CS2080: Applied Privacy for Data Science
Membership Inference Attacks: Theory

James Honaker, Priyanka Nanayakkara, Salil Vadhan

School of Engineering & Applied Sciences
Harvard University

February 5, 2025

Discussion

Consider the simulation experiment performed by Ruggles & van Riper and Hullman's blogpost response. Ruggles & van Riper aimed to cast doubt on the severity of the Census Bureau's findings from their reconstruction attack by comparing to a "null model" (simulating the individual-level 2010 records and finding matches between these and random age-sex draws combined with guesses about race & ethnicity based on previous Census distributions). Hullman argues for a different experiment: compare reconstruction rates on differentially-private data vs. non-differentially-private data.

- Do you agree with Ruggles & van Riper's claims?
- If you were to run your own experiment investigating the need for differential privacy, how would you design it?

Fill in post-discussion Google form!

The Debate Continues…

- Keyes & Flaxman. “How Census Data Put Trans Children at Risk.” Scientific American 2022.

- Hotz et al. “Balancing data privacy and usability in the federal statistical system.” PNAS 2022.

- Jarmin et al. “An in-depth examination of requirements for disclosure risk assessment.” PNAS 2023.
  - Appendix points out severe flaw in Ruggles & van Riper methodology.
  - Several disagreeing response letters.

- Dick et al. “Confidence-Ranked Reconstruction of Census Microdata from Published Statistics.” PNAS 2023.

How to Defend Against Reconstruction

- **Q:** what is a way that we can release many pretty-accurate estimates of proportions (counts divided by n) on a dataset while ensuring that an adversary can only reconstruct a small fraction of our dataset?

- **A:**

The Utility of Subsampling

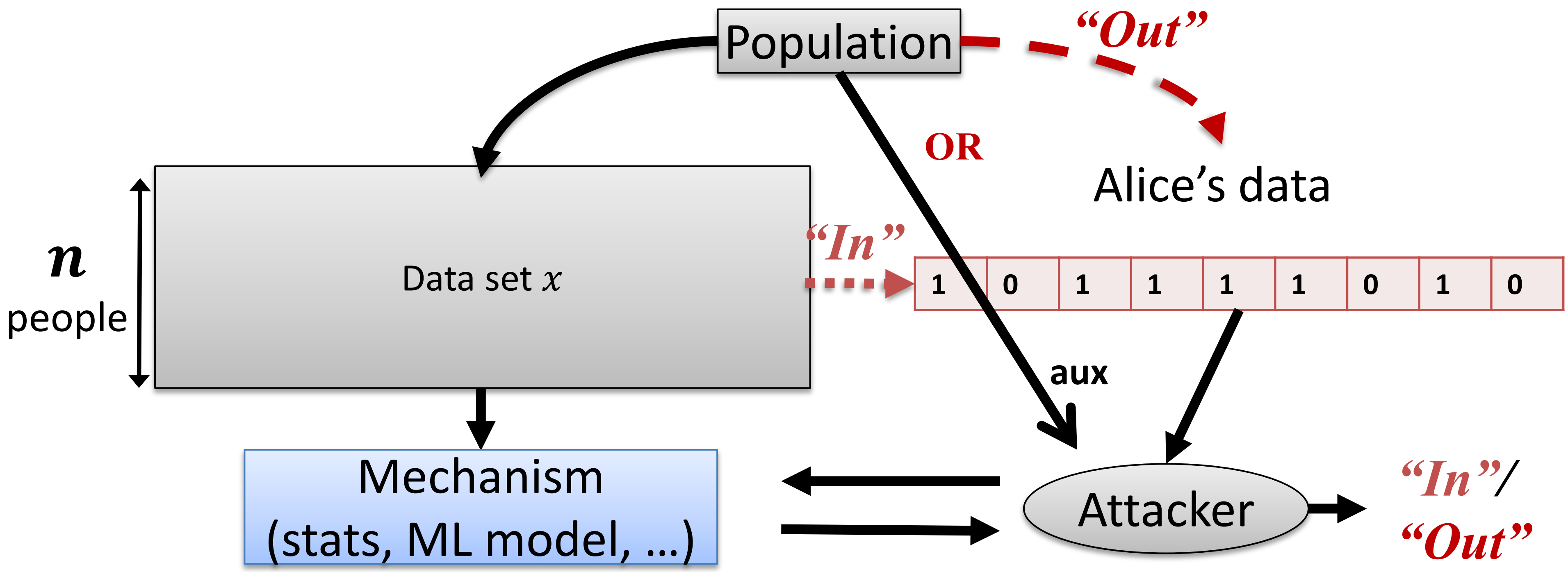
Q: why doesn't the subsampling defense disprove the Dinur-Nissim reconstruction theorem?

