CS208: Applied Privacy for Data Science
Membership Attacks

School of Engineering & Applied Sciences
Harvard University

February 1, 2022

Motivation

- Last time: on a dataset with n individuals, releasing $m = n$ counts with error $E = o(\sqrt{n})$ allows for reconstructing $1 - o(1)$ fraction of sensitive attributes. [Dinur-Nissim `03]

What is this $\sqrt{\boldsymbol{n}}$ threshold?

- if $X = X_1 + \cdots + X_n$ for independent random variables X_i each with standard deviation σ, then the standard deviation of X is $\Theta(\sqrt{n})$.

- If the X_i's are bounded (or "subgaussian"), then X will have Gaussian-like concentration around its expectation $\mathrm{n}\mu$:
  $\Pr[|X - n\mu| > t \cdot \sqrt{n}] \leq e^{-\Omega(t^2)}$ [Chernoff-Hoeffding Bound]

This is why subsampling k out of n rows allows us to approximate m counts each to within $\pm O\left(\sqrt{\underbrace{k}_{\text{std dev}}\ \underbrace{\log m}_{\text{concentration}}}\right)$

Normalized Counts (i.e. Averages)

- if $X = (X_1 + \cdots + X_n)/n$ for independent random variables X_i each with standard deviation σ, then the standard deviation of X is $\Theta(1/\sqrt{n})$
- If the X_i's are bounded (or "subgaussian"), then X will have Gaussian-like concentration around its mean μ:
  $\Pr[|X-\mu| > t/\sqrt{n}] \leq e^{-\Omega(t^2)}$ [Chernoff-Hoeffding Bound]

This is why subsampling k out of n rows allows us to approximate m averages each to within $\pm O\left(\underbrace{\left(\frac{1}{\sqrt{k}}\right)}_{\text{std dev}} \cdot \underbrace{\sqrt{\log m}}_{\text{concentration}}\right)$

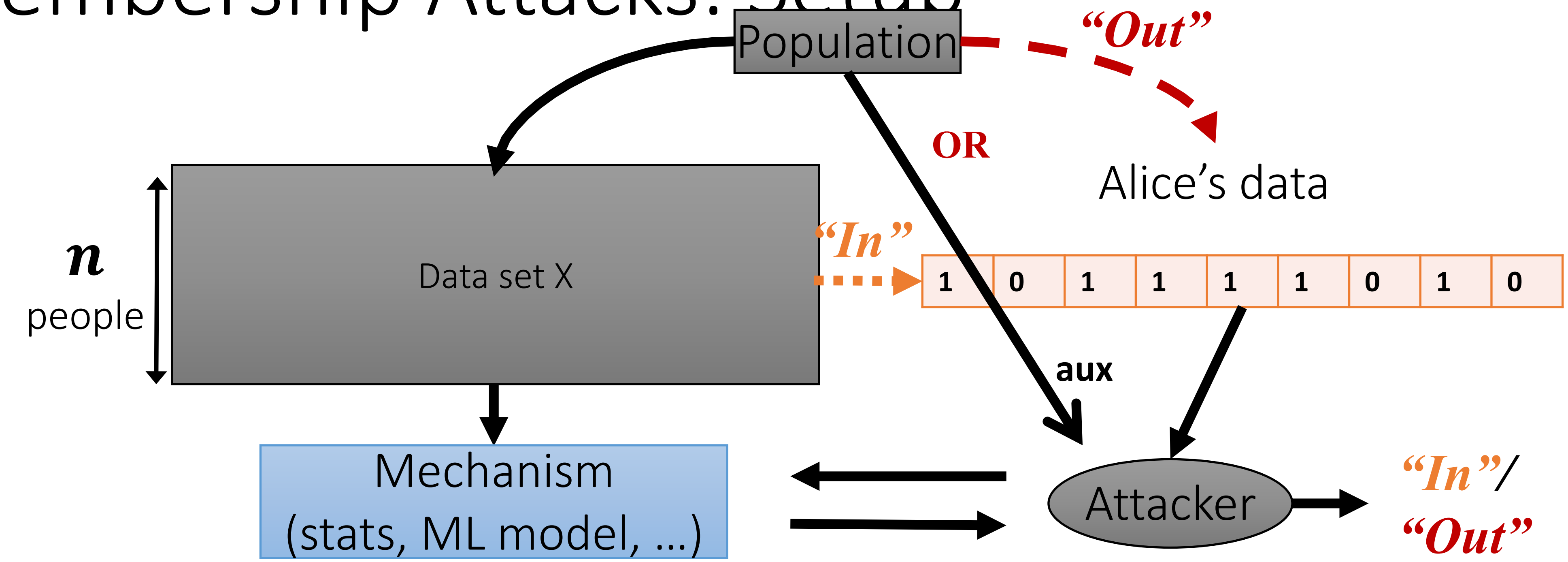
Motivation

- Last time: on a dataset with n individuals, releasing $m = n$ averages with error $E = o(1/\sqrt{n})$ allows for reconstructing $1 - o(1)$ fraction of sensitive attributes.

- Q: what happens if we allow error $\Omega(1/\sqrt{n}) \leq E \leq o(1)$?

- A (today): if we release $m = n^2$ counts, can be vulnerable to "membership attacks".

