu), \Sigma^{-\frac{1}{2}}(\hat{\mu}-\mu) \right\rangle < c\dfrac{d}{n} \end{cases}$$

Overview of Proof: Attempts for MIA

- Strategy of an **informed attacker**

$$\psi_{\mathrm{NP}}(X,\hat{\mu}) = \begin{cases} IN\colon \left\langle \Sigma^{-\frac{1}{2}}(X-\mu), \Sigma^{-\frac{1}{2}}(\hat{\mu}-\mu) \right\rangle > c\dfrac{d}{n} \\ OUT\colon \left\langle \Sigma^{-\frac{1}{2}}(X-\mu), \Sigma^{-\frac{1}{2}}(\hat{\mu}-\mu) \right\rangle < c\dfrac{d}{n} \end{cases}$$

- [DSSUV15] shows that if only mean is unknown, designing MIA is easy: Let $\mu_0 = \frac{1}{m}\sum_{i=2}^{m+1} Y_i$

$$\left\langle \Sigma^{-\frac{1}{2}}(X-\mu), \Sigma^{-\frac{1}{2}}(\hat{\mu}-\mu) \right\rangle \approx \left\langle \Sigma^{-\frac{1}{2}}(X-Y_1), \Sigma^{-\frac{1}{2}}(\hat{\mu}-\mu_0) \right\rangle$$

Overview of Proof: Attempts for MIA

- Strategy of an **informed attacker**

$$\psi_{\mathrm{NP}}(X,\hat{\mu}) = \begin{cases} IN\colon \left\langle \Sigma^{-\frac{1}{2}}(X-\mu), \Sigma^{-\frac{1}{2}}(\hat{\mu}-\mu) \right\rangle > c\dfrac{d}{n} \\ OUT\colon \left\langle \Sigma^{-\frac{1}{2}}(X-\mu), \Sigma^{-\frac{1}{2}}(\hat{\mu}-\mu) \right\rangle < c\dfrac{d}{n} \end{cases}$$

- [DSSUV15] shows that if only mean is unknown, designing MIA is easy: Let $\mu_0 = \frac{1}{m}\sum_{i=2}^{m+1} Y_i$

$$\left\langle \Sigma^{-\frac{1}{2}}(X-\mu), \Sigma^{-\frac{1}{2}}(\hat{\mu}-\mu) \right\rangle \approx \left\langle \Sigma^{-\frac{1}{2}}(X-Y_1), \Sigma^{-\frac{1}{2}}(\hat{\mu}-\mu_0) \right\rangle$$

- Attempt for MIA 1: Provide an estimate of the covariance

Overview of Proof: Attempts for MIA

- Strategy of an **informed attacker**

$$\psi_{\mathrm{NP}}(X,\hat{\mu}) = \begin{cases} IN\colon \ \left\langle \Sigma^{-\frac{1}{2}}(X-\mu), \Sigma^{-\frac{1}{2}}(\hat{\mu}-\mu) \right\rangle > c\dfrac{d}{n} \\ OUT\colon \ \left\langle \Sigma^{-\frac{1}{2}}(X-\mu), \Sigma^{-\frac{1}{2}}(\hat{\mu}-\mu) \right\rangle < c\dfrac{d}{n} \end{cases}$$

- [DSSUV15] shows that if only mean is unknown, designing MIA is easy: Let $\mu_0 = \frac{1}{m}\sum_{i=2}^{m+1} Y_i$

$$\left\langle \Sigma^{-\frac{1}{2}}(X-\mu), \Sigma^{-\frac{1}{2}}(\hat{\mu}-\mu) \right\rangle \approx \left\langle \Sigma^{-\frac{1}{2}}(X-Y_1), \Sigma^{-\frac{1}{2}}(\hat{\mu}-\mu_0) \right\rangle$$

- Attempt for MIA 1: Provide an estimate of the covariance

$$\left\| \Sigma^{-\frac{1}{2}} H \Sigma^{-\frac{1}{2}} - I_d \right\|_{op} \approx 1 \longrightarrow \left\langle \mathrm{H}^{-\frac{1}{2}}(X-\mu), \mathrm{H}^{-\frac{1}{2}}(\hat{\mu}-\mu) \right\rangle \approx \left\langle \Sigma^{-\frac{1}{2}}(X-\mu), \Sigma^{-\frac{1}{2}}(\hat{\mu}-\mu) \right\rangle$$

Overview of Proof: Attempts for MIA

- Strategy of an **informed attacker**

$$\psi_{\mathrm{NP}}(X,\hat{\mu}) = \begin{cases} IN: \ \left\langle \Sigma^{-\frac{1}{2}}(X-\mu), \Sigma^{-\frac{1}{2}}(\hat{\mu}-\mu) \right\rangle > c\dfrac{d}{n} \\ OUT: \ \left\langle \Sigma^{-\frac{1}{2}}(X-\mu), \Sigma^{-\frac{1}{2}}(\hat{\mu}-\mu) \right\rangle < c\dfrac{d}{n} \end{cases}$$

- [DSSUV15] shows that if only mean is unknown, designing MIA is easy: Let $\mu_0 = \frac{1}{m}\sum_{i=2}^{m+1} Y_i$

$$\left\langle \Sigma^{-\frac{1}{2}}(X-\mu), \Sigma^{-\frac{1}{2}}(\hat{\mu}-\mu) \right\rangle \approx \left\langle \Sigma^{-\frac{1}{2}}(X-Y_1), \Sigma^{-\frac{1}{2}}(\hat{\mu}-\mu_0) \right\rangle$$

- Attempt for MIA 1: Provide an estimate of the covariance

$$\left\| \Sigma^{-\frac{1}{2}} H \Sigma^{-\frac{1}{2}} - I_d \right\|_{op} \approx 1 \longrightarrow \left\langle \mathrm{H}^{-\frac{1}{2}}(X-\mu), \mathrm{H}^{-\frac{1}{2}}(\hat{\mu}-\mu) \right\rangle \approx \left\langle \Sigma^{-\frac{1}{2}}(X-\mu), \Sigma^{-\frac{1}{2}}(\hat{\mu}-\mu) \right\rangle$$

Requires $\boldsymbol{m \geq \Omega(d)}$.

Overview of Proof: Attempts for MIA

- Strategy of an **informed attacker**

$$\psi_{\mathrm{NP}}(X,\hat{\mu}) = \begin{cases} IN: \ \left\langle \Sigma^{-\frac{1}{2}}(X-\mu), \Sigma^{-\frac{1}{2}}(\hat{\mu}-\mu) \right\rangle > c\dfrac{d}{n} \\ OUT: \ \left\langle \Sigma^{-\frac{1}{2}}(X-\mu), \Sigma^{-\frac{1}{2}}(\hat{\mu}-\mu) \right\rangle < c\dfrac{d}{n} \end{cases}$$

- [DSSUV15] shows that if only mean is unknown, designing MIA is easy: Let $\mu_0 = \frac{1}{m}\sum_{i=2}^{m+1} Y_i$

$$\left\langle \Sigma^{-\frac{1}{2}}(X-\mu), \Sigma^{-\frac{1}{2}}(\hat{\mu}-\mu) \right\rangle \approx \left\langle \Sigma^{-\frac{1}{2}}(X-Y_1), \Sigma^{-\frac{1}{2}}(\hat{\mu}-\mu_0) \right\rangle$$

- Attempt for MIA 1: Provide an estimate of the covariance

$$\left\| \Sigma^{-\frac{1}{2}} H \Sigma^{-\frac{1}{2}} - I_d \right\|_{op} \approx 1 \longrightarrow \left\langle \mathrm{H}^{-\frac{1}{2}}(X-\mu), \mathrm{H}^{-\frac{1}{2}}(\hat{\mu}-\mu) \right\rangle \approx \left\langle \Sigma^{-\frac{1}{2}}(X-\mu), \Sigma^{-\frac{1}{2}}(\hat{\mu}-\mu) \right\rangle$$

Requires $\boldsymbol{m \geq \Omega(d)}$.

- Attempt for MIA 2: Provide an estimate of the inner product in geometry Σ^{-1}

Overview of Proof: Attempts for MIA

- Strategy of an **informed attacker**

$$\psi_{\mathrm{NP}}(X,\hat{\mu}) = \begin{cases} IN: \ \left\langle \Sigma^{-\frac{1}{2}}(X-\mu), \Sigma^{-\frac{1}{2}}(\hat{\mu}-\mu) \right\rangle > c\frac{d}{n} \\ OUT: \ \left\langle \Sigma^{-\frac{1}{2}}(X-\mu), \Sigma^{-\frac{1}{2}}(\hat{\mu}-\mu) \right\rangle < c\frac{d}{n} \end{cases}$$

- [DSSUV15] shows that if only mean is unknown, designing MIA is easy: Let $\mu_0 = \frac{1}{m}\sum_{i=2}^{m+1} Y_i$

$$\left\langle \Sigma^{-\frac{1}{2}}(X-\mu), \Sigma^{-\frac{1}{2}}(\hat{\mu}-\mu) \right\rangle \approx \left\langle \Sigma^{-\frac{1}{2}}(X-Y_1), \Sigma^{-\frac{1}{2}}(\hat{\mu}-\mu_0) \right\rangle$$

- Attempt for MIA 1: Provide an estimate of the covariance

$$\left\| \Sigma^{-\frac{1}{2}} H \Sigma^{-\frac{1}{2}} - I_d \right\|_{op} \approx 1 \longrightarrow \left\langle \mathrm{H}^{-\frac{1}{2}}(X-\mu), \mathrm{H}^{-\frac{1}{2}}(\hat{\mu}-\mu) \right\rangle \approx \left\langle \Sigma^{-\frac{1}{2}}(X-\mu), \Sigma^{-\frac{1}{2}}(\hat{\mu}-\mu) \right\rangle$$

Requires $\boldsymbol{m \geq \Omega(d)}$.