A:

Q: are attacks still possible if we allow error larger than $1/\sqrt{n}$?

A:

Membership Inference Attacks: Setup



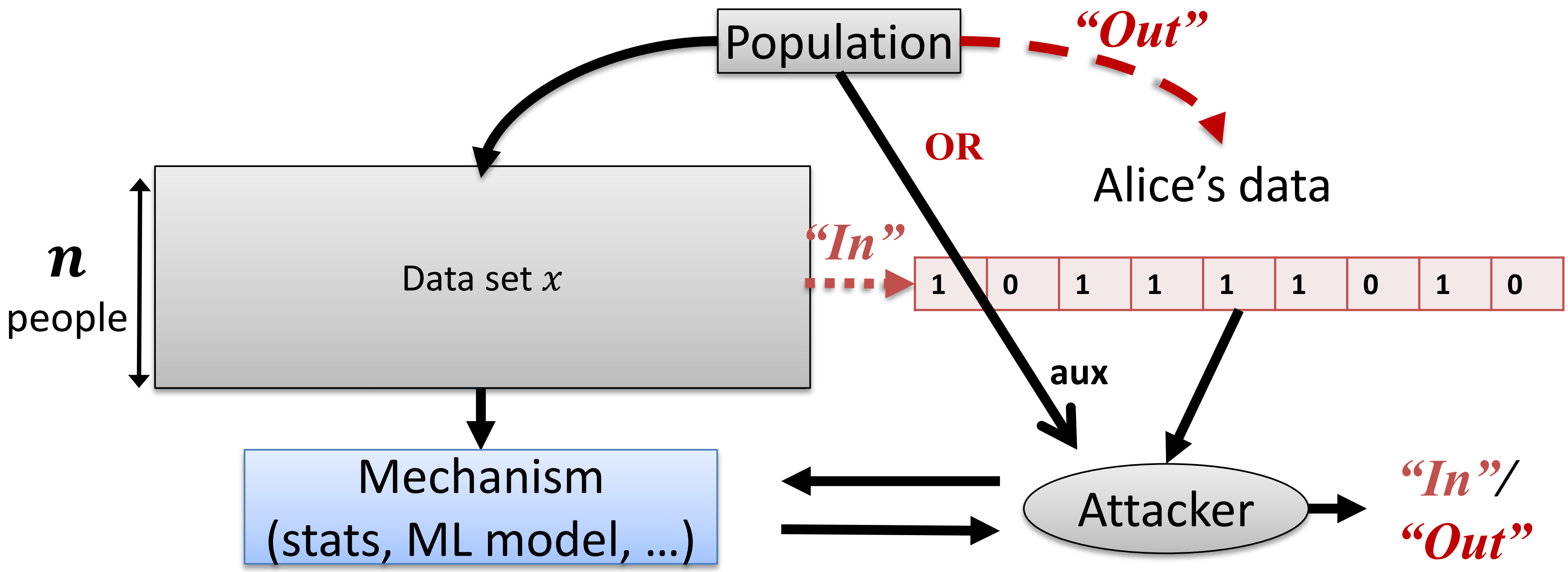
Attacker gets:

- Access to mechanism outputs
- (Some of) Alice's data
- (Possibly) auxiliary info about population
- (Possibly) the code for the mechanism (cf. Kerkhoff's Principle)

Then decides: if Alice is in the dataset x

[slide based on one from Adam Smith]

