

CS208: Applied Privacy for Data Science Membership & Other Attacks

James Honaker & Salil Vadhan
School of Engineering & Applied Sciences
Harvard University

February 8, 2019



CRCS Center for Research on
Computation and Society

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.
- **Q:** what happens if we allow error $\Omega(\sqrt{n}) \leq E \leq o(n)$?
- **A (today):** if we release $m = n^2$ counts, can be vulnerable to “membership attacks”.

What is this \sqrt{n} threshold?

- if $X = X_1 + \dots + X_n$ for independent random variables X_i each with standard deviation σ , then the standard deviation of X is $\sigma \cdot \sqrt{n}$.
- So the “sampling error” for a sum is typically $\Theta(\sqrt{n})$.
- If the X_i ’s are bounded (or “subgaussian”), then X will have Gaussian-like concentration around its mean μ :
 $\Pr[|X - \mu| > t \cdot \sqrt{n}] \leq e^{-\Omega(t^2)}$ [Chernoff-Hoeffding Bound]

Normalized Counts (i.e. Averages)

- if $X = (X_1 + \dots + X_n)/n$ for independent random variables X_i each with standard deviation σ , then the standard deviation of X is σ/\sqrt{n} .
- So the “sampling error” for a sum is typically $\Theta(1/\sqrt{n})$.
- 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

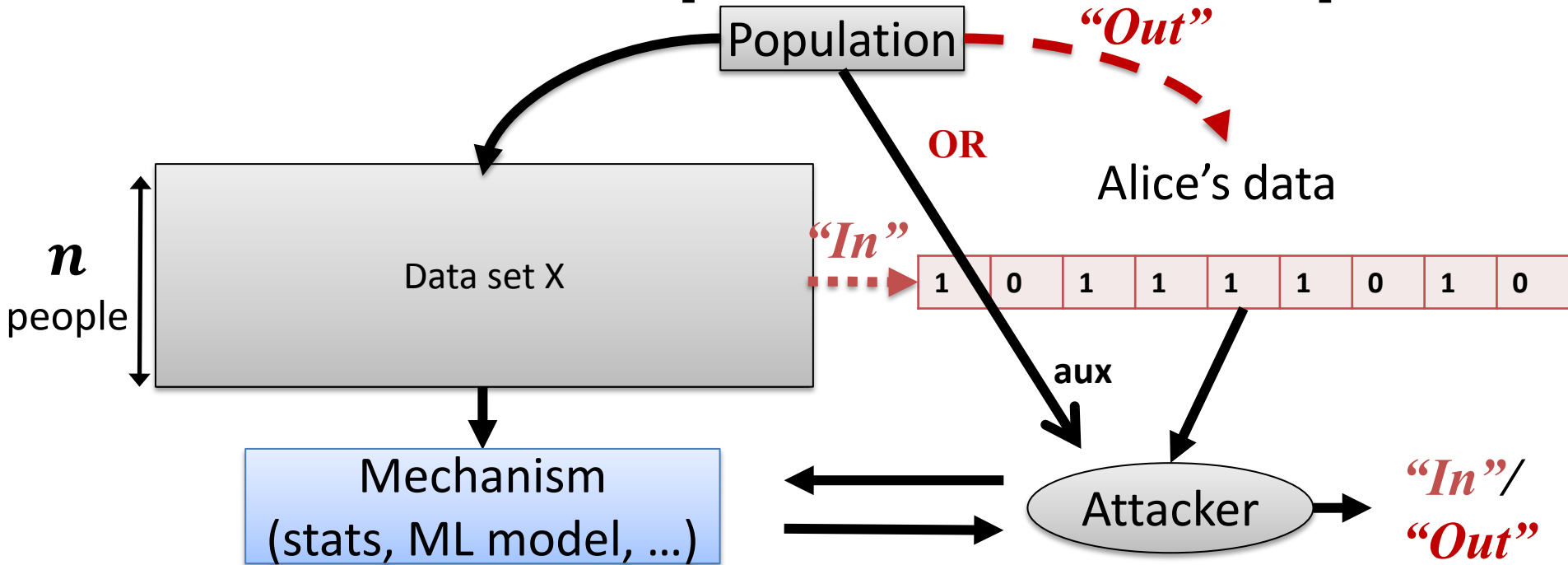
approximate m averages each to within $\pm O\left(\underbrace{\left(\frac{1}{\sqrt{k}}\right)}_{\text{std dev}} \cdot \underbrace{\sqrt{\log m}}_{\text{concentration}}\right)$

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”.