- Attempt for MIA 2: Provide an estimate of the inner product in geometry Σ^{-1}

For two fixed vectors w_1, w_2 how many samples from $N(0,\Sigma)$ is required to approximate $w_1^T \Sigma^{-1} w_2$?

Overview of Proof: Attempts for MIA

- Strategy of an **informed attacker**

$$\psi_{\mathrm{NP}}(X,\hat{\mu}) = \begin{cases} IN: \ \left\langle \Sigma^{-\frac{1}{2}}(X-\mu), \Sigma^{-\frac{1}{2}}(\hat{\mu}-\mu) \right\rangle > c\frac{d}{n} \\ OUT: \ \left\langle \Sigma^{-\frac{1}{2}}(X-\mu), \Sigma^{-\frac{1}{2}}(\hat{\mu}-\mu) \right\rangle < c\frac{d}{n} \end{cases}$$

- [DSSUV15] shows that if only mean is unknown, designing MIA is easy: Let $\mu_0 = \frac{1}{m}\sum_{i=2}^{m+1} Y_i$

$$\left\langle \Sigma^{-\frac{1}{2}}(X-\mu), \Sigma^{-\frac{1}{2}}(\hat{\mu}-\mu) \right\rangle \approx \left\langle \Sigma^{-\frac{1}{2}}(X-Y_1), \Sigma^{-\frac{1}{2}}(\hat{\mu}-\mu_0) \right\rangle$$

- Attempt for MIA 1: Provide an estimate of the covariance

$$\left\| \Sigma^{-\frac{1}{2}} H \Sigma^{-\frac{1}{2}} - I_d \right\|_{op} \approx 1 \longrightarrow \left\langle \mathrm{H}^{-\frac{1}{2}}(X-\mu), \mathrm{H}^{-\frac{1}{2}}(\hat{\mu}-\mu) \right\rangle \approx \left\langle \Sigma^{-\frac{1}{2}}(X-\mu), \Sigma^{-\frac{1}{2}}(\hat{\mu}-\mu) \right\rangle$$

Requires $\boldsymbol{m \geq \Omega(d)}$.

- Attempt for MIA 2: Provide an estimate of the inner product in geometry Σ^{-1}

For two fixed vectors w_1, w_2 how many samples from $N(0,\Sigma)$ is required to approximate $w_1^T \Sigma^{-1} w_2$?

Our Result:
Requires $\boldsymbol{m \geq \Omega(d)}$.

Intuition Behind Hard Distributions

Intuition Behind Hard Distributions

$$\psi_{\mathrm{NP}}(X,\hat{\mu}) = \begin{cases} IN\colon \ (X-\mu)^{\mathrm{T}}\Sigma^{-1}(\hat{\mu}-\mu) \geq c\dfrac{d}{n} \\ OUT\colon \ (X-\mu)^{\mathrm{T}}\Sigma^{-1}(\hat{\mu}-\mu) < c\dfrac{d}{n} \end{cases}$$

Intuition Behind Hard Distributions

$$\psi_{\mathrm{NP}}(X,\hat{\mu}) = \begin{cases} IN: \quad (X-\mu)^{\mathrm{T}}\Sigma^{-1}(\hat{\mu}-\mu) \geq c\dfrac{d}{n} \\ OUT: \quad (X-\mu)^{\mathrm{T}}\Sigma^{-1}(\hat{\mu}-\mu) < c\dfrac{d}{n} \end{cases}$$

What is the intuition behind $(X-\mu)^T\Sigma^{-1}(\hat{\mu}-\mu)$?

Intuition Behind Hard Distributions

$$\psi_{\mathrm{NP}}(X,\hat{\mu}) = \begin{cases} IN: \quad (X-\mu)^{\mathrm{T}}\Sigma^{-1}(\hat{\mu}-\mu) \geq c\dfrac{d}{n} \\ OUT: \quad (X-\mu)^{\mathrm{T}}\Sigma^{-1}(\hat{\mu}-\mu) < c\dfrac{d}{n} \end{cases}$$

What is the intuition behind $(X-\mu)^T\Sigma^{-1}(\hat{\mu}-\mu)$?

It measures meaningful correlation rather than misleading correlation from high variance directions.

Intuition Behind Hard Distributions

$$\psi_{\mathrm{NP}}(X,\hat{\mu}) = \begin{cases} IN: \quad (X-\mu)^{\mathrm{T}}\Sigma^{-1}(\hat{\mu}-\mu) \geq c\dfrac{d}{n} \\ OUT: \quad (X-\mu)^{\mathrm{T}}\Sigma^{-1}(\hat{\mu}-\mu) < c\dfrac{d}{n} \end{cases}$$

What is the intuition behind $(X-\mu)^T\Sigma^{-1}(\hat{\mu}-\mu)$?

It measures meaningful correlation rather than misleading correlation from high variance directions.

Intuition: sample-based MIA may need to find any directions with high variance in the data!

Intuition Behind Hard Distributions

$$\psi_{\mathrm{NP}}(X,\hat{\mu}) = \begin{cases} IN: \quad (X-\mu)^{\mathrm{T}}\Sigma^{-1}(\hat{\mu}-\mu) \geq c\dfrac{d}{n} \\ OUT: \quad (X-\mu)^{\mathrm{T}}\Sigma^{-1}(\hat{\mu}-\mu) < c\dfrac{d}{n} \end{cases}$$

What is the intuition behind $(X-\mu)^T\Sigma^{-1}(\hat{\mu}-\mu)$?

It measures meaningful correlation rather than misleading correlation from high variance directions.

Intuition: sample-based MIA may need to find any directions with high variance in the data!

$\mathcal{P}_{d,k-\text{spiked}} = \{\mathcal{N}(0,\Sigma) : \Sigma = I_d + \sigma^2 UU^T$ where $U \in \mathbb{R}^{d\times k}$ with orthonormal columns$\}$

Intuition Behind Hard Distributions

$$\psi_{\mathrm{NP}}(X,\hat{\mu}) = \begin{cases} IN: \ \ (X-\mu)^{\mathrm{T}}\Sigma^{-1}(\hat{\mu}-\mu) \geq c\dfrac{d}{n} \\ OUT: \ (X-\mu)^{\mathrm{T}}\Sigma^{-1}(\hat{\mu}-\mu) < c\dfrac{d}{n} \end{cases}$$

What is the intuition behind $(X-\mu)^T\Sigma^{-1}(\hat{\mu}-\mu)$?

It measures meaningful correlation rather than misleading correlation from high variance directions.

Intuition: sample-based MIA may need to find any directions with high variance in the data!

$$\mathcal{P}_{d,k-\mathrm{spiked}} = \{\mathcal{N}(0,\Sigma) : \Sigma = I_d + \sigma^2 UU^T \text{ where } U \in \mathbb{R}^{d\times k} \text{ with orthonormal columns}\}$$

Intuition Behind Hard Distributions

$$\psi_{\mathrm{NP}}(X,\hat{\mu}) = \begin{cases} IN: \ \ (X-\mu)^{\mathrm{T}}\Sigma^{-1}(\hat{\mu}-\mu) \geq c\dfrac{d}{n} \\ OUT: \ (X-\mu)^{\mathrm{T}}\Sigma^{-1}(\hat{\mu}-\mu) < c\dfrac{d}{n} \end{cases}$$

What is the intuition behind $(X-\mu)^T\Sigma^{-1}(\hat{\mu}-\mu)$?

It measures meaningful correlation rather than misleading correlation from high variance directions.

Intuition: sample-based MIA may need to find any directions with high variance in the data!

$$\mathcal{P}_{d,k-\mathrm{spiked}} = \{\mathcal{N}(0,\Sigma) : \Sigma = I_d + \sigma^2 UU^T \text{ where } U \in \mathbb{R}^{d\times k} \text{ with orthonormal columns}\}$$

Difficult algorithm to attack: $\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i + N(0,\Sigma)$

Intuition Behind Hard Distributions

$$\psi_{\mathrm{NP}}(X,\hat{\mu}) = \begin{cases} IN: \ \ (X-\mu)^{\mathrm{T}}\Sigma^{-1}(\hat{\mu}-\mu) \geq c\dfrac{d}{n} \\ OUT: \ (X-\mu)^{\mathrm{T}}\Sigma^{-1}(\hat{\mu}-\mu) < c\dfrac{d}{n} \end{cases}$$

What is the intuition behind $(X-\mu)^T\Sigma^{-1}(\hat{\mu}-\mu)$?

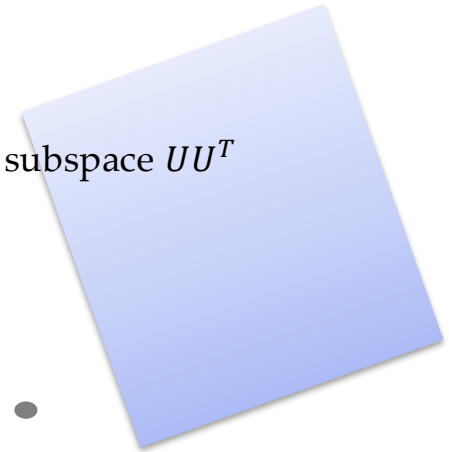
It measures meaningful correlation rather than misleading correlation from high variance directions.