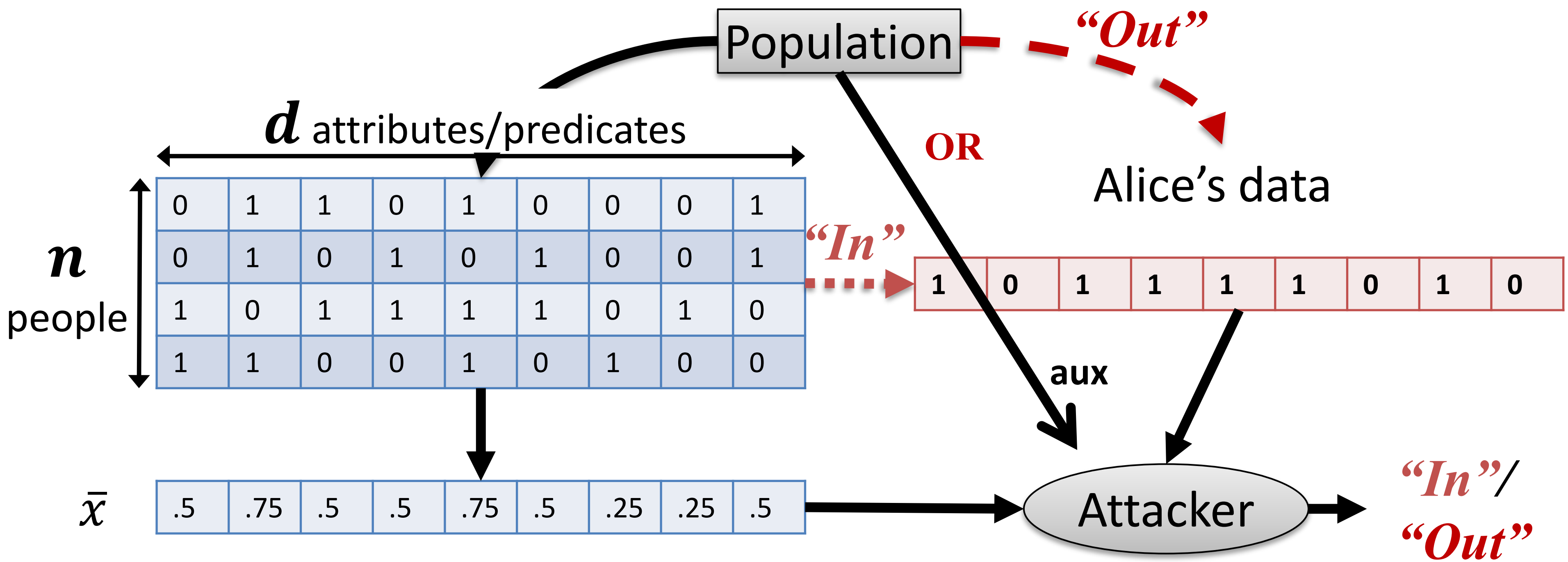
MIAs: Examples



- Genome-wide Association Studies [Homer et al. `08]
  - release frequencies of SNP's (individual positions)
  - determine whether Alice is in "case group" [w/a particular diagnosis]
- ML as a service [Shokri et al. `17]
  - apply models trained on x to Alice's data

[slide based on one from Adam Smith]

MIAs from Means



Some possible aux:

- The vector $p = (p_1, \ldots, p_d)$ of population means
- Or the data of several random individuals from the population

Q: how should the Attacker decide "In" vs. "Out"?

A:

[slide based on one from Adam Smith]

MIAs as Hypothesis Testing

Attacker wants to *reject*

The Null Hypothesis H_0: Alice is not in the dataset, and the dataset is drawn iid from population (given Alice's data and aux)

False Positive Rate (aka Significance Level α, False Alarm, Type I error):

$$\mathrm{FPR} = \Pr[\text{MIA says "In"} \mid H_0]$$

Q: Suppose we have an MIA with a very low FPR (e.g. 10^{-9}) and it outputs "In" on a real-world data release. What do we need to know to be confident that Alice is in the dataset?

A:

Why is a Low FPR Important?

- **A:**

- **Q:** Suppose an attacker goes on a fishing expedition and tries the MIA out on k people, and the MIA says "In" on one of them. Can the attacker be confident that they're in the dataset?