Membership Attacks: Setup



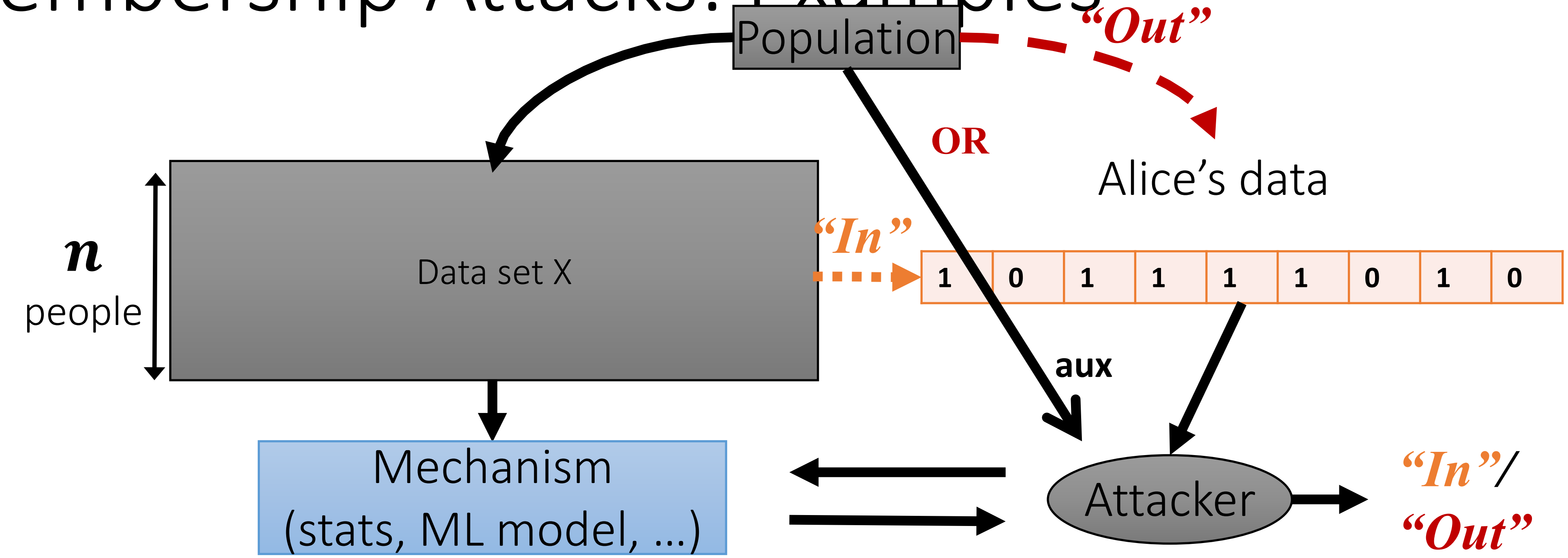
Attacker gets:

- Access to mechanism outputs
- Alice's data
- (Possibly) auxiliary info about population

Then decides: if Alice is in the dataset X

[slide based on one from Adam Smith]

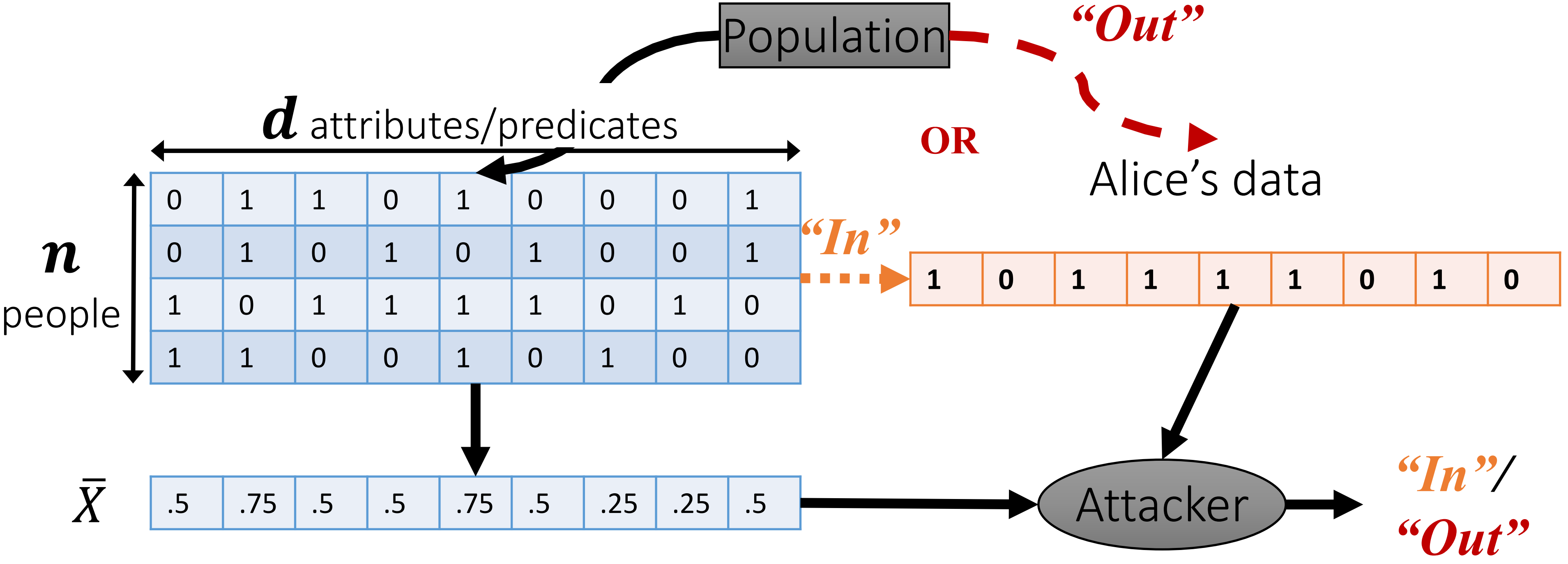
Membership Attacks: Examples



- 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

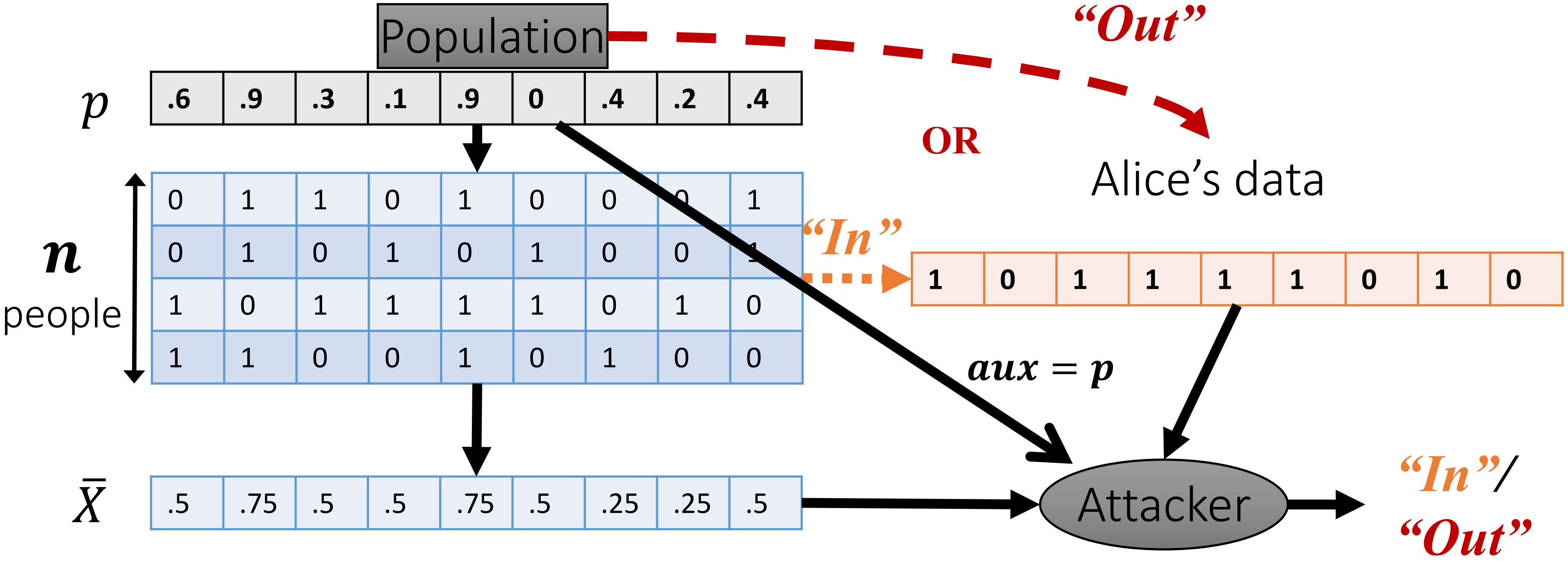
[slide based on one from Adam Smith]