Intuition: sample-based MIA may need to find any directions with high variance in the data!

$$\mathcal{P}_{d,k-\text{spiked}} = \{\mathcal{N}(0,\Sigma) : \Sigma = I_d + \sigma^2 UU^T \text{ where } U \in \mathbb{R}^{d\times k} \text{ with orthonormal columns}\}$$

Difficult algorithm to attack: $\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i + N(0,\Sigma)$

Noisy subspace UU^T



Intuition Behind Hard Distributions

$$\psi_{\mathrm{NP}}(X,\hat{\mu}) = \begin{cases} IN: \ \ (X-\mu)^{\mathrm{T}}\Sigma^{-1}(\hat{\mu}-\mu) \geq c\dfrac{d}{n} \\ OUT: \ (X-\mu)^{\mathrm{T}}\Sigma^{-1}(\hat{\mu}-\mu) < c\dfrac{d}{n} \end{cases}$$

What is the intuition behind $(X-\mu)^T\Sigma^{-1}(\hat{\mu}-\mu)$?

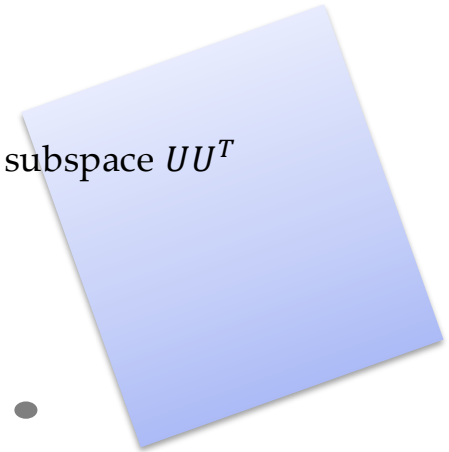
It measures meaningful correlation rather than misleading correlation from high variance directions.

Intuition: sample-based MIA may need to find any directions with high variance in the data!

$$\mathcal{P}_{d,k-\text{spiked}} = \{\mathcal{N}(0,\Sigma) : \Sigma = I_d + \sigma^2 UU^T \text{ where } U \in \mathbb{R}^{d\times k} \text{ with orthonormal columns}\}$$

Difficult algorithm to attack: $\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i + N(0,\Sigma)$

Noisy subspace UU^T



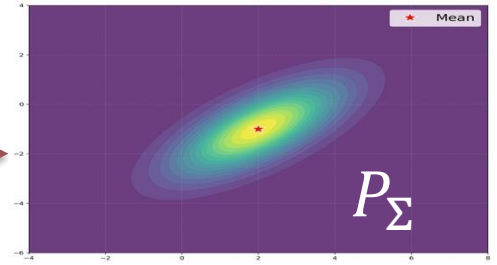
Main Challenge:
We want to show the attacker needs to identify this hidden subspace
How to show it for every sample-based MIA?

Approach to Prove the Lower Bound

Approach to Prove the Lower Bound

Dist. π on $\mathcal{P}_{d,k-\text{spiked}}$:
Uniformly subspace

$\Sigma \sim \pi$
$P_\Sigma = \mathcal{N}(0, \Sigma)$



+ Auxiliary Information

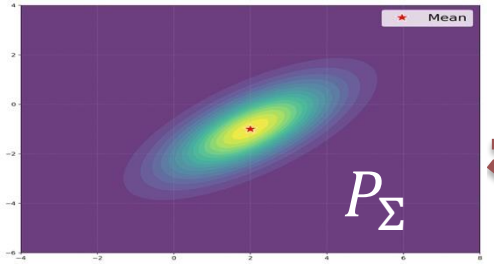
$$\mathcal{P}_{d,k-\text{spiked}} = \{\mathcal{N}(0,\Sigma) : \Sigma = I_d + \sigma^2 UU^T \text{ where } U \in \mathbb{R}^{d\times k} \text{ has orthonormal columns}\}$$

Approach to Prove the Lower Bound

Dist. π on $\mathcal{P}_{d,k-\text{spiked}}$: Uniformly subspace

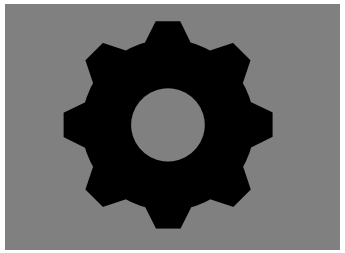
$\Sigma \sim \pi$

$P_\Sigma = \mathcal{N}(0, \Sigma)$



$S_n = (X_1, \dots, X_n)$

X_0



$\hat{\mu}$

+ Auxiliary Information

Auxiallry Information $Y^m = (Y_1, \dots, Y_m)$

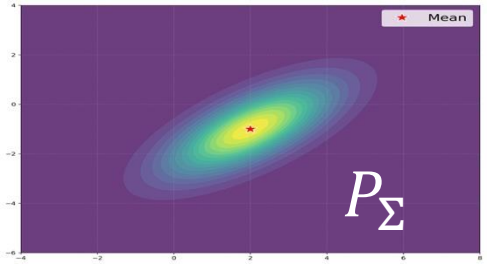
$$\mathcal{P}_{d,k-\text{spiked}} = \{\mathcal{N}(0, \Sigma) : \Sigma = I_d + \sigma^2 UU^T \text{ where } U \in \mathbb{R}^{d \times k} \text{ has orthonormal columns}\}$$

Approach to Prove the Lower Bound

Dist. π on $\mathcal{P}_{d,k-\text{spiked}}$: Uniformly subspace

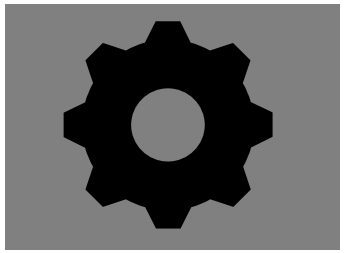
$\Sigma \sim \pi$

$P_\Sigma = \mathcal{N}(0, \Sigma)$



$S_n = (X_1, \dots, X_n)$

X_0



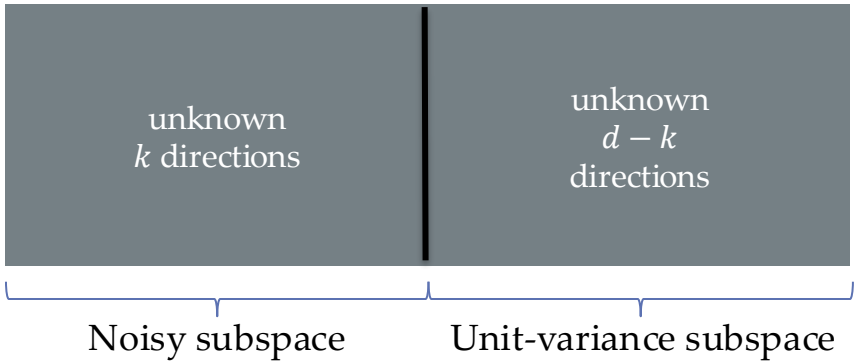
$\hat{\mu}$

+ Auxiliary Information

Auxiallry Information $Y^m = (Y_1, \dots, Y_m)$

$\mathcal{P}_{d,k-\text{spiked}} = \{\mathcal{N}(0, \Sigma) : \Sigma = I_d + \sigma^2 UU^T \text{ where } U \in \mathbb{R}^{d \times k} \text{ has orthonormal columns}\}$

$\mathbb{R}^d =$

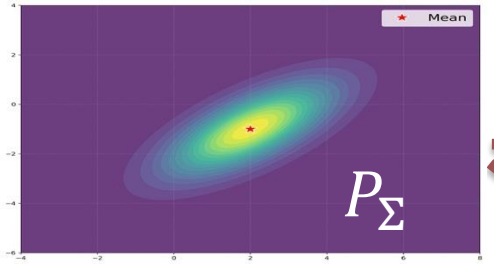


Approach to Prove the Lower Bound

Dist. π on $\mathcal{P}_{d,k-\text{spiked}}$: Uniformly subspace

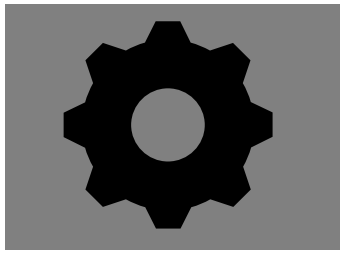
$\Sigma \sim \pi$

$P_\Sigma = \mathcal{N}(0, \Sigma)$



$S_n = (X_1, \dots, X_n)$

X_0



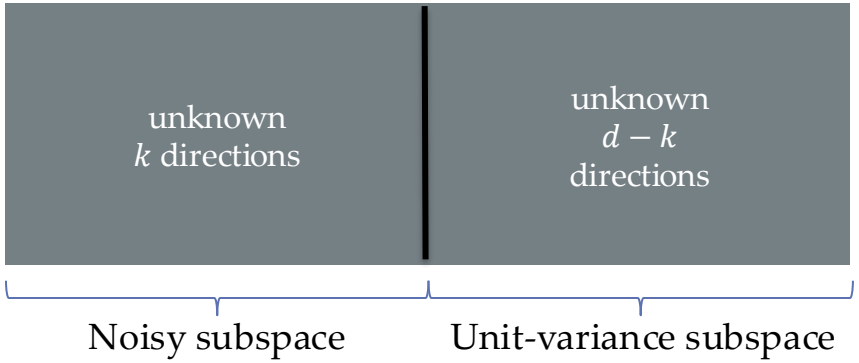
$\hat{\mu}$

+ Auxiliary Information

Auxiallry Information $Y^m = (Y_1, \dots, Y_m)$

$\mathcal{P}_{d,k-\text{spiked}} = \{\mathcal{N}(0, \Sigma) : \Sigma = I_d + \sigma^2 UU^T \text{ where } U \in \mathbb{R}^{d \times k} \text{ has orthonormal columns}\}$

$\mathbb{R}^d =$



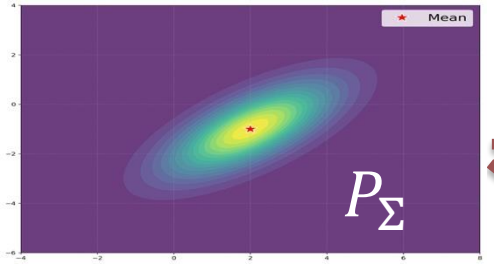
Fix a mean estimator algorithm. The hardness of MIA is controlled by:

Approach to Prove the Lower Bound

Dist. π on $\mathcal{P}_{d,k-\text{spiked}}$: Uniformly subspace

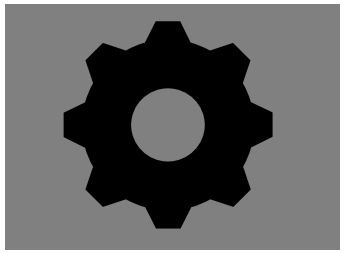
$\Sigma \sim \pi$

$P_\Sigma = \mathcal{N}(0, \Sigma)$



$S_n = (X_1, \dots, X_n)$

X_0



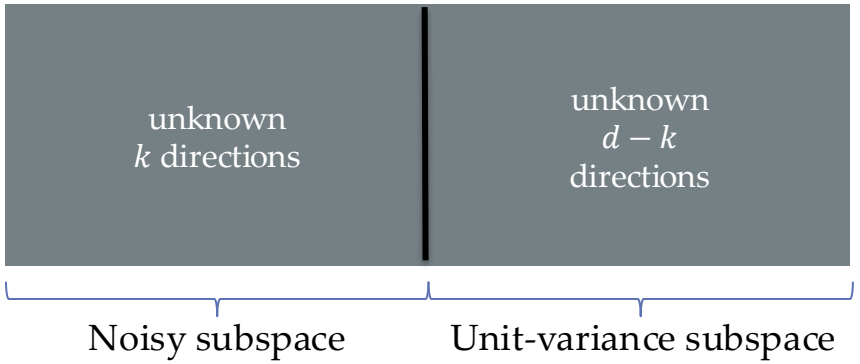
$\hat{\mu}$

+ Auxiliary Information

Auxiallry Information $Y^m = (Y_1, \dots, Y_m)$

$\mathcal{P}_{d,k-\text{spiked}} = \{\mathcal{N}(0, \Sigma) : \Sigma = I_d + \sigma^2 UU^T \text{ where } U \in \mathbb{R}^{d \times k} \text{ has orthonormal columns}\}$

$\mathbb{R}^d =$



Fix a mean estimator algorithm. The hardness of MIA is controlled by:

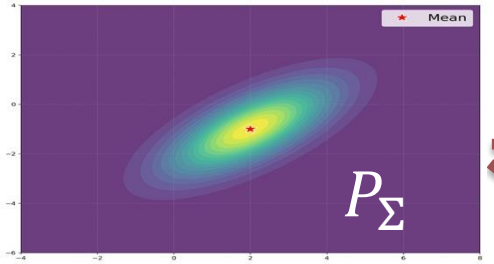
$$TV\left(P_{OUT,\pi} \middle\| P_{IN,\pi}\right)$$

Approach to Prove the Lower Bound

Dist. π on $\mathcal{P}_{d,k-\text{spiked}}$: Uniformly subspace

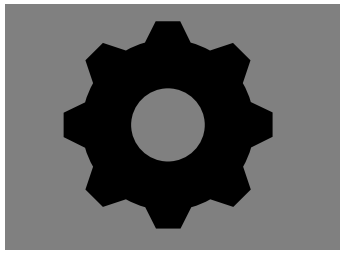
$\Sigma \sim \pi$

$P_\Sigma = \mathcal{N}(0, \Sigma)$



$S_n = (X_1, \dots, X_n)$

X_0



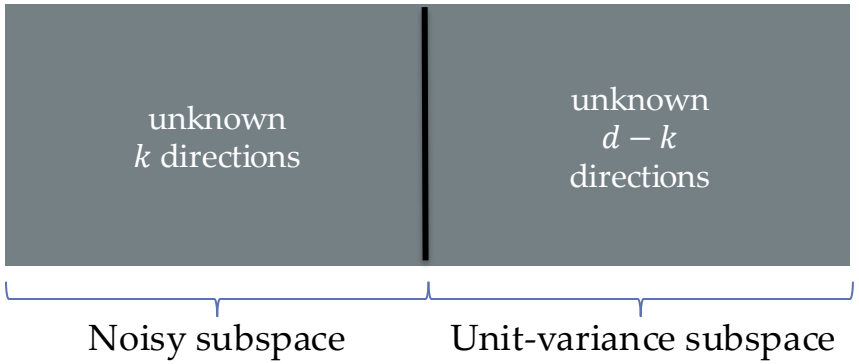
$\hat{\mu}$

+ Auxiliary Information

Auxiallry Information $Y^m = (Y_1, \dots, Y_m)$

$\mathcal{P}_{d,k-\text{spiked}} = \{\mathcal{N}(0, \Sigma) : \Sigma = I_d + \sigma^2 UU^T \text{ where } U \in \mathbb{R}^{d \times k} \text{ has orthonormal columns}\}$

$\mathbb{R}^d =$



Fix a mean estimator algorithm. The hardness of MIA is controlled by:

$$TV\left(P_{OUT,\pi} \middle\| P_{IN,\pi}\right)$$

$P_{OUT,\pi} = \mathbb{E}_{\Sigma\sim\pi}[P_\Sigma(Y^m, \hat{\mu}, X_0)]$

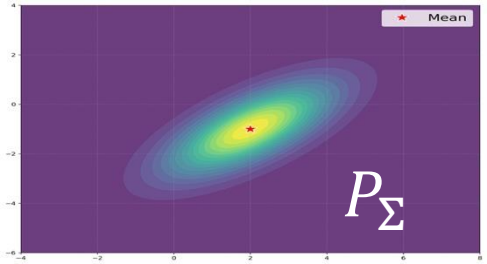
$P_{IN,\pi} = \mathbb{E}_{\Sigma\sim\pi}[P_\Sigma(Y^m, \hat{\mu}, X_1)]$

Approach to Prove the Lower Bound

Dist. π on $\mathcal{P}_{d,k-\text{spiked}}$: Uniformly subspace

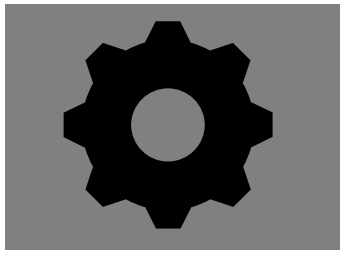
$\Sigma \sim \pi$

$P_\Sigma = \mathcal{N}(0, \Sigma)$



$S_n = (X_1, \dots, X_n)$

X_0



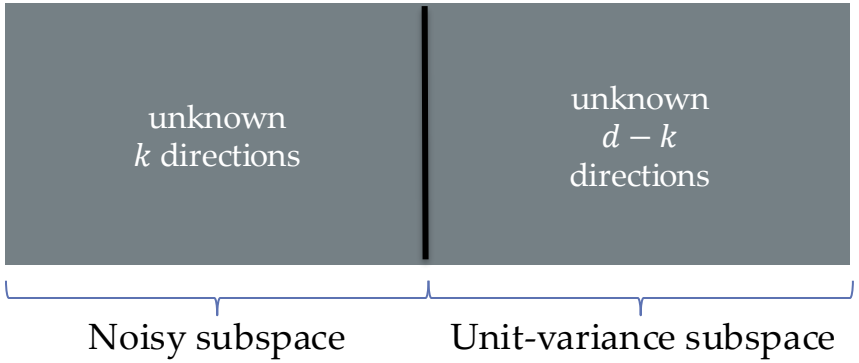
$\hat{\mu}$

+ Auxiliary Information

Auxiallry Information $Y^m = (Y_1, \dots, Y_m)$

$\mathcal{P}_{d,k-\text{spiked}} = \{\mathcal{N}(0, \Sigma) : \Sigma = I_d + \sigma^2 UU^T \text{ where } U \in \mathbb{R}^{d \times k} \text{ has orthonormal columns}\}$

$\mathbb{R}^d =$



Fix a mean estimator algorithm. The hardness of MIA is controlled by:

$TV(P_{OUT,\pi} \| P_{IN,\pi})$

$P_{OUT,\pi} = \mathbb{E}_{\Sigma \sim \pi}[P_\Sigma(Y^m, \hat{\mu}, X_0)]$

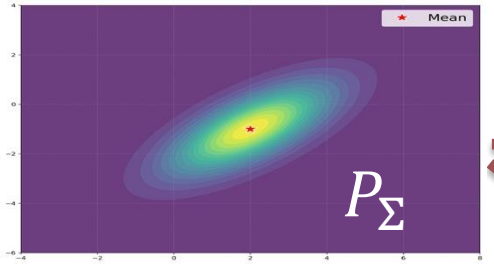
$P_{IN,\pi} = \mathbb{E}_{\Sigma \sim \pi}[P_\Sigma(Y^m, \hat{\mu}, X_1)]$