- **A:**

True Positive Rate

Alternative Hypothesis H_1: Alice is a random member of the dataset, which is drawn iid from the population

True Positive Rate (aka "Power", "Sensitivity", "Recall"):

$$\mathrm{TPR} = \Pr[\mathrm{MIA\ says\ "In"} \mid H_1] = 1 - \mathrm{FNR}$$

FNR = "false negative rate", "type II error β", "missed detection"

What FPR & TPR are Meaningful?

- Hypothesis tests only useful if $\mathrm{TPR} > \mathrm{FPR}$.

- MIAs only useful if $\mathrm{TPR} \gtrsim 1/n$, where $n =$ size of dataset

- There are very non-private mechanisms w/best $\mathrm{TPR} = 1/n$.

Salil's Opinion: $\mathrm{TPR} \gtrsim 1/n \gg \mathrm{FPR}$ is most relevant for privacy.

Comparing Attack Frameworks

	Dinur-Nissim Reconstruction	Membership Inference
What is reconstructed?	Explicit attributes	"In" or "Out" attribute
Parameter regime	$\mathrm{FPR} = o(1)$, $\mathrm{TPR} = 1 - o(1)$	$\mathrm{TPR} \gtrsim 1/n \gg \mathrm{FPR}$

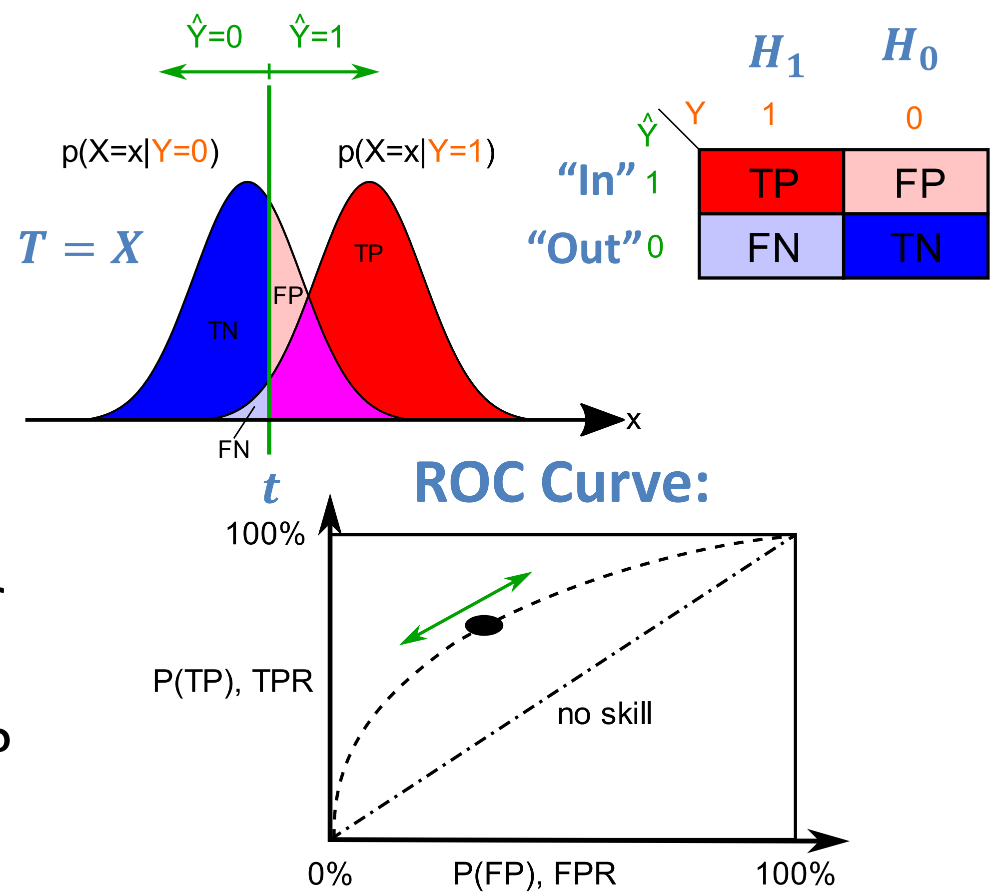
- Reconstruction and Membership Inference Attacks are endpoints on a common spectrum.
  - MIAs ↔ "high-confidence partial reconstruction"

