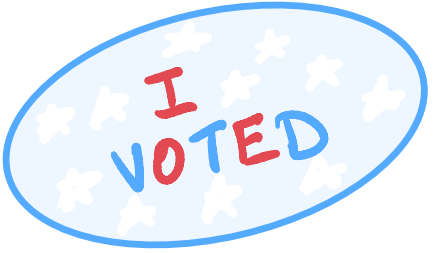
Foundations of Trustworthy AI

Jonathan Ullman

CS7180/7880 · Fall 2026

Class 16: Membership Inference Attacks



Nov 3, 2026

Outline

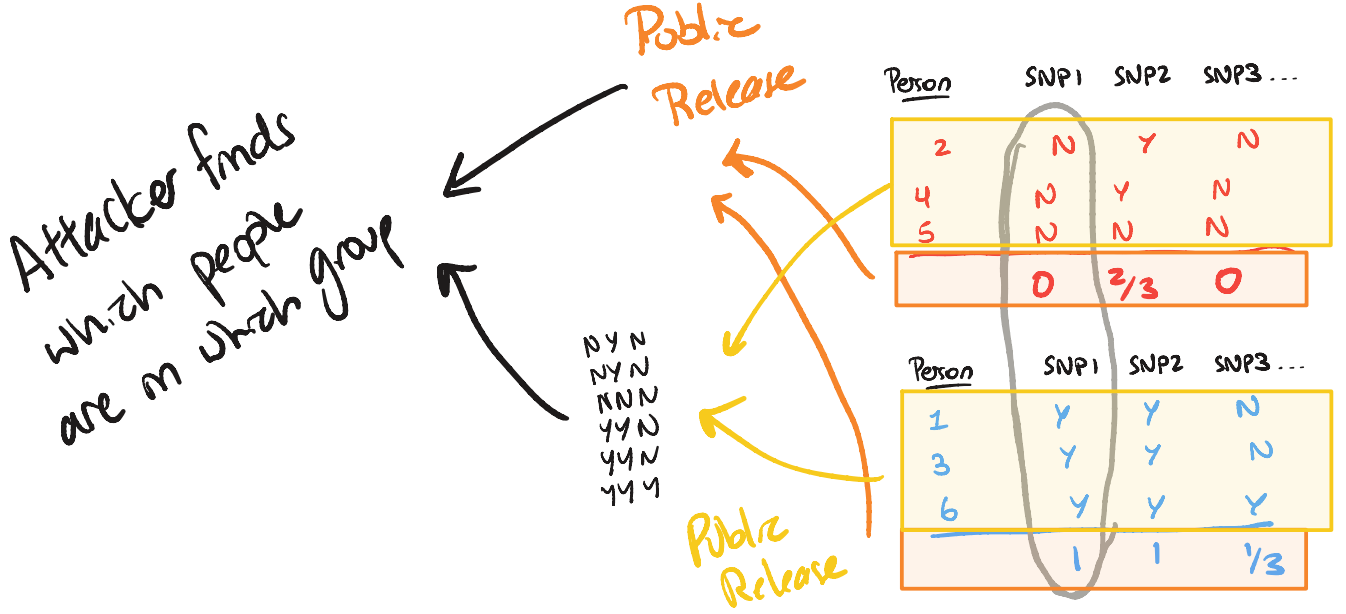
- ~~The OG membership inference~~
- A model of membership inference
- A theoretical example
  - ↳ mechanics
  - ↳ connection to hypothesis testing
  - ↳ connection to membership inference in ML/AI
- "Robust" membership inference

Homer et al. 08 (The OG MIA)