PAUSE FOR INTRODUCTIONS

Membership Attacks: Setup

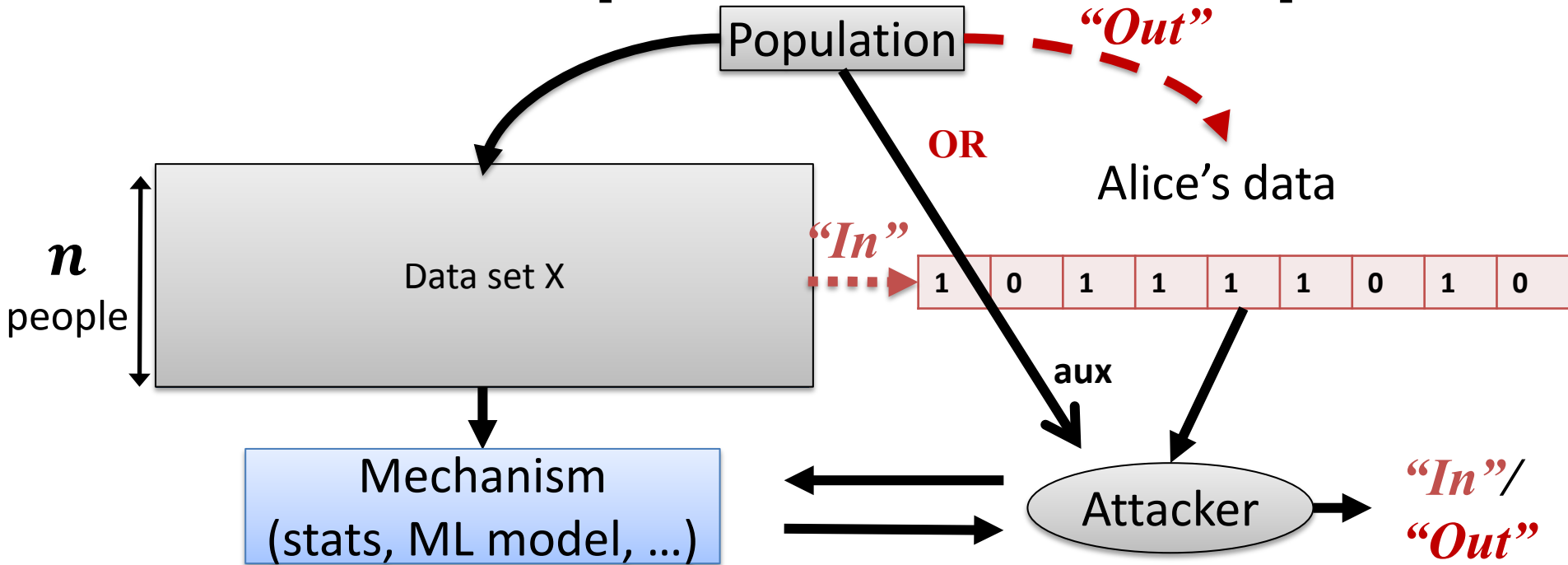


Attacker gets:

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

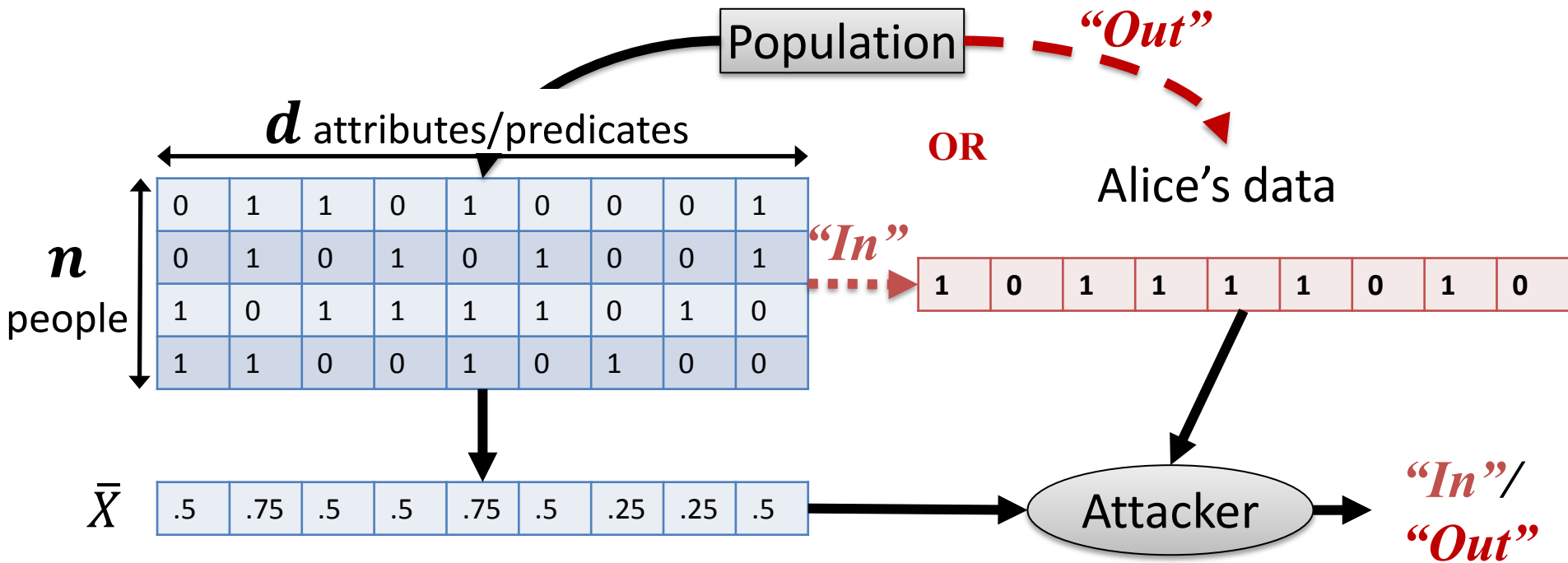
Then decides: if Alice is in the dataset X