- Important variables for both:
  - Distributional assumptions
  - Quantity & quality of mechanism outputs needed
  - Auxiliary information used by attacker
  - Comparisons to appropriate baselines

How to Design MIAs

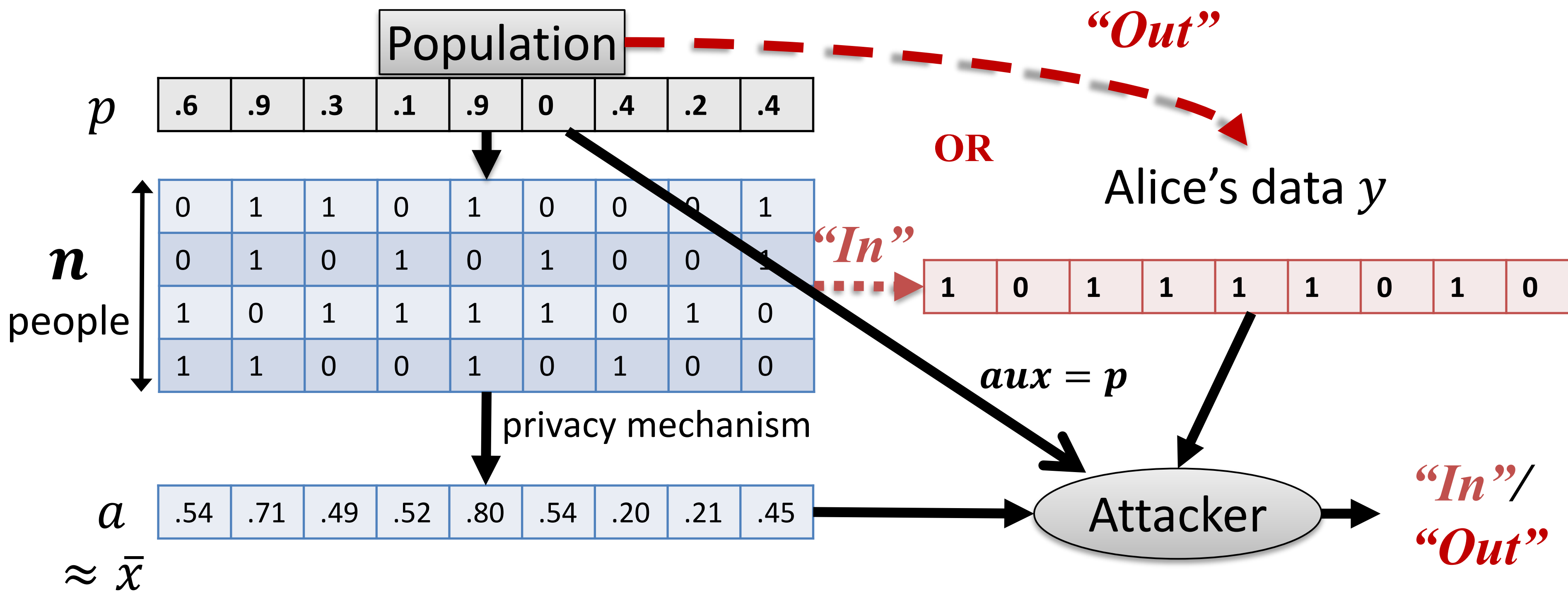
- Design Test Statistic $T = T(\text{everything given to attacker})$ that you expect to be larger under H_1 than H_0.

- Declare "In" if $T \geq t$
  "Out" otherwise
  for a threshold t
  carefully selected
  to tune FPR and TPR.



- **Q:** Why is the "Area Under the ROC Curve" (AUC) not so informative for privacy?