Approach to Prove the Lower Bound

Dist. π on $\mathcal{P}_{d,k-\text{spiked}}$: Uniformly subspace

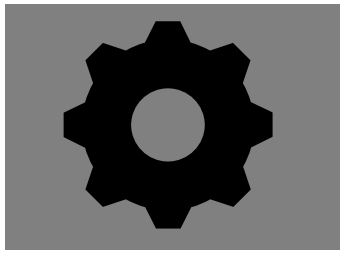
$\Sigma \sim \pi$

$P_\Sigma = \mathcal{N}(0, \Sigma)$



$S_n = (X_1, \dots, X_n)$

X_0



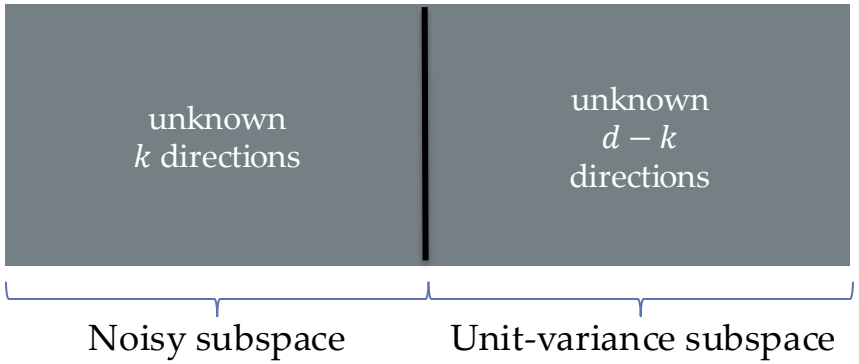
$\hat{\mu}$

+ Auxiliary Information

Auxiallry Information $Y^m = (Y_1, \dots, Y_m)$

$\mathcal{P}_{d,k-\text{spiked}} = \{\mathcal{N}(0, \Sigma) : \Sigma = I_d + \sigma^2 UU^T \text{ where } U \in \mathbb{R}^{d \times k} \text{ has orthonormal columns}\}$

$\mathbb{R}^d =$



Fix a mean estimator algorithm. The hardness of MIA is controlled by:

$TV(P_{OUT,\pi} \| P_{IN,\pi})$

$P_{OUT,\pi} = \mathbb{E}_{\Sigma\sim\pi}[P_\Sigma(Y^m, \hat{\mu}, X_0)]$

$P_{IN,\pi} = \mathbb{E}_{\Sigma\sim\pi}[P_\Sigma(Y^m, \hat{\mu}, X_1)]$

Main Technical Challenge: Analyzing TV distance between mixtures is challenging.

Step 1: Reduction to Zero Auxiliary Sample

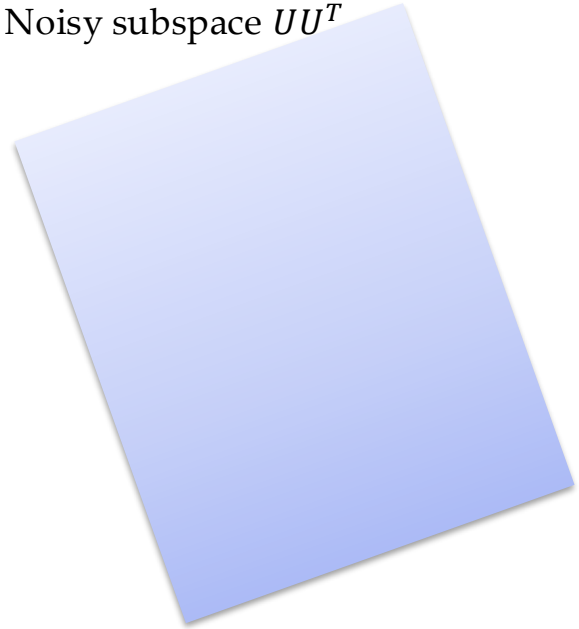
Step 1: Reduction to Zero Auxiliary Sample

$$\mathcal{P}_{d,k-\text{spiked}} = \{\mathcal{N}(0,\Sigma) : \Sigma = I_d + \sigma^2 UU^T \text{ where } U \in \mathbb{R}^{d\times k} \text{ with orthonormal columns}\}$$

Step 1: Reduction to Zero Auxiliary Sample

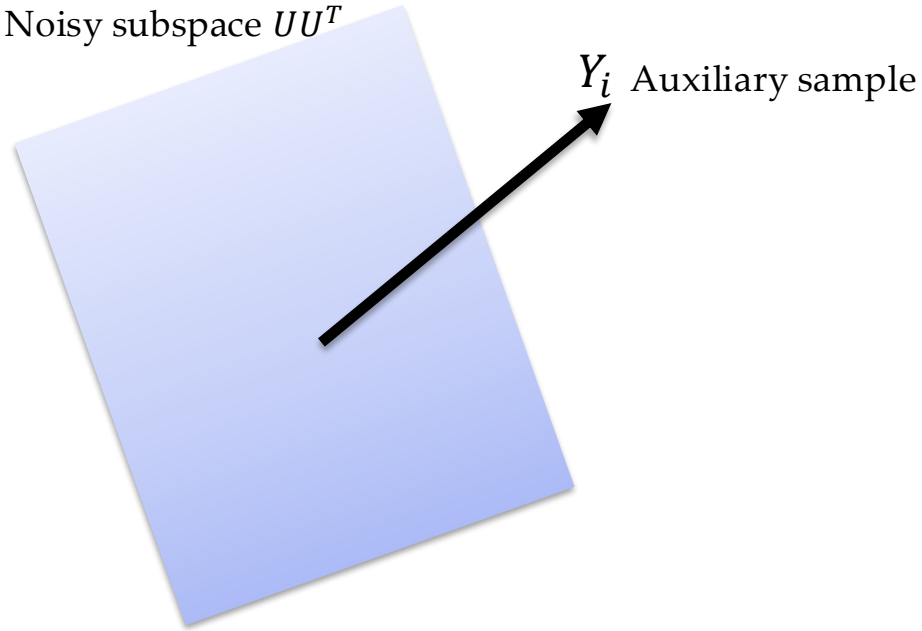
$$\mathcal{P}_{d,k-\text{spiked}} = \{\mathcal{N}(0,\Sigma) : \Sigma = I_d + \sigma^2 UU^T \text{ where } U \in \mathbb{R}^{d\times k} \text{ with orthonormal columns}\}$$

Noisy subspace UU^T



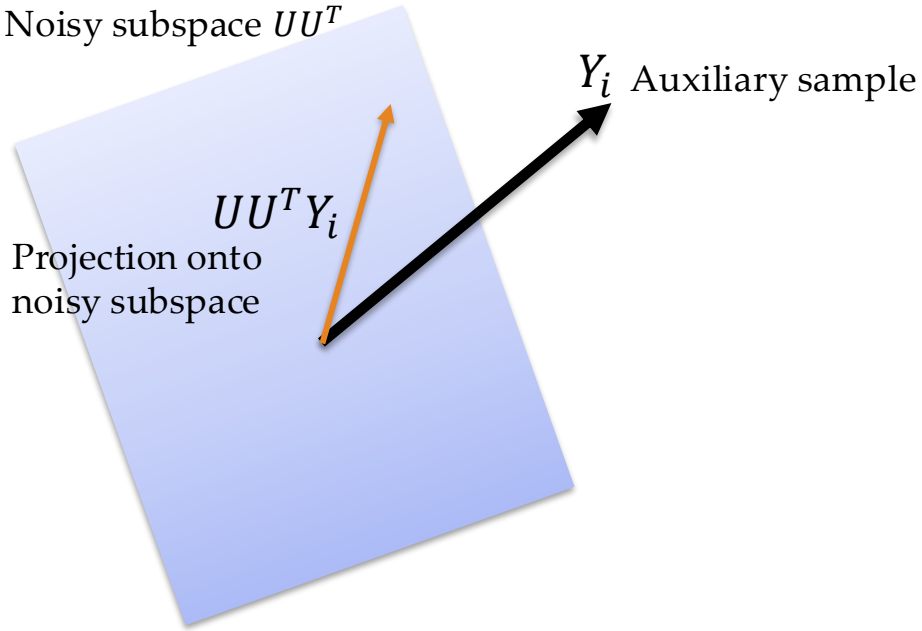
Step 1: Reduction to Zero Auxiliary Sample

$$\mathcal{P}_{d,k-\text{spiked}} = \{\mathcal{N}(0,\Sigma) : \Sigma = I_d + \sigma^2 UU^T \text{ where } U \in \mathbb{R}^{d\times k} \text{ with orthonormal columns}\}$$



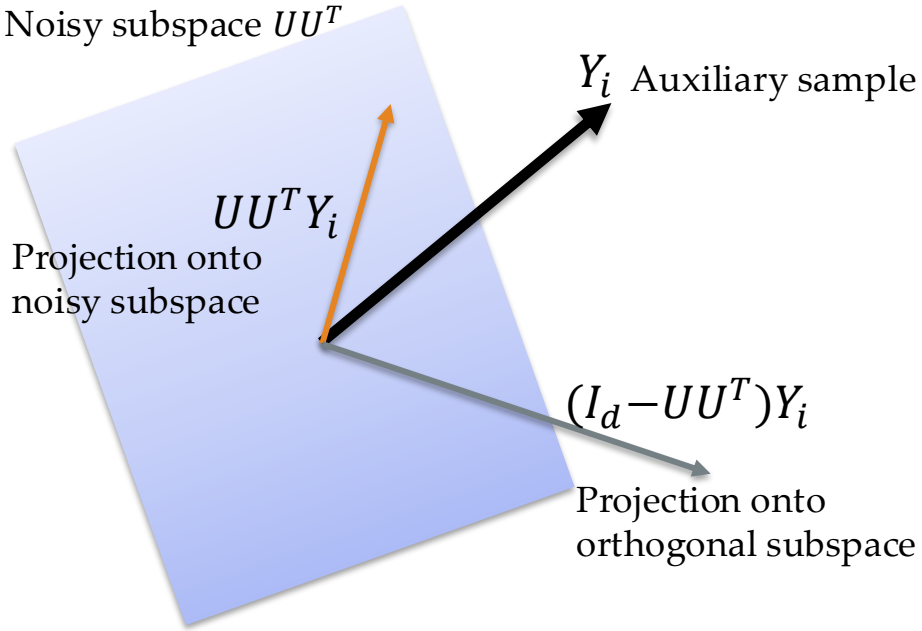
Step 1: Reduction to Zero Auxiliary Sample

$$\mathcal{P}_{d,k-\text{spiked}} = \{\mathcal{N}(0,\Sigma) : \Sigma = I_d + \sigma^2 UU^T \text{ where } U \in \mathbb{R}^{d\times k} \text{ with orthonormal columns}\}$$



