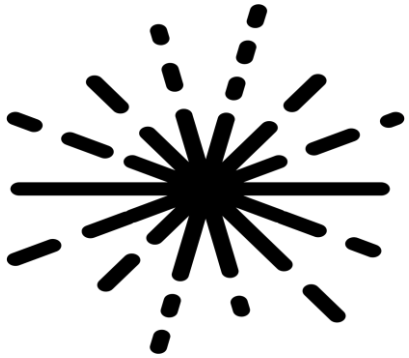


CS2080: Applied Privacy for Data Science
Intro to Membership Inference Attacks

James Honaker, Priyanka Nanayakkara, Salil Vadhan

School of Engineering & Applied Sciences
Harvard University

February 3, 2025

Takeaway Message on Reconstruction Attacks

- Every statistic released yields a (hard or soft) constraint on the dataset.
  - Sometimes have nonlinear or logical constraints ⇒ use fancier solvers (e.g. SAT or SMT solvers)
- Releasing too many statistics with too much accuracy necessarily determines almost the entire dataset.
- This works in theory and in practice (see readings, ps2).

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:**

Subsampling vs. Reconstruction

- **Q:** If the adversary is just trying to reconstruct a single sensitive bit per individual, what fraction of the dataset should we expect the adversary to reconstruct if we subsample k rows and answer arbitrarily many counts?
- **Guess 1:**
- **Guess 2:**
- **A:**

- **Q:** is subsampling a satisfactory privacy defense?

The Utility of Subsampling

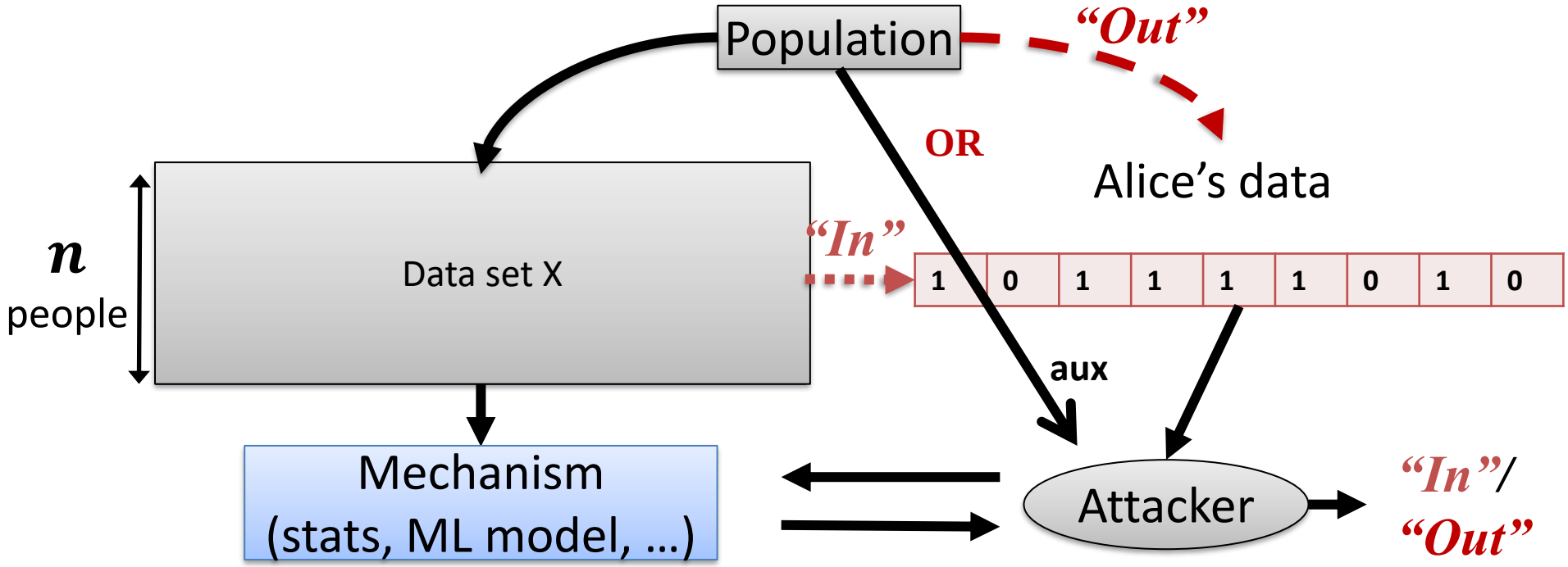
Q: why doesn't the subsampling defense disprove the Dinur-Nissim reconstruction theorem?

