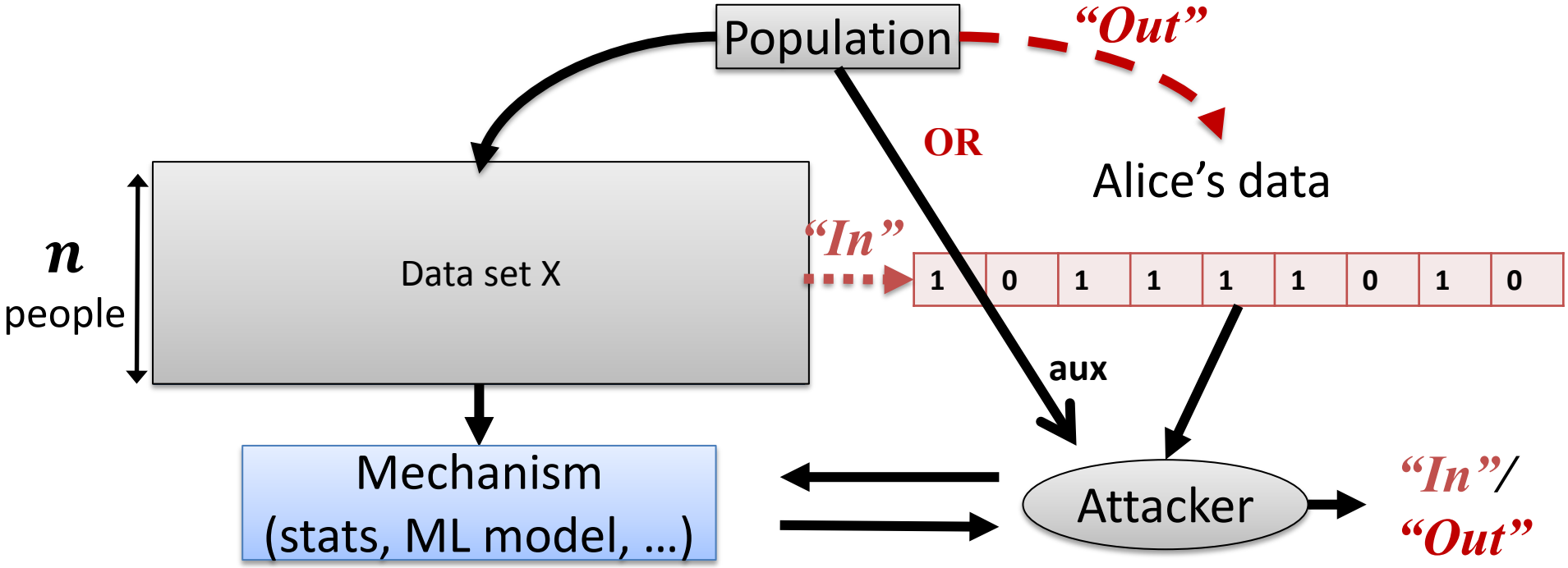
CS208: Applied Privacy for Data Science
Membership & Other Attacks (cont.)
And Introduction to Differential Privacy

James Honaker & Salil Vadhan

School of Engineering & Applied Sciences
Harvard University

February 15, 2019

Recap: Membership Attacks



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]

Attacks on Aggregate Stats

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

[slide based on one from Adam Smith]

Reconstruction vs. Membership

- Reconstruction Attack ⇒ Membership Attack
  - Take sensitive bit = 1 iff in dataset.
  - Use form of reconstruction attack that only requires knowing identifier for person being attacked (PS1 bonus).
  - Reconstruction failure probability bounds false positive and false negative probabilities.
- Membership Attack ⇒ Reconstruction Attack
  - Test membership in sub-datasets where sensitive bit is 0, and where sensitive bit is 1.
  - Pr[reconstruct correctly] ≈ true positive prob.
  - Pr[reconstruct incorrectly] ≈ false positive prob
  - Reconstruction fails (⊥) if both tests say “OUT”.