A Test Statistic for Means



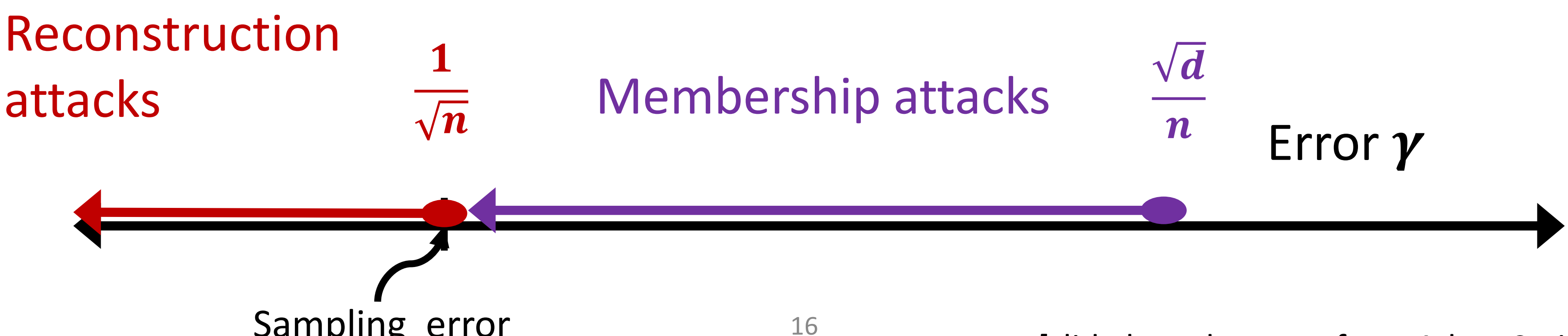
$$A(y, a, p) = \begin{cases} \text{IN} & \text{if } \langle y - p, a - p \rangle \geq t \\ \text{OUT} & \text{if } \langle y - p, a - p \rangle < t \end{cases}$$

Thm [Dwork et al. `15]: under natural distributional assumptions, if mechanism outputs have error smaller than $\gamma < ½$, can achieve

- $\text{FPR} = \exp(-\Omega(d/(\gamma n)^2))$ [very small when $d \gg (\gamma n)^2$]
- $\text{TPR} = \Omega(1/(\gamma^2 n))$ [declare “In” for $k = \Omega(1/\gamma^2)$ members of dataset]

Attacks on Aggregate Stats

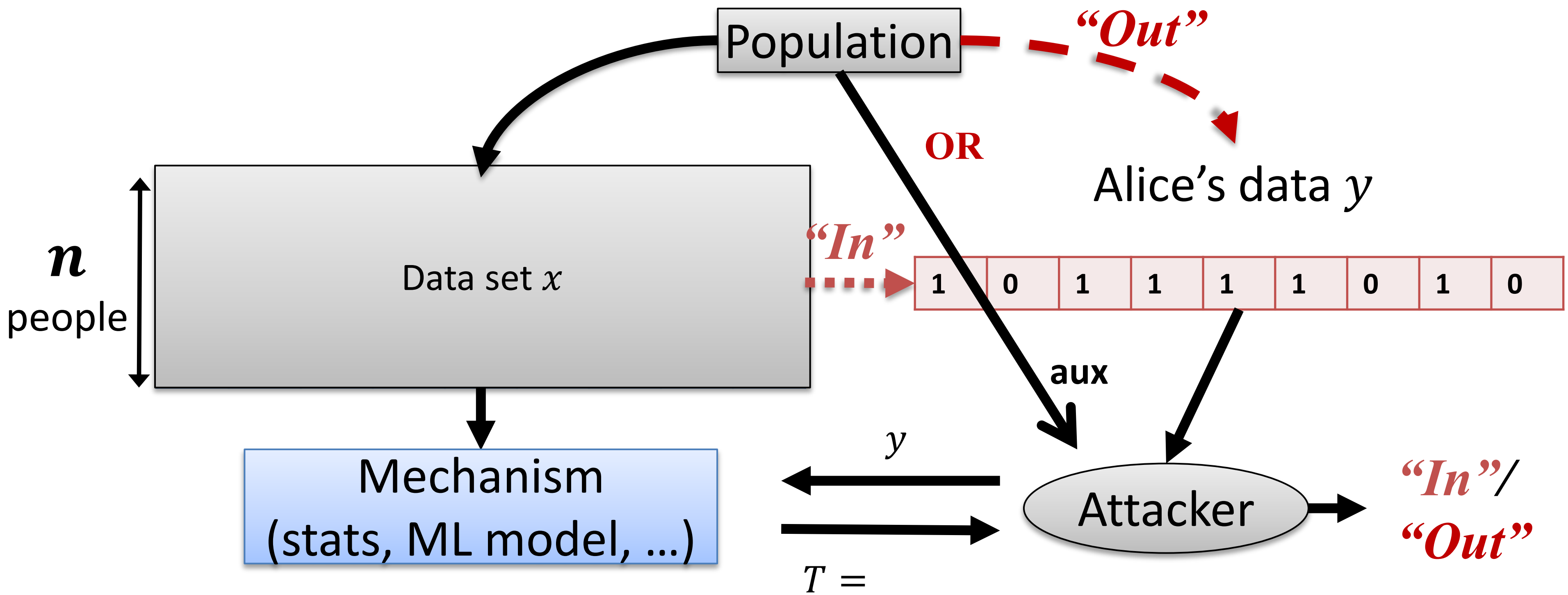
- What error γ makes sense?
  - Estimation error due to sampling $\approx 1/\sqrt{n}$
  - Reconstruction attacks require $\gamma \lesssim 1/\sqrt{n}, d \geq n$
  - Robust membership attacks: $\gamma \lesssim \sqrt{\boldsymbol{d}}/\boldsymbol{n}$
- Lessons
  - "Too many, ~~too accurate~~" statistics reveal individual data
  - "Aggregate" is hard to pin down