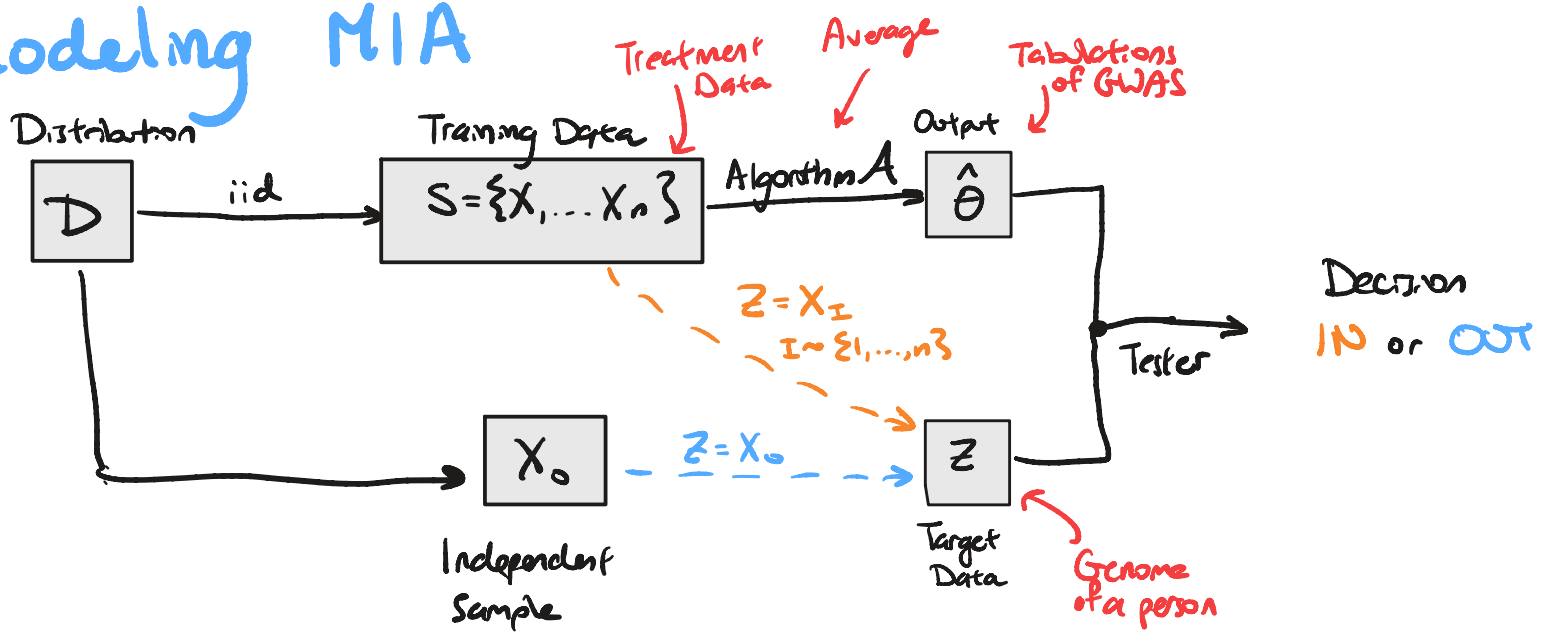
Genome Wide Association Study (GWAS)

(treatment)
People with disease X

(control)
Random people



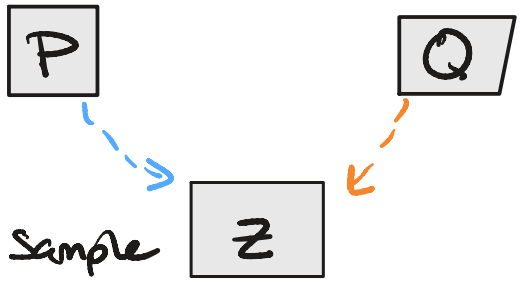
Modeling MIA



Key Observation: Any advantage over random guessing is only due to leakage from the model