Membership Attacks: Examples

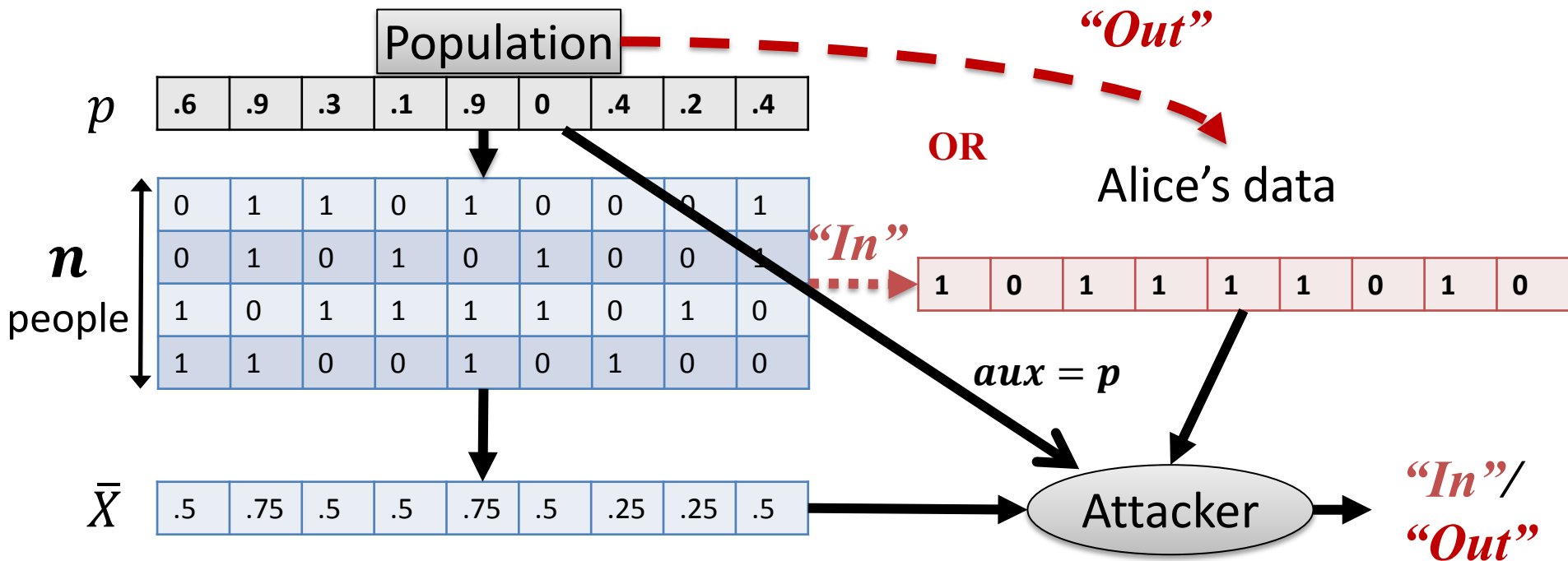


- Genome-wide Association Studies [Homer et al. `08]
 - release frequencies of SNP's (individual positions)
 - determine whether Alice is in "case group" [w/a particular diagnosis]
- ML as a service [Shokri et al. `17]
 - apply models trained on X to Alice's data