[slide based on one from Adam Smith]

A Test Statistic for ML Models

[Shokri et al. `17]



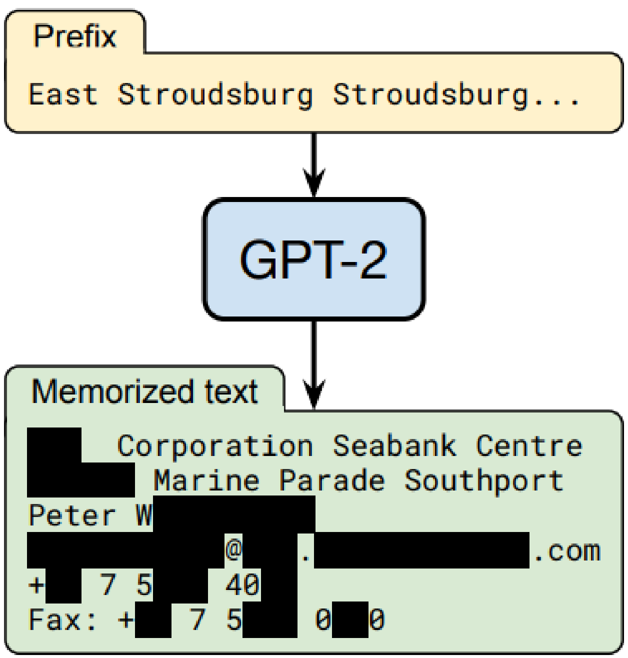
- Exploits "overfitting" of ML models
- Q: how to set threshold t?
- A:

An Optimal Test Statistic

- The Likelihood Ratio: $T_{\mathrm{LR}}(z) = \frac{\Pr[z|H_1]}{\Pr[z|H_0]}$
  - where z=everything the attacker sees
  - Well-defined if H_0, H_1 fully determine probability distribution of z ("simple hypothesis testing")
  - Neyman-Pearson: using $T_{\mathrm{LR}}$with appropriate thresholds t achieves maximum TPR at all FPR, among all hypothesis tests

- T_{LR} be calculated if attacker has full knowledge of mechanism M (e.g. ML training algorithm) and population distribution.
  - Computationally expensive!
  - Much work on efficient approximations to T_{LR} for practical attacks. [Carlini et al. `22, Zarifzadeh et al. `24]

Extracting Training Data from AI Models



[Carlini, Tramèr, Wallace et al. 2021]

[Carlini, Hayes, Nasr et al. 2023]