A:

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

A:

Membership Inference Attacks: Setup



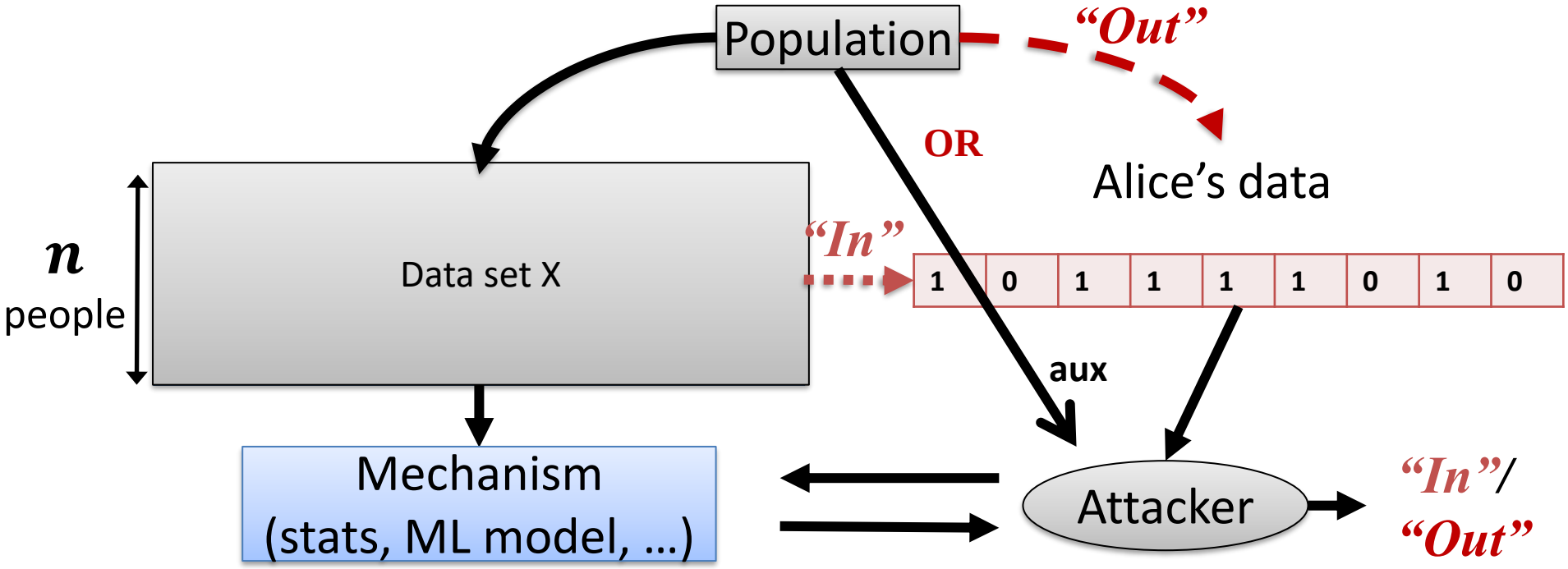
Attacker gets:

- 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]

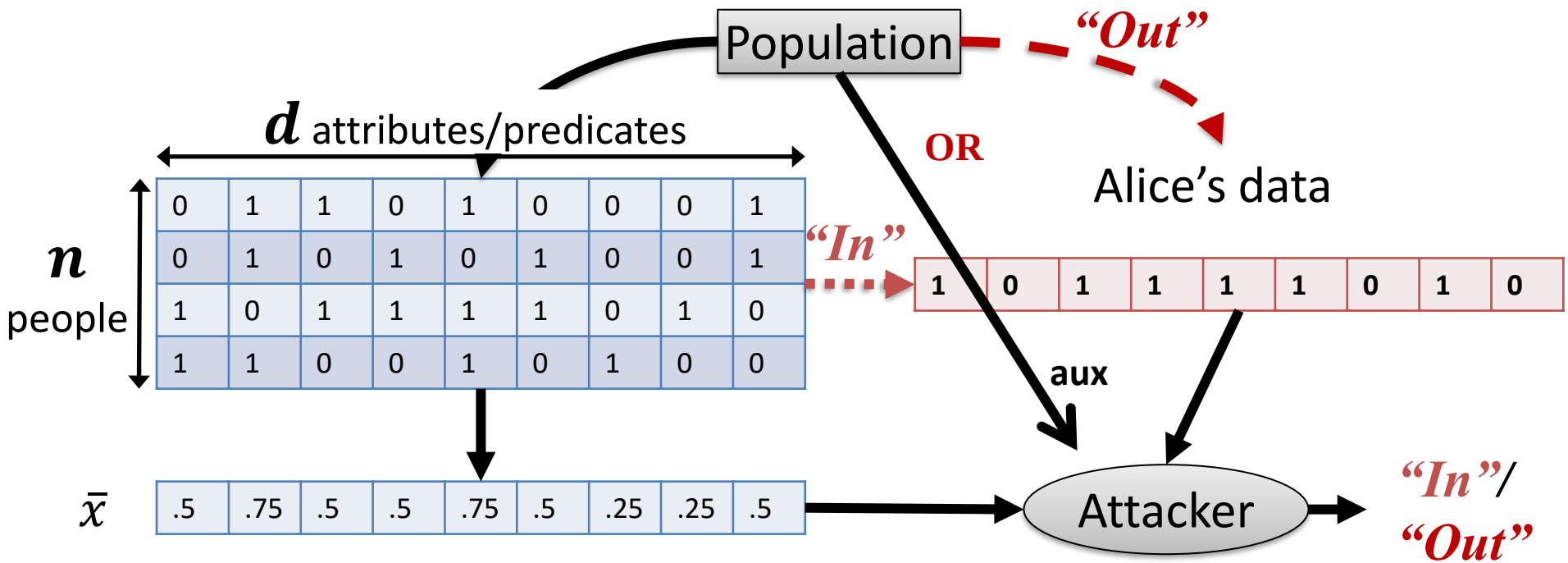
MIAs: Examples



- 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):

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

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 "Sensitivity"):

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

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.

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