Membership Attacks from Means



[slide based on one from Adam Smith]

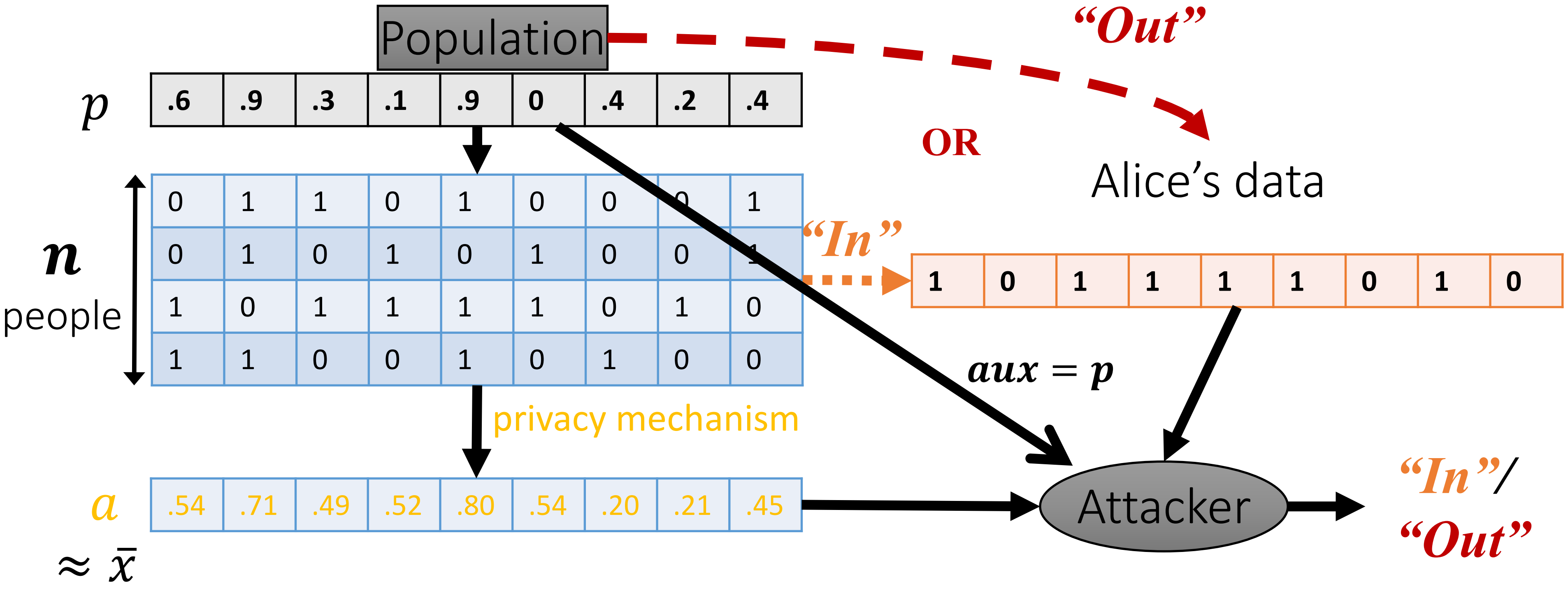
Membership Attacks from Means



- Population = [vector $p = (p_1, \ldots, p_d)$ of probabilities]
  - j'th attribute = iid Bernoulli(p_j), independent across j
  - Attacker gets p (or a few random draws)

[slide based on one from Adam Smith]