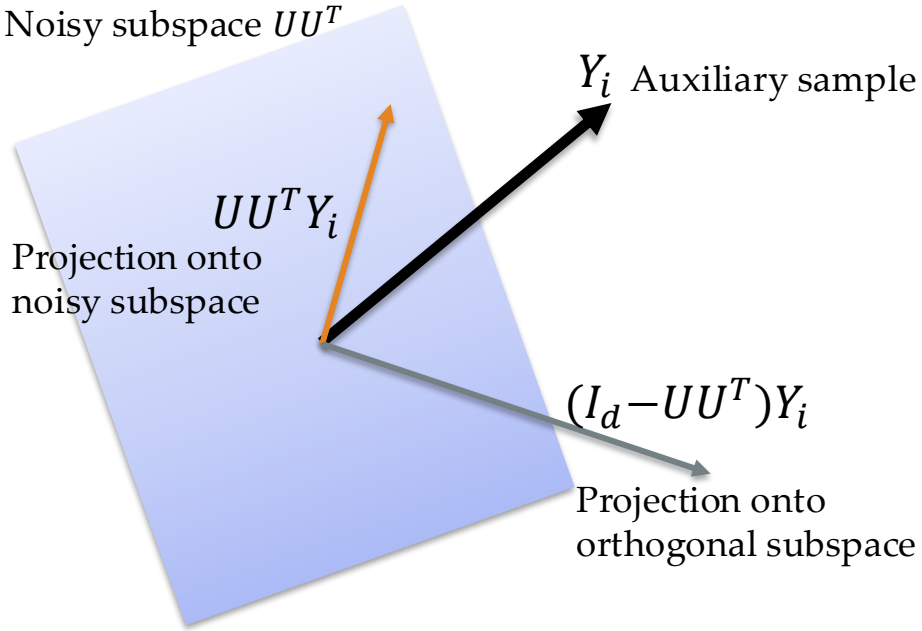
Step 1: Reduction to Zero Auxiliary Sample

$$\mathcal{P}_{d,k-\text{spiked}} = \{\mathcal{N}(0,\Sigma) : \Sigma = I_d + \sigma^2 UU^T \text{ where } U \in \mathbb{R}^{d\times k} \text{ with orthonormal columns}\}$$



Step 1: Reduction to Zero Auxiliary Sample

$$\mathcal{P}_{d,k-\text{spiked}} = \{\mathcal{N}(0,\Sigma) : \Sigma = I_d + \sigma^2 UU^T \text{ where } U \in \mathbb{R}^{d\times k} \text{ with orthonormal columns}\}$$

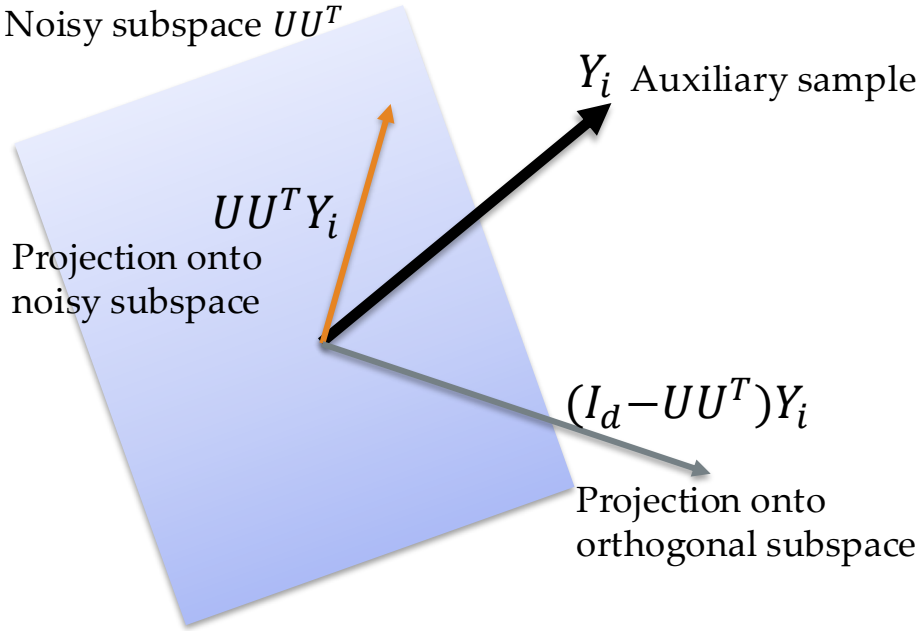


Assume MIA also knows $\{UU^TY_i\}_{i\in[m]}$

Informally, it can identify m directions in the noisy subspace and m directions in unit-variance subspace.

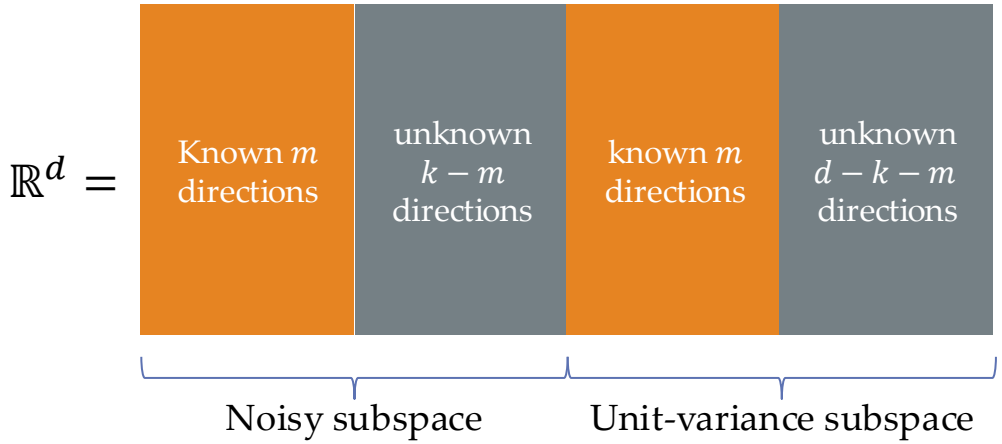
Step 1: Reduction to Zero Auxiliary Sample

$$\mathcal{P}_{d,k-\text{spiked}} = \{\mathcal{N}(0,\Sigma) : \Sigma = I_d + \sigma^2 UU^T \text{ where } U \in \mathbb{R}^{d\times k} \text{ with orthonormal columns}\}$$



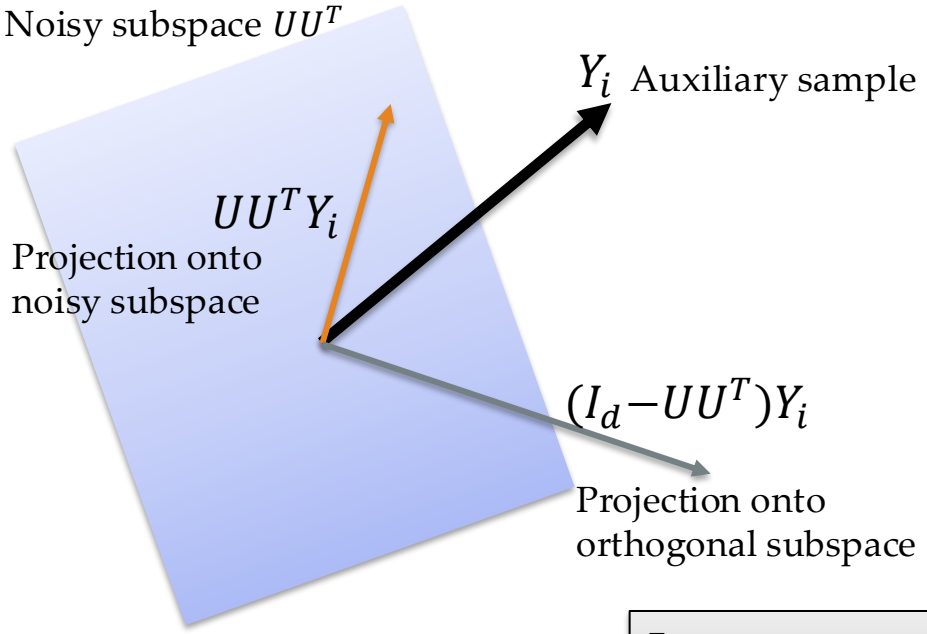
Assume MIA also knows $\{UU^T Y_i\}_{i\in[m]}$

Informally, it can identify m directions in the noisy subspace and m directions in unit-variance subspace.



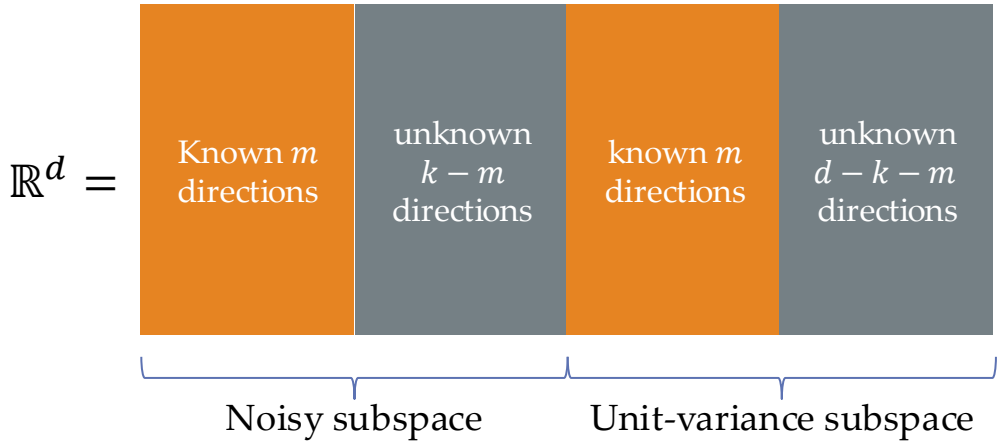
Step 1: Reduction to Zero Auxiliary Sample

$$\mathcal{P}_{d,k-\text{spiked}} = \{\mathcal{N}(0,\Sigma) : \Sigma = I_d + \sigma^2 UU^T \text{ where } U \in \mathbb{R}^{d\times k} \text{ with orthonormal columns}\}$$



Assume MIA also knows $\{UU^T Y_i\}_{i\in[m]}$

Informally, it can identify m directions in the noisy subspace and m directions in unit-variance subspace.