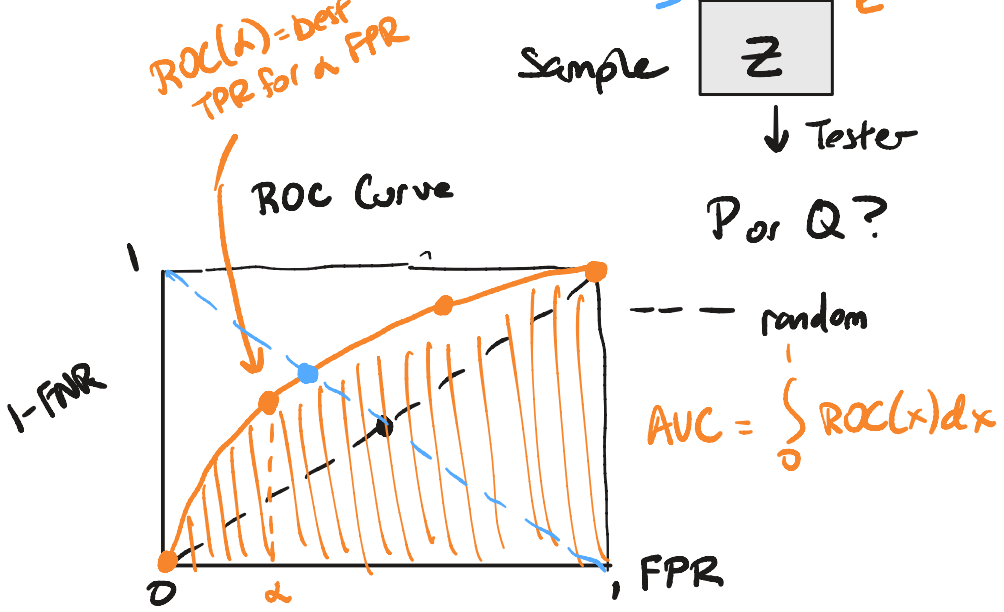
Hypothesis Testing

Two Distributions: P and Q on some domain $\mathcal{X}$



$$FPR = \underset{Z \sim P}{\mathbb{P}}[Test(Z) = Q]$$

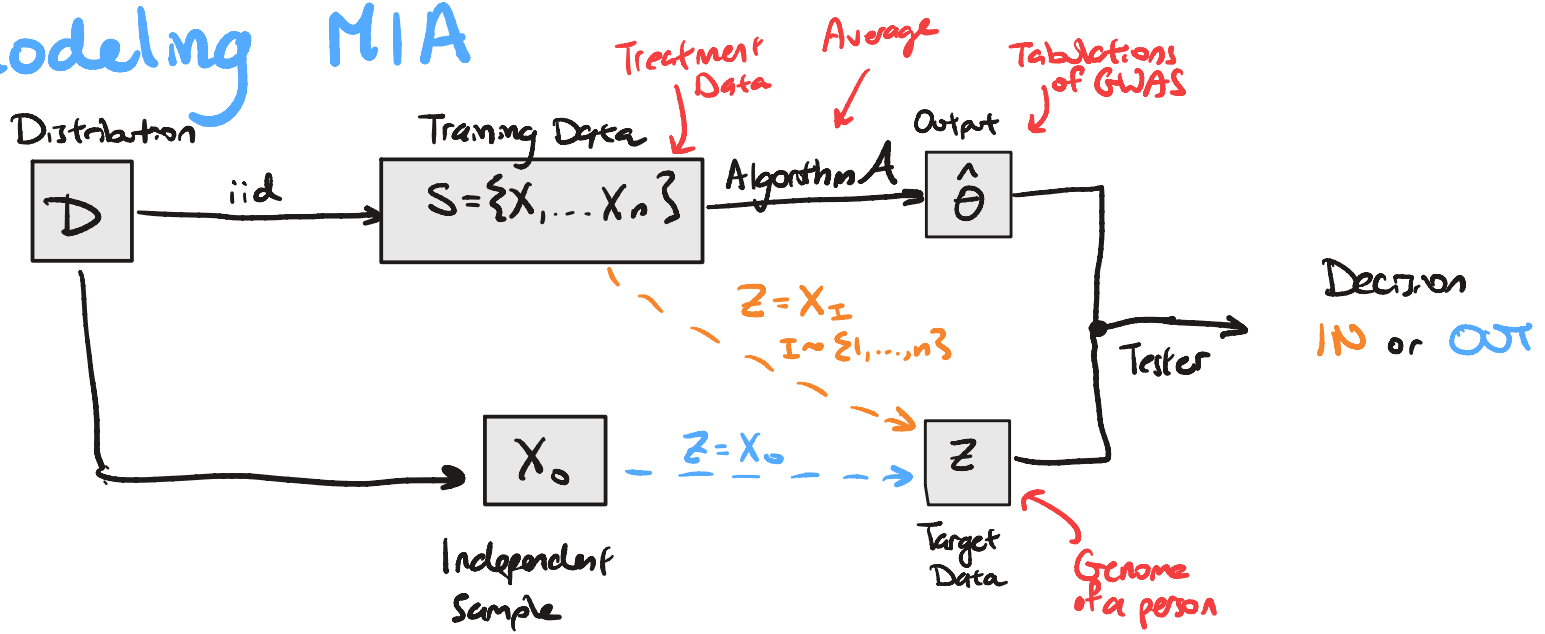
$$FNR = \underset{Z \sim Q}{\mathbb{P}}[Test(Z) = P]$$



Neyman Pearson Lemma

The optimal test has the form $Test(z) = \begin{cases} \frac{\mathbb{P}[Q=z]}{\mathbb{P}[P=z]} \geq t \end{cases}$