Membership Attacks from Means

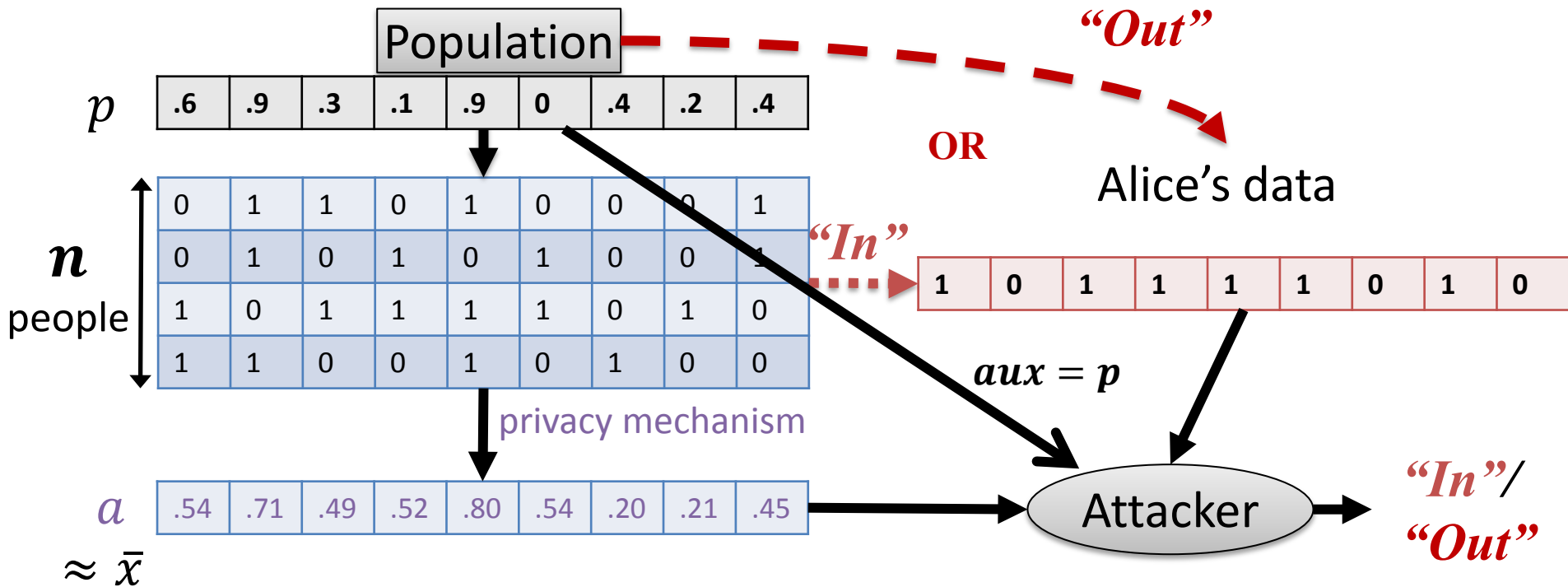


Membership Attacks from Means



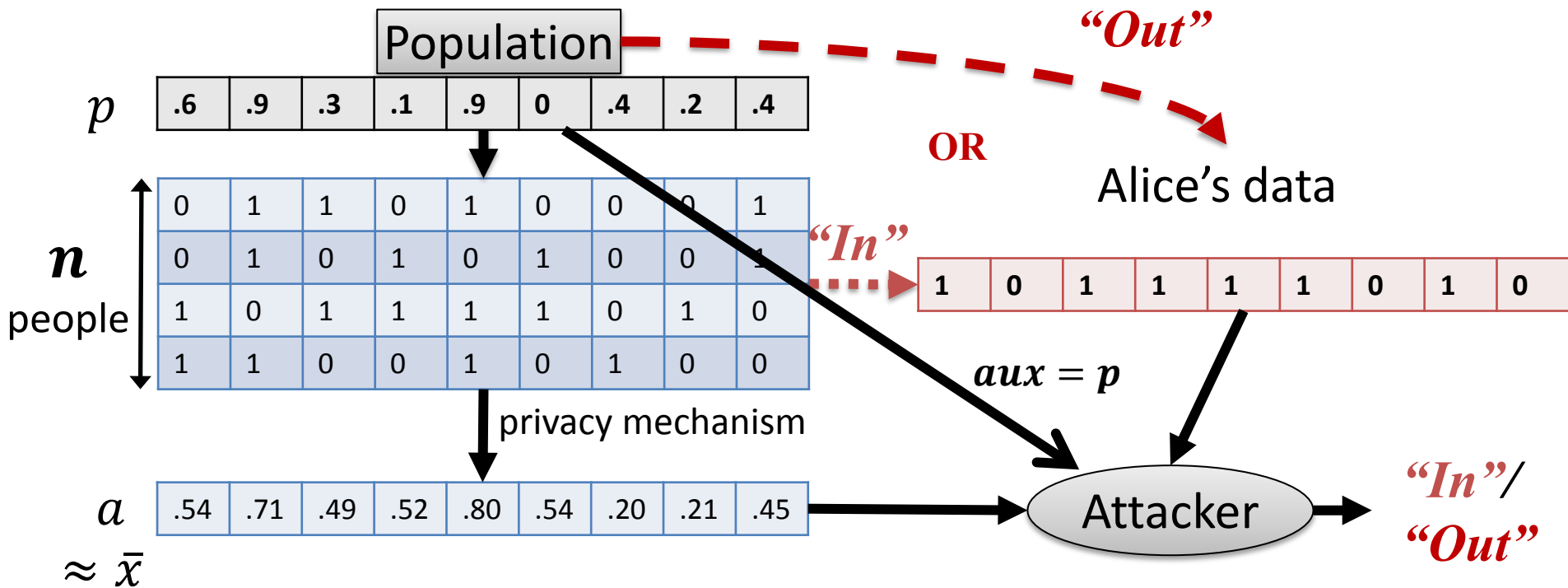
- Population = [vector $p = (p_1, \dots, p_d)$ of probabilities]
 - j 'th attribute = iid Bernoulli(p_j), independent across j
 - Adversary gets p (or a few random draws) given to adversary