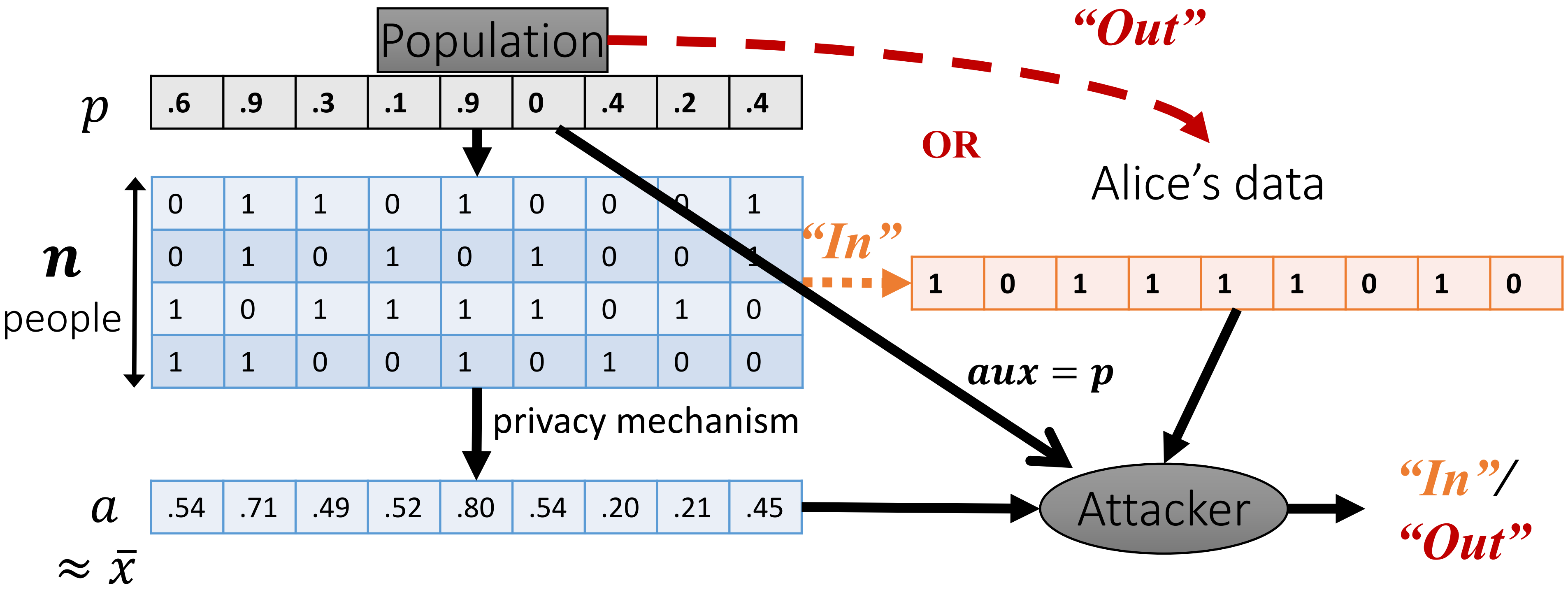
Membership Attacks from Noisy Means



- Population = vector $p = (p_1, \dots, p_d)$ of probabilities
  - j'th attribute = iid Bernoulli(p_j), independent across j
  - Adversary gets $a \approx \bar{x}$ and p (or a few random draws)
  - Only assume that $a = M(x)$ has $|a_j - \bar{x}_j| \leq \alpha$ whp.
    ("Noise" need not be independent or unbiased.)

[slide based on one from Adam Smith]

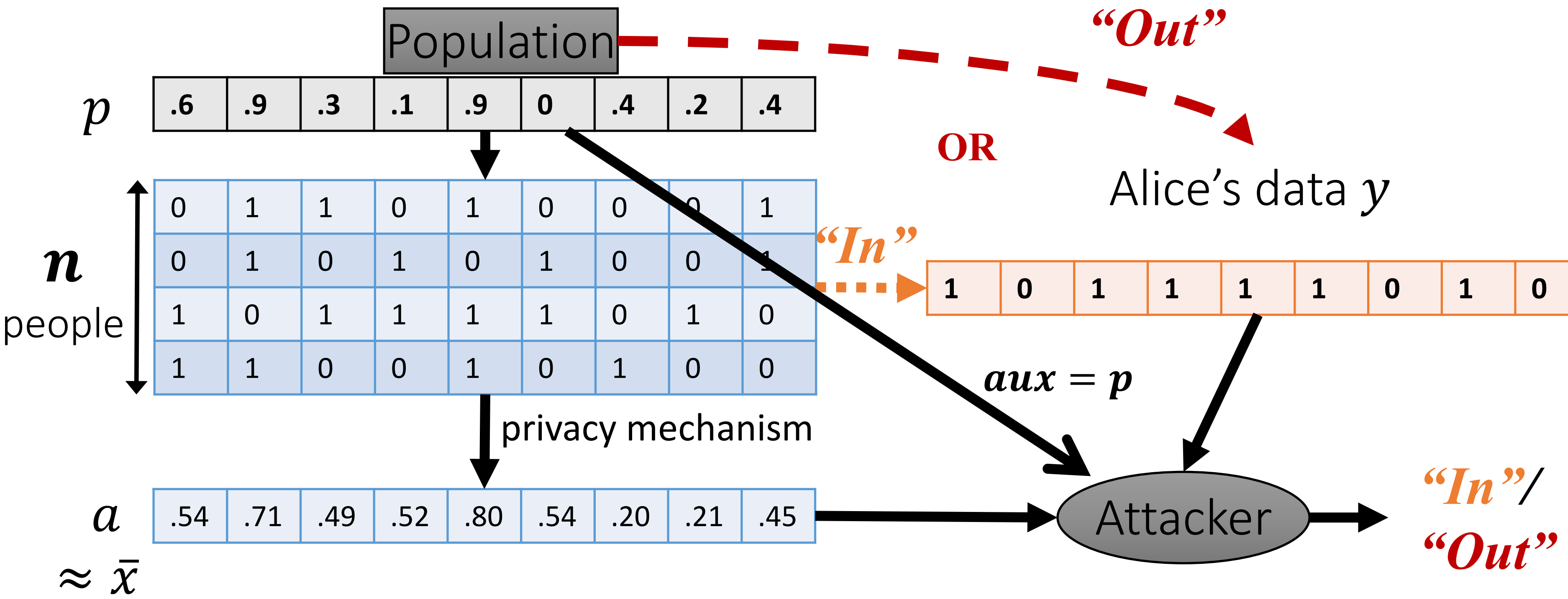
Membership Attacks from Noisy Means



- We are interested in $\alpha > 1/\sqrt{n}$.
- In this regime, if p known to mechanism, can prevent attack. (Q: Why?)
- So we will assume random p_j's (e.g. iid uniform in $[0,1]$).

[slide based on one from Adam Smith]

Membership Attacks from Noisy Means



Theorem [Dwork et al. `15]: There is a constant c and an attacker A such that when $d \geq cn$ and $\alpha = |a - \bar{x}| < \min\left\{\sqrt{d/O(n^2 \log(1/\delta))}, 1/2\right\}$:

- If Alice is IN, then $\Pr[A(y, a, p) = \text{IN}] \geq \Omega\left(\frac{1}{\alpha^2 n}\right)$.
- If Alice is OUT, then $\Pr[A(y, a, p) = \text{IN}] \leq \delta$.