Modeling MIA



P_{OUT}: $(X_0, A(X_1,\dots X_n)) = (Z, \hat{\theta})$

P_{IN}: $(X_I, A(X_1,\dots X_n)) = (Z, \hat{\theta})$

If D is known and A is fixed then

MIA is a hypothesis testing problem

The Simplest MIA

$D = \mathcal{N}(\mu, \mathbb{I})$ in $\mathbb{R}^d$

$S = \{X_1, \dots X_n\} \overset{iid}{\sim} D \quad X_0 \overset{iid}{\sim} D$

$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$

Computing the LR:

P_{OUT}: $(X_0, \bar{X})$ independent

$X_0 \sim \mathcal{N}(\mu, \mathbb{I})$

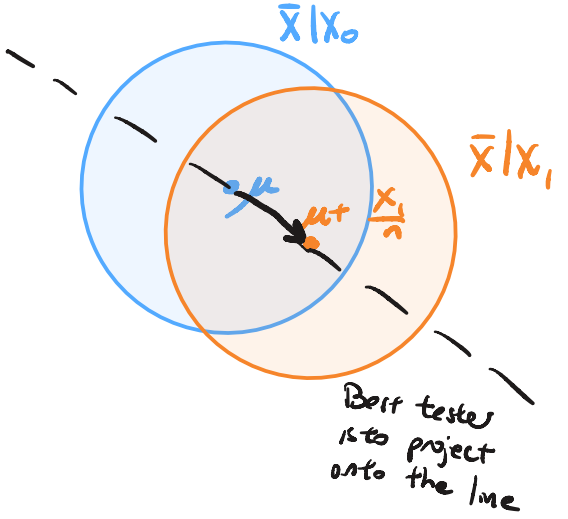
$\bar{X} \sim \mathcal{N}(\mu, \frac{1}{n}\mathbb{I})$

P_{IN}: $(X_1, \bar{X})$ correlated

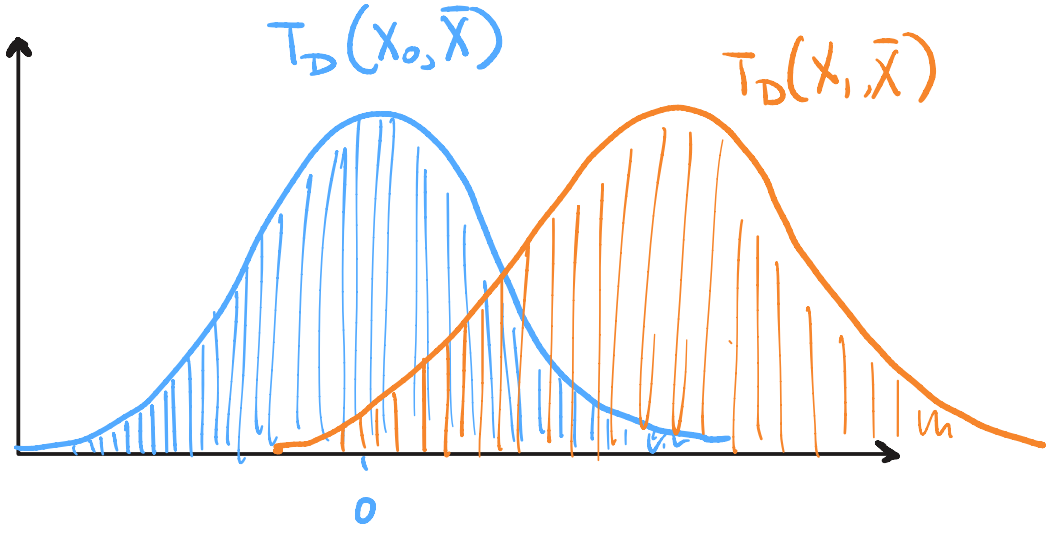
$X_1 \sim \mathcal{N}(\mu, \mathbb{I})$

$\bar{X}|X_1 \sim \mathcal{N}(\frac{X_1}{n} + \mu, \frac{1}{n}\mathbb{I})$

$\bar{X} = \frac{X_1}{n} + \frac{n-1}{n} \cdot \frac{X_2 + \dots + X_n}{n-1}$



$\text{Test}_D(z, \bar{X}) = \langle \bar{X} - \mu, z - \mu \rangle$


$T_D(X_0, \bar{X})$
$T_D(X_1, \bar{X})$
0