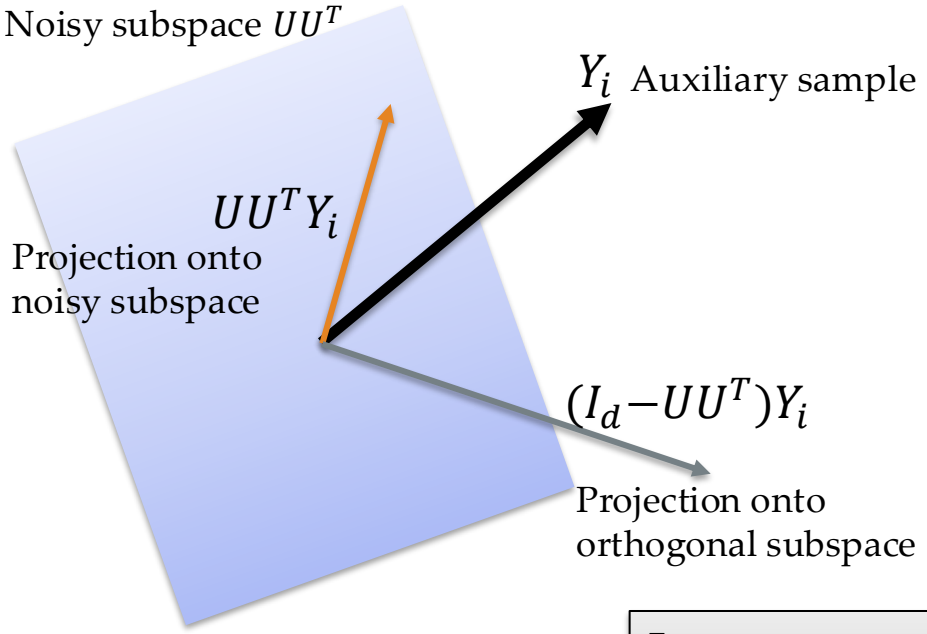


Lemma

$$TV\left(P_{OUT,\pi} \middle\| P_{IN,\pi}\right) \le \sqrt{\frac{m}{n+n^2\rho^2}} + TV\left(\widetilde{X_0}, \tilde{\hat{\mu}} \middle\| \widetilde{X_1}, \tilde{\hat{\mu}}\right)$$

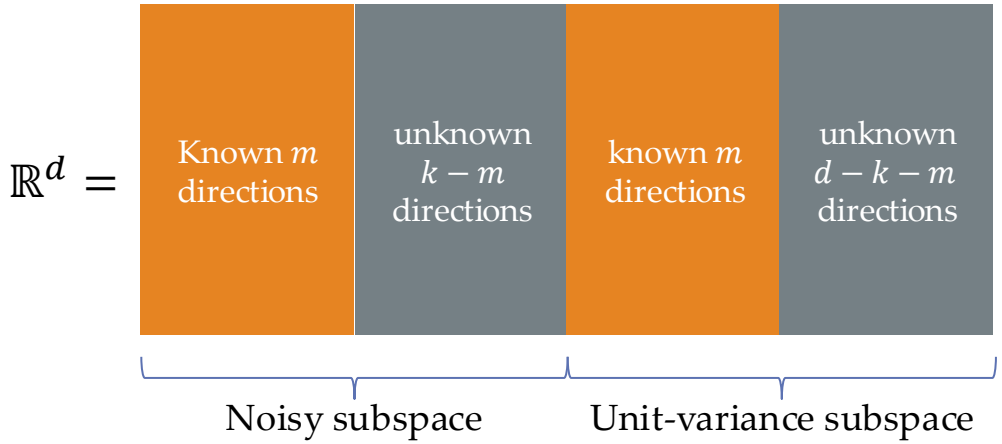
Step 1: Reduction to Zero Auxiliary Sample

$$\mathcal{P}_{d,k-\text{spiked}} = \{\mathcal{N}(0,\Sigma) : \Sigma = I_d + \sigma^2 UU^T \text{ where } U \in \mathbb{R}^{d\times k} \text{ with orthonormal columns}\}$$



Assume MIA also knows $\{UU^T Y_i\}_{i\in[m]}$

Informally, it can identify m directions in the noisy subspace and m directions in unit-variance subspace.



Lemma

$$TV\left(P_{OUT,\pi} \| P_{IN,\pi}\right) \le \sqrt{\frac{m}{n+n^2\rho^2}} + TV\left(\widetilde{X_0}, \tilde{\mu} \,\|\, \widetilde{X_1}, \tilde{\mu}\right)$$

(mixture) MI problem in $\mathbb{R}^{d-2m}$ where $k-m$ directions of covariance matrix is noisy, i.e., $\mathcal{P}_{d-2m,(k-m)-spiked}$

Step 2: Zero Auxiliary Sample Case

Step 2: Zero Auxiliary Sample Case

$$\mathcal{P}_{d-2m,k-m-\text{spiked}} = \left\{\mathcal{N}(0,\Sigma) : \Sigma = I_{d-2m} + \sigma^2 UU^T \text{ where } U \in \mathbb{R}^{(d-2m)\times(k-m)} \text{ with orthonormal columns}\right\}$$

Step 2: Zero Auxiliary Sample Case

$$\mathcal{P}_{d-2m,k-m-\text{spiked}} = \left\{\mathcal{N}(0,\Sigma) : \Sigma = I_{d-2m} + \sigma^2 UU^T \text{ where } U \in \mathbb{R}^{(d-2m)\times(k-m)} \text{ with orthonormal columns}\right\}$$

$TV(X_0, \hat{\mu}\,; X_1, \hat{\mu})$ is the distance between two mixtures.

Step 2: Zero Auxiliary Sample Case

$$\mathcal{P}_{d-2m,k-m-\text{spiked}} = \left\{\mathcal{N}(0,\Sigma) : \Sigma = I_{d-2m} + \sigma^2 UU^T \text{ where } U \in \mathbb{R}^{(d-2m)\times(k-m)} \text{ with orthonormal columns}\right\}$$

$TV(X_0, \hat{\mu}\,; X_1, \hat{\mu})$ is the distance between two mixtures.

Lemma: Sufficient Statistics for Zero-Side Info. Case

$$TV(X_0, \hat{\mu}\,; X_1, \hat{\mu}) = TV(\|X_0\|, \|\hat{\mu}\|, \langle X_0, \hat{\mu}\rangle; \|X_1\|, \|\hat{\mu}\|, \langle X_1, \hat{\mu}\rangle)$$

Step 2: Zero Auxiliary Sample Case

$$\mathcal{P}_{d-2m,k-m-\text{spiked}} = \left\{\mathcal{N}(0,\Sigma) : \Sigma = I_{d-2m} + \sigma^2 UU^T \text{ where } U \in \mathbb{R}^{(d-2m)\times(k-m)} \text{ with orthonormal columns}\right\}$$

$TV(X_0, \hat{\mu}\,; X_1, \hat{\mu})$ is the distance between two mixtures.

Lemma: Sufficient Statistics for Zero-Side Info. Case

$$TV(X_0, \hat{\mu}\,; X_1, \hat{\mu}) = TV(\|X_0\|, \|\hat{\mu}\|, \langle X_0, \hat{\mu}\rangle; \|X_1\|, \|\hat{\mu}\|, \langle X_1, \hat{\mu}\rangle)$$

Idea: for $j \in \{0,1\}$, $(X_j, \hat{\mu})$ has the same distribution as $(RX_j, R\hat{\mu})$ for every rotation matrix R

Step 2: Zero Auxiliary Sample Case

$$\mathcal{P}_{d-2m,k-m-\text{spiked}} = \left\{\mathcal{N}(0,\Sigma) : \Sigma = I_{d-2m} + \sigma^2 UU^T \text{ where } U \in \mathbb{R}^{(d-2m)\times(k-m)} \text{ with orthonormal columns}\right\}$$

$TV(X_0, \hat{\mu}\,; X_1, \hat{\mu})$ is the distance between two mixtures.

Lemma: Sufficient Statistics for Zero-Side Info. Case

$$TV(X_0, \hat{\mu}\,; X_1, \hat{\mu}) = TV(\|X_0\|, \|\hat{\mu}\|, \langle X_0, \hat{\mu}\rangle; \|X_1\|, \|\hat{\mu}\|, \langle X_1, \hat{\mu}\rangle)$$

Idea: for $j \in \{0,1\}$, $(X_j, \hat{\mu})$ has the same distribution as $(RX_j, R\hat{\mu})$ for every rotation matrix R

Observation: The inner product and norm operators mix noisy and unit-variance directions.

Step 2: Zero Auxiliary Sample Case

$$\mathcal{P}_{d-2m,k-m-\text{spiked}} = \left\{\mathcal{N}(0,\Sigma) : \Sigma = I_{d-2m} + \sigma^2 UU^T \text{ where } U \in \mathbb{R}^{(d-2m)\times(k-m)} \text{ with orthonormal columns}\right\}$$

$TV(X_0, \hat{\mu}\,; X_1, \hat{\mu})$ is the distance between two mixtures.