[slide based on one from Adam Smith]

Membership Attacks from Noisy Means

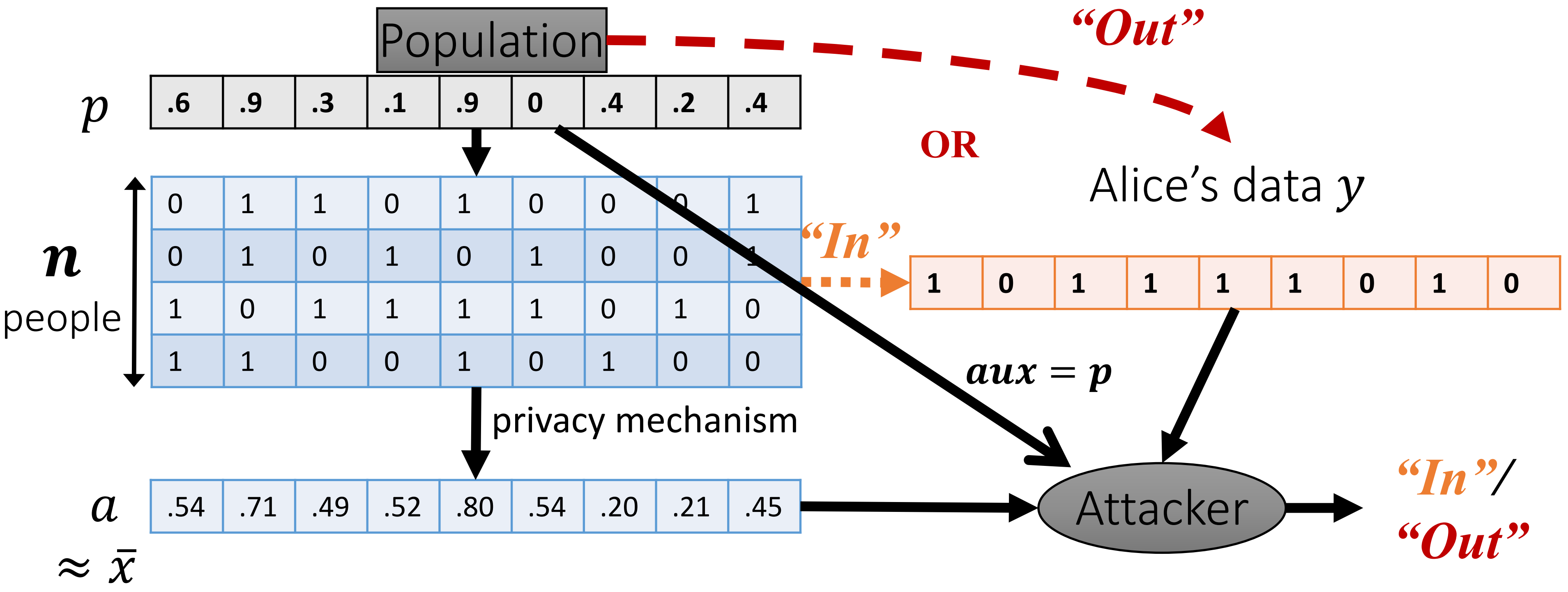
Theorem [Dwork et al. `15]: There is an attacker A such that when $d \geq O(n)$ and $\alpha < \min\left\{\sqrt{d/O(n^2 \log(1/\delta))}, 1/2\right\}$:

- If Alice is IN, then $\Pr[A(y, a, p) = \mathrm{IN}] \geq \Omega\left(\frac{1}{\alpha^2 n}\right)$. (true positive)
- If Alice is OUT, then $\Pr[A(y, a, p) = \mathrm{IN}] \leq \delta$. (false positive)

Remarks:

- Only interesting when $\delta < \Omega\left(\frac{1}{\alpha^2 n}\right)$.
- On average, successfully trace $\Omega\left(\frac{1}{\alpha^2}\right)$ members of dataset. This is the best possible. (Why?)
- Gives hope of safely release at most $\tilde{O}(n^2)$ means!

The Attacker



Q: How would you do the attack?

$$A(y, a, p) = \begin{cases} \text{IN} & \text{if } \langle y, a\rangle - \langle p, a\rangle > T \\ \text{OUT} & \text{if } \langle y, a\rangle - \langle p, a\rangle \leq T \end{cases}$$

Note: given p, a, can choose $T = T_{p,a} = O\left(\sqrt{d \log(1/\delta)}\right)$ to make false positive probability exactly δ.

[slide based on one from Adam Smith]

Attacks on Aggregate Stats (mean)

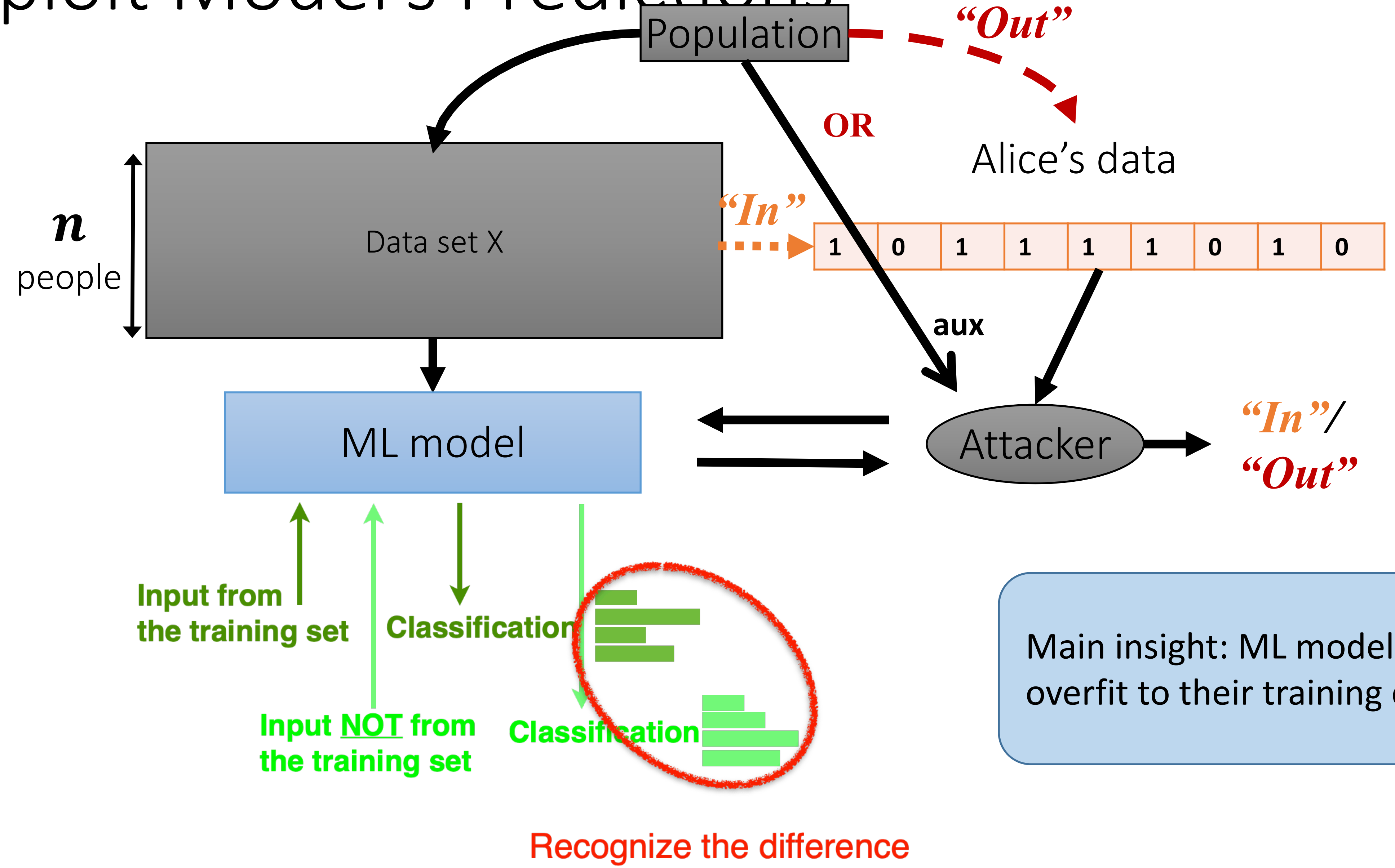
- What error α makes sense?
  - Estimation error due to sampling $\approx 1/\sqrt{n}$
  - Reconstruction attacks require $\alpha \lesssim 1/\sqrt{n}, d \geq n$
  - Robust membership attacks: $\alpha \lesssim \sqrt{\boldsymbol{d}}/\boldsymbol{n}$
- Lessons
  - "Too many, ~~too accurate~~" statistics reveal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