Membership Attacks from Noisy Means



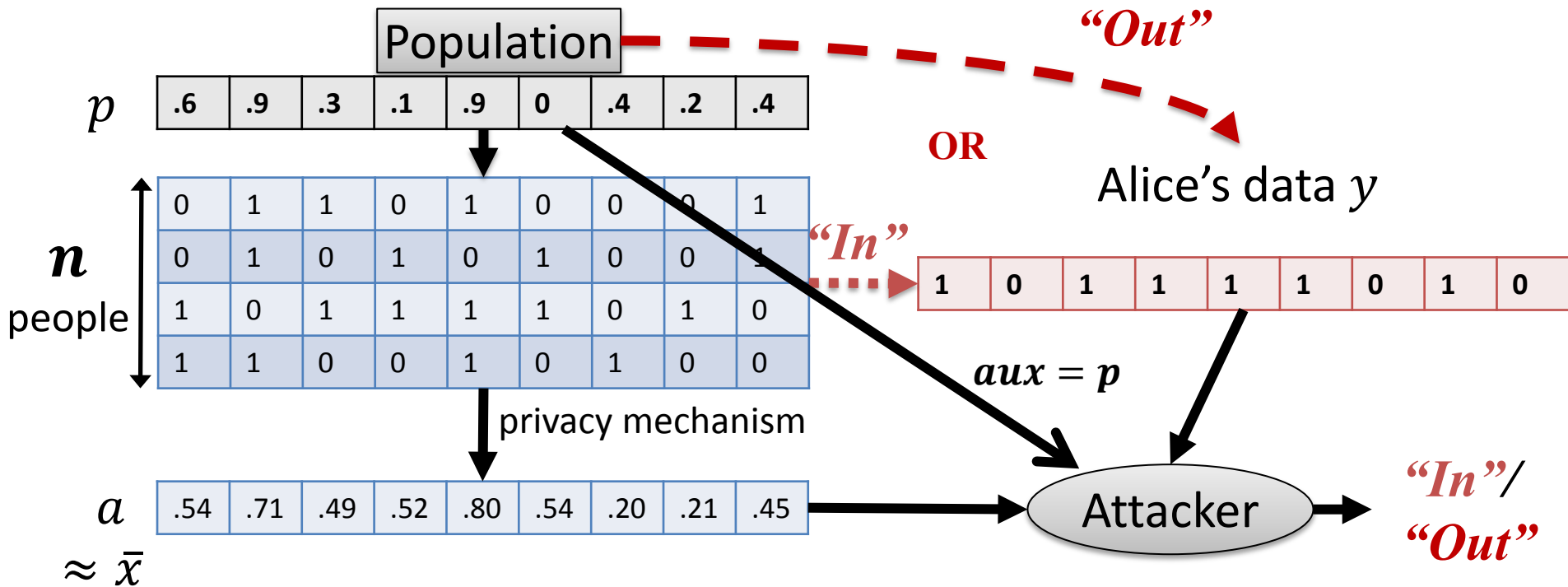
- Population = vector $p = (p_1, \dots, p_d)$ of probabilities
 - j 'th attribute = iid Bernoulli(p_j), independent across j
 - Adversary gets $a \approx \bar{x}$ and p (or a few random draws)
 - Only assume that $a = M(x)$ has $|a_j - \bar{x}_j| \leq \alpha$ whp. ("Noise" need not be independent or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

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 $[0,1]$).

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 $\alpha < \min \left\{ \sqrt{d / O(n^2 \log(1/\delta))}, 1/2 \right\}$:

- 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 $\Pr[A(y, a, p) = \text{IN}] \leq \delta$.

Membership Attacks from Noisy Means

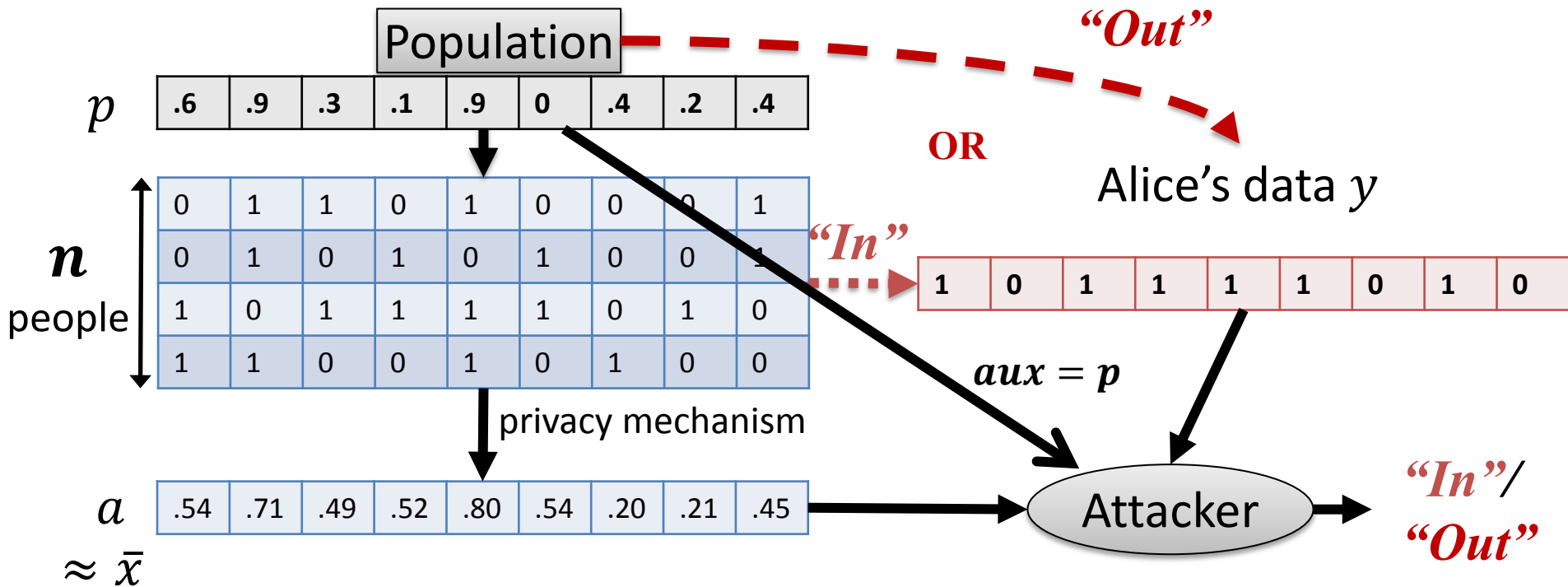
Theorem [Dwork et al. '15]: There is an attacker A such that when $d \geq O(n)$ and $\alpha < \min \left\{ \sqrt{d/O(n^2 \log(1/\delta))}, 1/2 \right\}$:

- If Alice is IN, then $\Pr[A(y, a, p) = \text{IN}] \geq \Omega\left(\frac{1}{\alpha^2 n}\right)$. (true positive)
- If Alice is OUT, then $\Pr[A(y, a, p) = \text{IN}] \leq \delta$. (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?)
- Can safely release at most $\tilde{O}(n^2)$ means!

The Attacker



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

Note: given p, a , can choose $T = T_{p,a} = 0 \left(\sqrt{d \log(1/\delta)} \right)$ to make false positive probability exactly δ .

Attacks on Aggregate Stats

- 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: $\alpha \lesssim \sqrt{d}/n$
- Lessons
 - “Too many, too accurate” statistics reveal individual data
 - “Aggregate” is hard to pin down



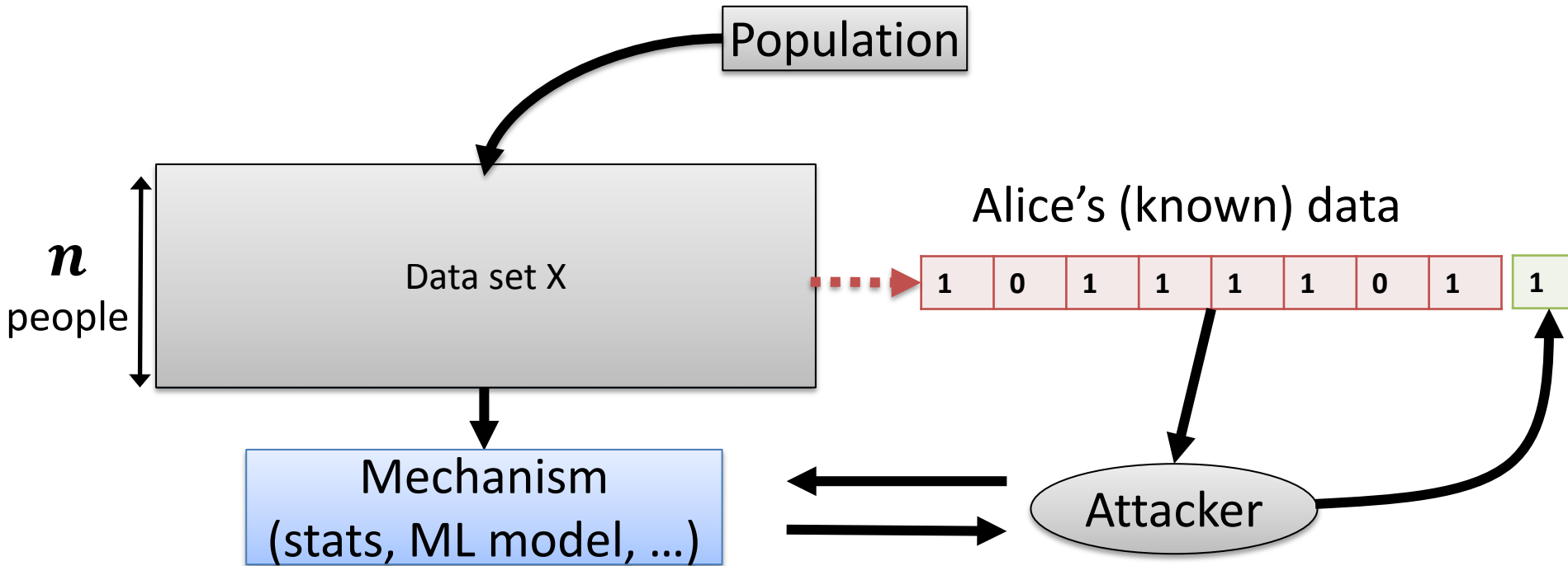
Membership Attacks on ML as a Service

[Shokri et al. 2017]

Switch to slides from Reza Shokri's talk

Another Attack on ML?

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



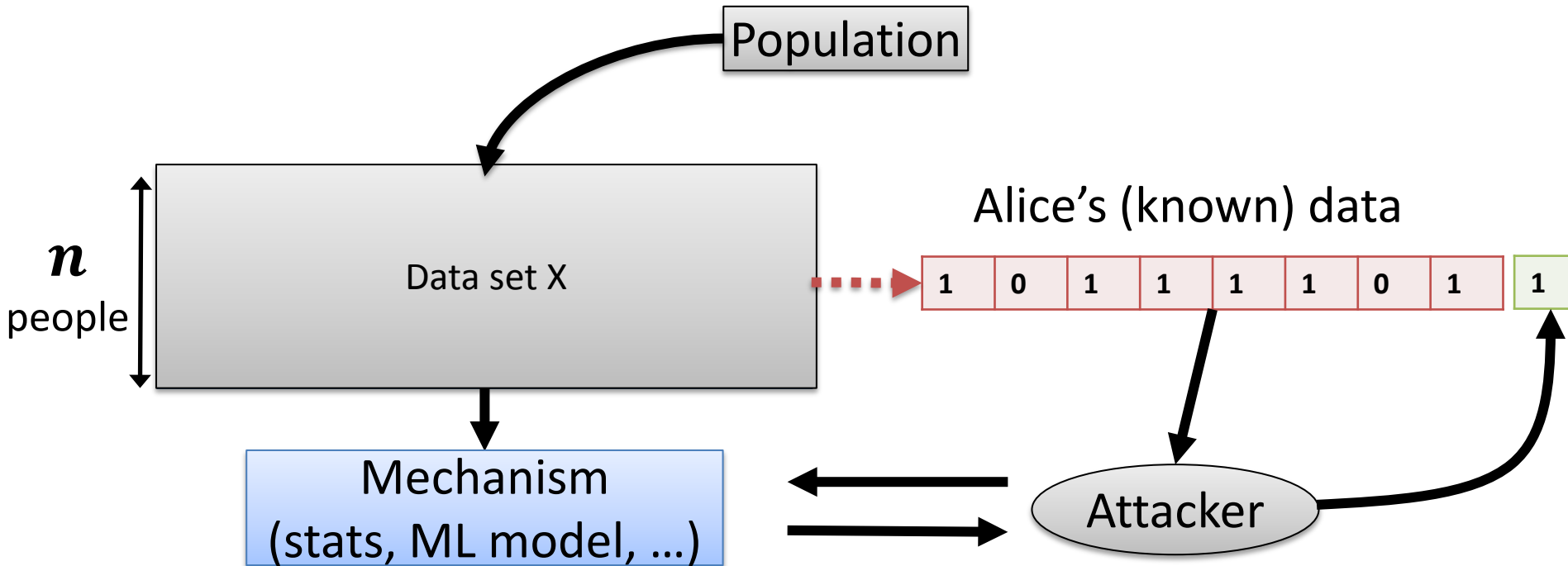
Attacker gets:

- 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.
- Q:** Can it help with privacy in microtargetted advertising?
[Korolova attacks]
- inference from impressions?
 - inference from clicks?
 - displaying intrusive ads?

“Five Views” Responses to Membership Attacks on GWAS