Lemma: Sufficient Statistics for Zero-Side Info. Case

$$TV(X_0, \hat{\mu}\,; X_1, \hat{\mu}) = TV(\|X_0\|, \|\hat{\mu}\|, \langle X_0, \hat{\mu}\rangle; \|X_1\|, \|\hat{\mu}\|, \langle X_1, \hat{\mu}\rangle)$$

Idea: for $j \in \{0,1\}$, $(X_j, \hat{\mu})$ has the same distribution as $(RX_j, R\hat{\mu})$ for every rotation matrix R

Observation: The inner product and norm operators mix noisy and unit-variance directions.

$$TV(\|X_0\|, \|\hat{\mu}\|, \langle X_0, \hat{\mu}\rangle; \|X_1\|, \|\hat{\mu}\|, \langle X_1, \hat{\mu}\rangle) \leq \mathbb{E}[TV((\|X_0\|, \|\hat{\mu}\|, \langle X_0, \hat{\mu}\rangle)|UU^T; (\|X_1\|, \|\hat{\mu}\|, \langle X_1, \hat{\mu}\rangle)|UU^T)]$$

Step 2: Zero Auxiliary Sample Case

$$\mathcal{P}_{d-2m,k-m-\text{spiked}} = \left\{\mathcal{N}(0,\Sigma) : \Sigma = I_{d-2m} + \sigma^2 UU^T \text{ where } U \in \mathbb{R}^{(d-2m)\times(k-m)} \text{ with orthonormal columns}\right\}$$

$TV(X_0, \hat{\mu}\,; X_1, \hat{\mu})$ is the distance between two mixtures.

Lemma: Sufficient Statistics for Zero-Side Info. Case

$$TV(X_0, \hat{\mu}\,; X_1, \hat{\mu}) = TV(\|X_0\|, \|\hat{\mu}\|, \langle X_0, \hat{\mu}\rangle; \|X_1\|, \|\hat{\mu}\|, \langle X_1, \hat{\mu}\rangle)$$

Idea: for $j \in \{0,1\}$, $(X_j, \hat{\mu})$ has the same distribution as $(RX_j, R\hat{\mu})$ for every rotation matrix R

Observation: The inner product and norm operators mix noisy and unit-variance directions.

$$TV(\|X_0\|, \|\hat{\mu}\|, \langle X_0, \hat{\mu}\rangle; \|X_1\|, \|\hat{\mu}\|, \langle X_1, \hat{\mu}\rangle) \leq \mathbb{E}[TV((\|X_0\|, \|\hat{\mu}\|, \langle X_0, \hat{\mu}\rangle)|UU^T; (\|X_1\|, \|\hat{\mu}\|, \langle X_1, \hat{\mu}\rangle)|UU^T)]$$

Conditioning on UU^T makes the distribution simple.

Step 2: Zero Auxiliary Sample Case

$$\mathcal{P}_{d-2m,k-m-\text{spiked}} = \left\{\mathcal{N}(0,\Sigma) : \Sigma = I_{d-2m} + \sigma^2 UU^T \text{ where } U \in \mathbb{R}^{(d-2m)\times(k-m)} \text{ with orthonormal columns}\right\}$$

$TV(X_0, \hat{\mu}\,; X_1, \hat{\mu})$ is the distance between two mixtures.

Lemma: Sufficient Statistics for Zero-Side Info. Case

$$TV(X_0, \hat{\mu}\,; X_1, \hat{\mu}) = TV(\|X_0\|, \|\hat{\mu}\|, \langle X_0, \hat{\mu}\rangle; \|X_1\|, \|\hat{\mu}\|, \langle X_1, \hat{\mu}\rangle)$$

Idea: for $j \in \{0,1\}$, $(X_j, \hat{\mu})$ has the same distribution as $(RX_j, R\hat{\mu})$ for every rotation matrix R

Observation: The inner product and norm operators mix noisy and unit-variance directions.

$$TV(\|X_0\|, \|\hat{\mu}\|, \langle X_0, \hat{\mu}\rangle; \|X_1\|, \|\hat{\mu}\|, \langle X_1, \hat{\mu}\rangle) \leq \mathbb{E}[TV((\|X_0\|, \|\hat{\mu}\|, \langle X_0, \hat{\mu}\rangle)|UU^T; (\|X_1\|, \|\hat{\mu}\|, \langle X_1, \hat{\mu}\rangle)|UU^T)]$$

Conditioning on UU^T makes the distribution simple.

Step 2: Zero Auxiliary Sample Case

$$\mathcal{P}_{d-2m,k-m-\text{spiked}} = \left\{\mathcal{N}(0,\Sigma) : \Sigma = I_{d-2m} + \sigma^2 UU^T \text{ where } U \in \mathbb{R}^{(d-2m)\times(k-m)} \text{ with orthonormal columns}\right\}$$

$TV(X_0, \hat{\mu}\,; X_1, \hat{\mu})$ is the distance between two mixtures.

Lemma: Sufficient Statistics for Zero-Side Info. Case

$$TV(X_0, \hat{\mu}\,; X_1, \hat{\mu}) = TV(\|X_0\|, \|\hat{\mu}\|, \langle X_0, \hat{\mu}\rangle; \|X_1\|, \|\hat{\mu}\|, \langle X_1, \hat{\mu}\rangle)$$

Idea: for $j \in \{0,1\}$, $(X_j, \hat{\mu})$ has the same distribution as $(RX_j, R\hat{\mu})$ for every rotation matrix R

Observation: The inner product and norm operators mix noisy and unit-variance directions.

$$TV(\|X_0\|, \|\hat{\mu}\|, \langle X_0, \hat{\mu}\rangle; \|X_1\|, \|\hat{\mu}\|, \langle X_1, \hat{\mu}\rangle) \leq \mathbb{E}[TV((\|X_0\|, \|\hat{\mu}\|, \langle X_0, \hat{\mu}\rangle)|UU^T; (\|X_1\|, \|\hat{\mu}\|, \langle X_1, \hat{\mu}\rangle)|UU^T)]$$

Conditioning on UU^T makes the distribution simple.

Technical Steps:

1) TV distance between two random variables perturbed by chi-squared RVs.
2) Expected value of ratio of Gaussian quadratic forms.

Outline of the Talk

- Membership Inference Attack and Auxiliary Information ✓

Membership Inference Attacks: identify the training samples.
Auxiliary Information: samples from the same population that the training set comes from.
Current paradigm: distribution-free score function and distribution-dependent threshold

- Membership Inference Attack on Gaussian Mean Estimators ✓

MIA on Gaussian mean estimators requires $\Omega(n^2\rho^2)$ auxillary samples!
This result challenges the current practice of design of MIAs.

- Proof Overview ✓

Idea 1: Reduction from m side information case to zero side information case
Idea 2: Sufficient statistics for zero side information case

Takeaways and Future Directions

Takeaways and Future Directions

Takeaways:

Takeaways and Future Directions

Takeaways:

- We formalize the problem of sample complexity of MIAs.

Takeaways and Future Directions

Takeaways:

- We formalize the problem of sample complexity of MIAs.
- We investigate this question in the setting of Gaussian mean estimation

Takeaways and Future Directions

Takeaways:

- We formalize the problem of sample complexity of MIAs.
- We investigate this question in the setting of Gaussian mean estimation
- When attacker doesn't know the data distribution the number of auxiliary samples can be order-wise larger than the number of training samples!

Takeaways and Future Directions

Takeaways:

- We formalize the problem of sample complexity of MIAs.
- We investigate this question in the setting of Gaussian mean estimation
- When attacker doesn't know the data distribution the number of auxiliary samples can be order-wise larger than the number of training samples!

Takeaways and Future Directions

Takeaways:

- We formalize the problem of sample complexity of MIAs.
- We investigate this question in the setting of Gaussian mean estimation
- When attacker doesn't know the data distribution the number of auxiliary samples can be order-wise larger than the number of training samples!

Future Directions:

Takeaways and Future Directions

Takeaways:

- We formalize the problem of sample complexity of MIAs.
- We investigate this question in the setting of Gaussian mean estimation
- When attacker doesn't know the data distribution the number of auxiliary samples can be order-wise larger than the number of training samples!

Future Directions:

- Attacker with access to samples to a distribution with a small distance to the population.

Takeaways and Future Directions

Takeaways:

- We formalize the problem of sample complexity of MIAs.
- We investigate this question in the setting of Gaussian mean estimation
- When attacker doesn't know the data distribution the number of auxiliary samples can be order-wise larger than the number of training samples!

Future Directions:

- Attacker with access to samples to a distribution with a small distance to the population.
- Extending the results to linear regression, binary classification, etc.

Takeaways and Future Directions

Takeaways:

- We formalize the problem of sample complexity of MIAs.
- We investigate this question in the setting of Gaussian mean estimation
- When attacker doesn't know the data distribution the number of auxiliary samples can be order-wise larger than the number of training samples!

Future Directions:

- Attacker with access to samples to a distribution with a small distance to the population.
- Extending the results to linear regression, binary classification, etc.
- How can we design a better MIAs that can use more auxiliary samples?

Takeaways and Future Directions

Takeaways:

- We formalize the problem of sample complexity of MIAs.
- We investigate this question in the setting of Gaussian mean estimation
- When attacker doesn't know the data distribution the number of auxiliary samples can be order-wise larger than the number of training samples!

Future Directions:

- Attacker with access to samples to a distribution with a small distance to the population.
- Extending the results to linear regression, binary classification, etc.
- How can we design a better MIAs that can use more auxiliary samples?
- Sample complexity of data reconstruction attacks.