[slide based on one from Adam Smith]

Membership Attacks on ML

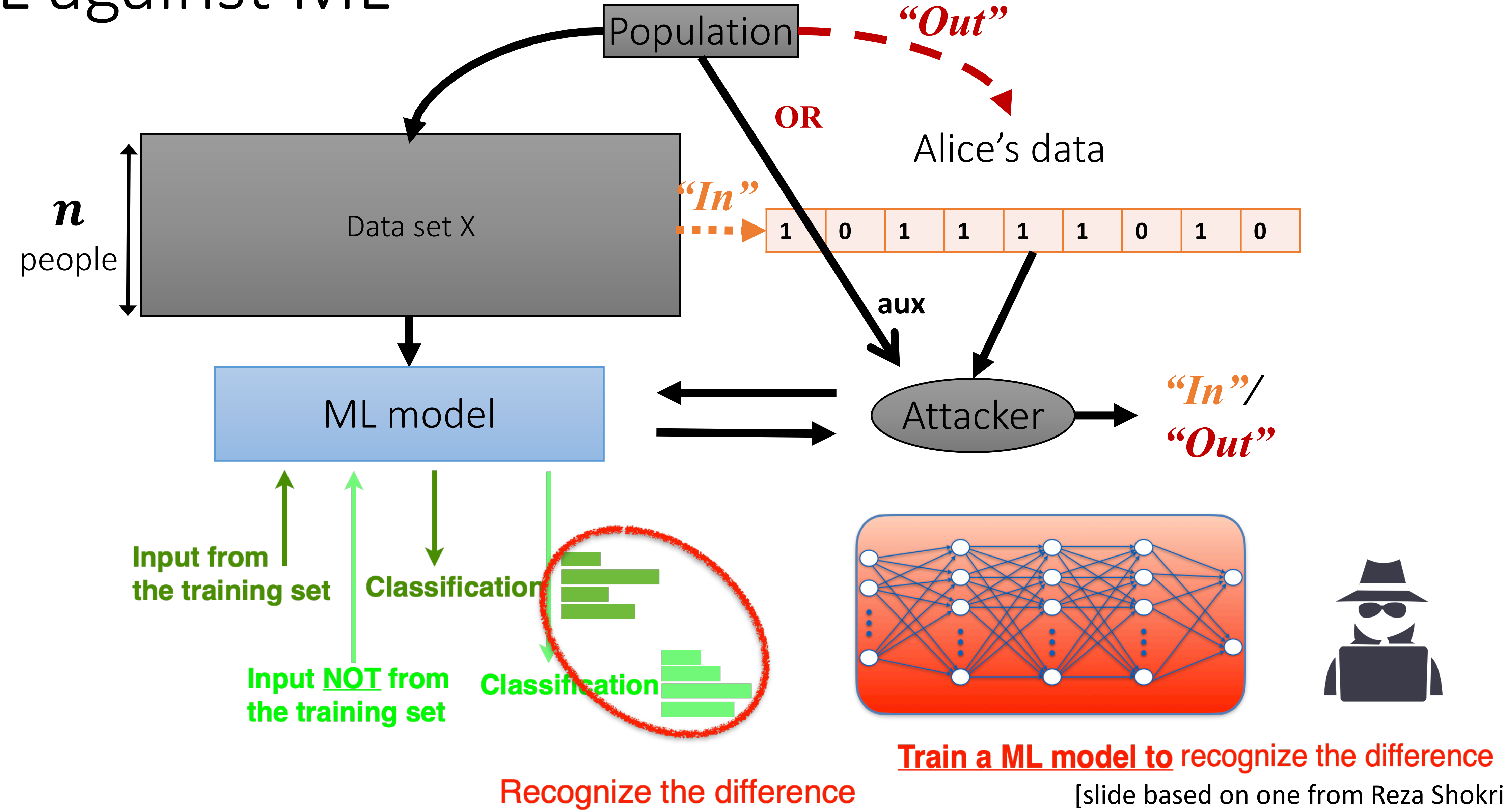
[Shokri et al. 2017]

Exploit Model's Predictions



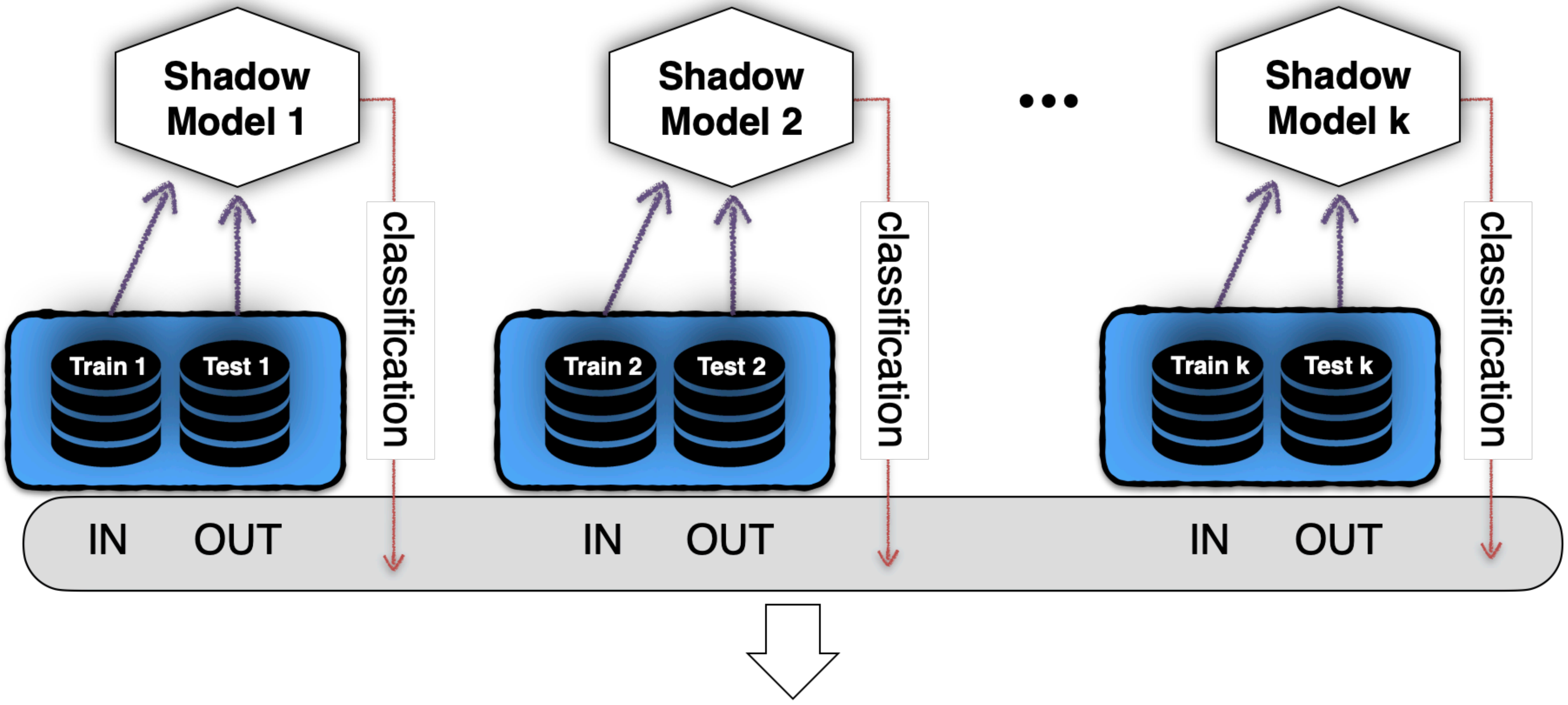
Main insight: ML models overfit to their training data

ML against ML



[slide based on one from Reza Shokri]

Attack Technique – Shadow Models



Train the attack model

to predict if an input was a member of the training set (in) or a non-member (out)

[slide based on one from Reza Shokri]

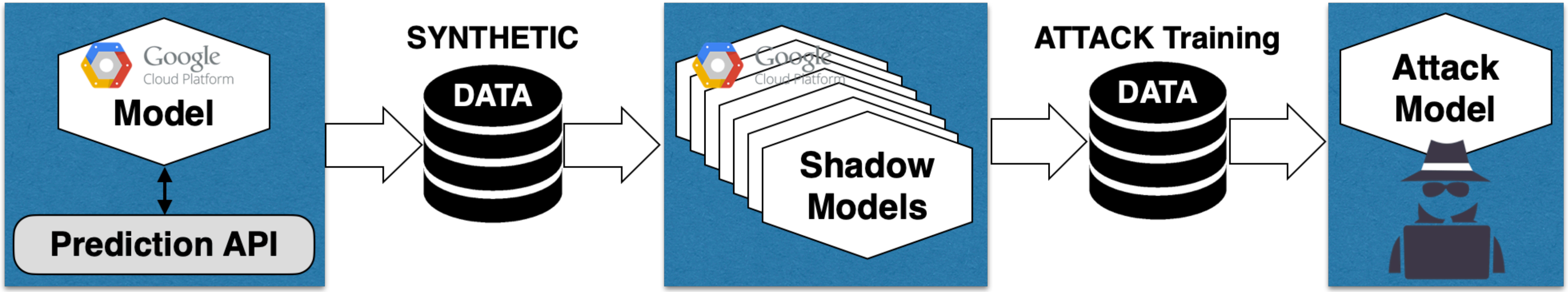
Obtaining Data for Training Shadow Models

- **Real**: similar to training data of the target model (i.e., drawn from same distribution)

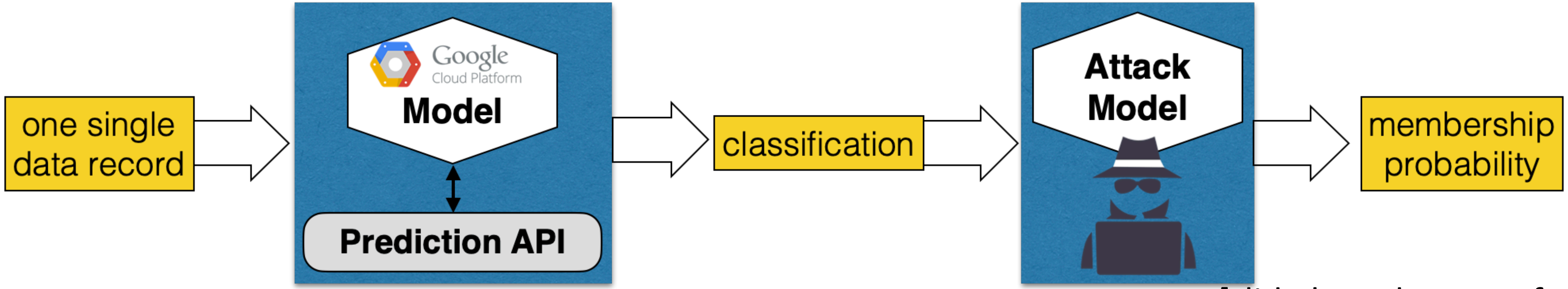
- **Synthetic**: use a sampling algorithm to obtain data classified with high confidence by the target model

[slide based on one from Reza Shokri]

Attack Pipeline



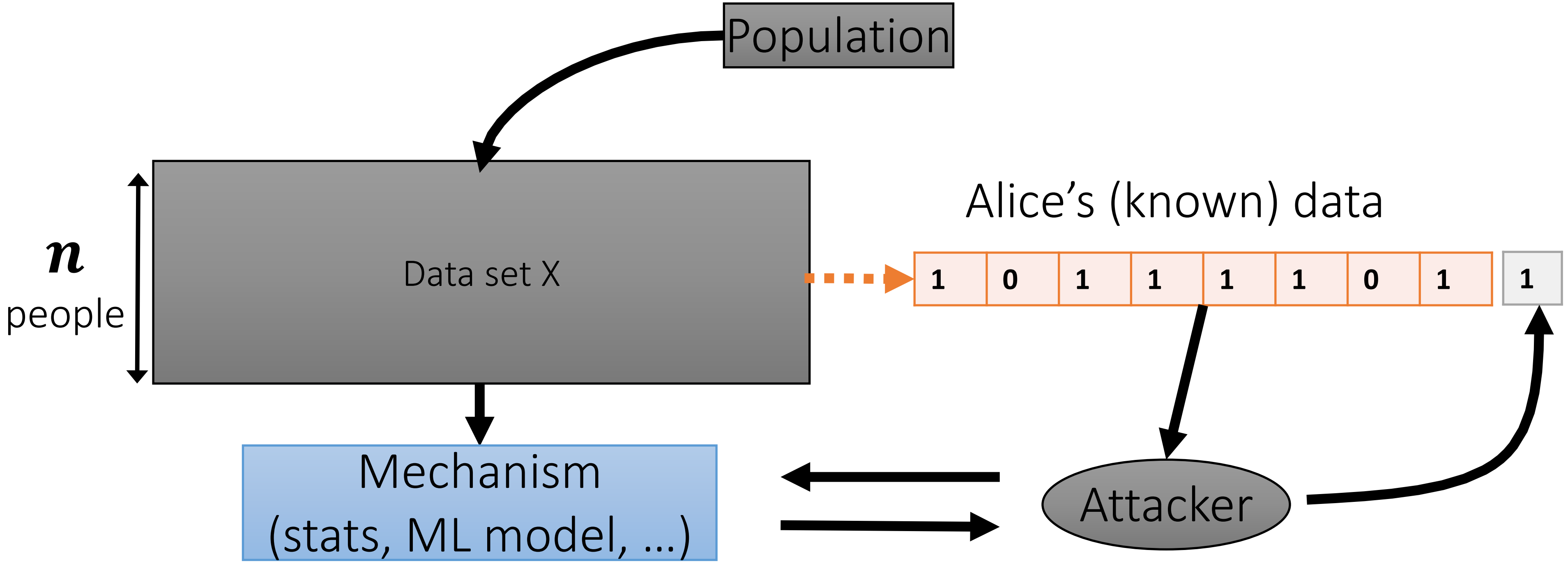
Using the Attack Model



[slide based on one from Reza Shokri]

Another Attack on ML?

[Frederickson et al. `14, cf. McSherry `16]



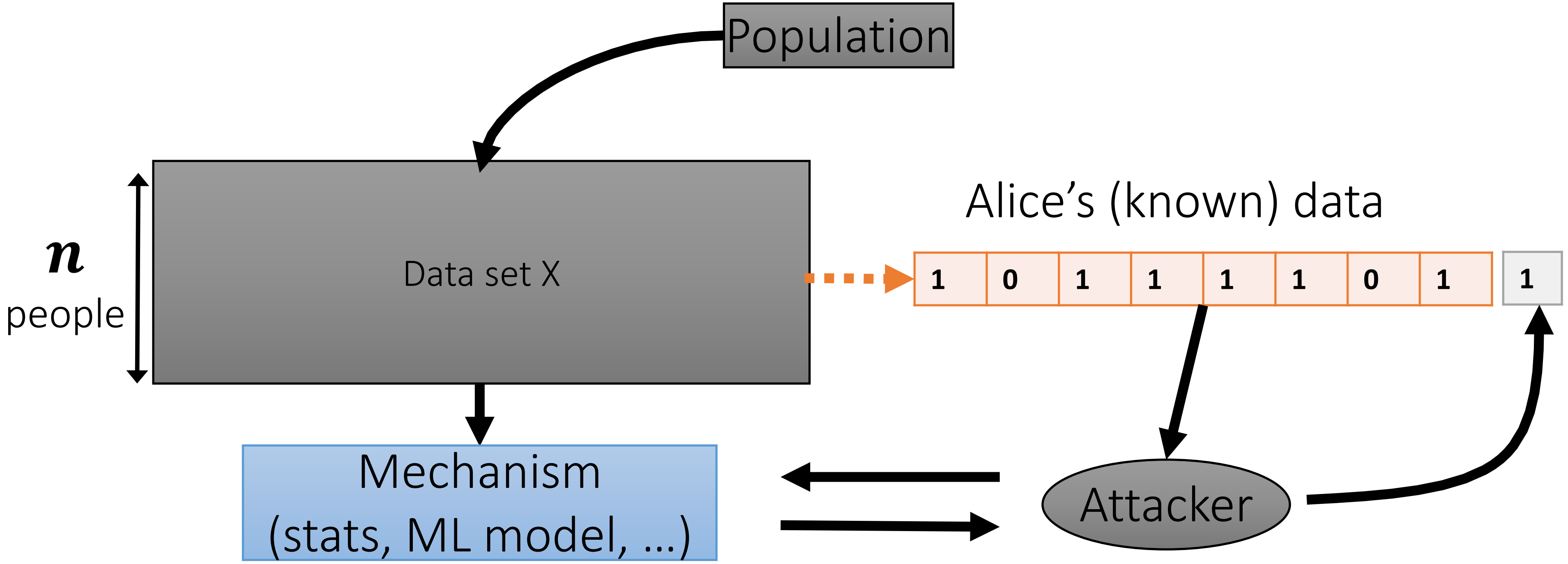
Attacker gets:

- Access to mechanism outputs
- Some of Alice's data
- (Possibly) auxiliary info about population

Then computes: a sensitive attribute of Alice

Another Attack on ML?

[Frederickson et al. `14, cf. McSherry `16]



Difference from reconstruction attacks:

- Above attack works even if Alice not in dataset. Based on correlation between known & sensitive attributes.
- Reconstruction attacks work even when sensitive bit uncorrelated.

Goals of Differential Privacy

- Utility: enable “statistical analysis” of datasets
  - e.g. inference about population, ML training, useful descriptive statistics
- Privacy: protect individual-level data
  - against “all” attack strategies, auxiliary info.