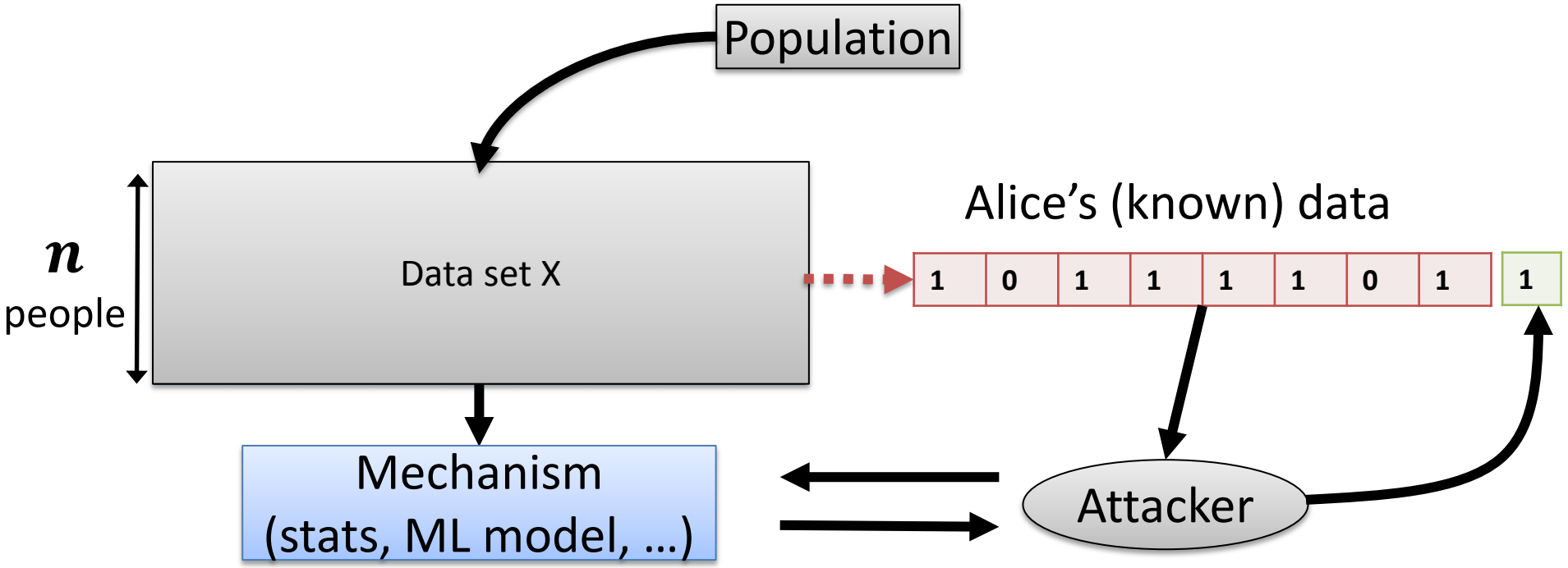
Membership Attacks on ML as a Service

[Shokri et al. 2017]

Switch to slides from Reza Shokri's talk

Another Attack on ML?

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



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.

“Five Views” Responses to Membership Attacks on GWAS

Some points raised:

- Limiting access to credentialed researchers
- Informed consent
- Privacy vs. utility
- Individual vs. group privacy
- Making reidentication illegal
- Maintaining trust and participation

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.

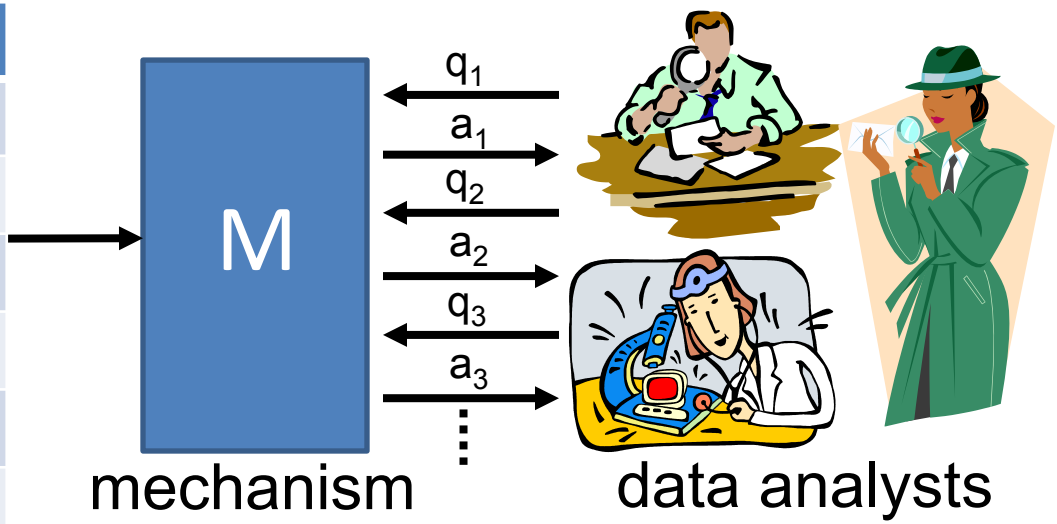
Q: Can it help with privacy in microtargetted advertising? [Korolova attacks]

- inference from impressions?
- inference from clicks?
- displaying intrusive ads?

Differential privacy

[Dinur-Nissim ’03+Dwork, Dwork-Nissim ’04, Blum-Dwork-McSherry-Nissim ’05, Dwork-McSherry-Nissim-Smith ’06]

Sex	Blood	…	HIV?
F	B	…	Y
M	A	…	N
M	O	…	N
M	O	…	Y
F	A	…	N
M	B	…	Y

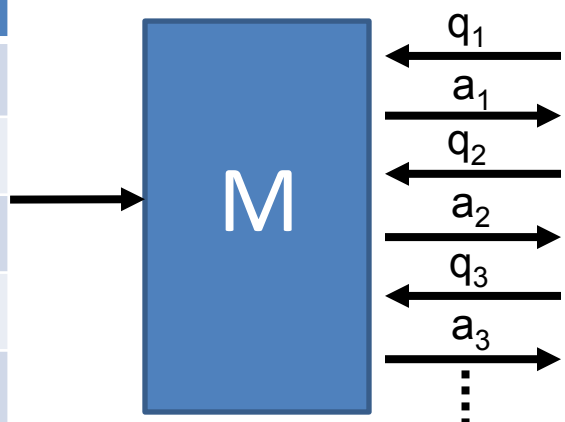


Requirement: effect of each individual should be “hidden”

Differential privacy

[Dinur-Nissim ’03+Dwork, Dwork-Nissim ’04, Blum-Dwork-McSherry-Nissim ’05, Dwork-McSherry-Nissim-Smith ’06]

Sex	Blood	…	HIV?
F	B	…	Y
M	A	…	N
M	O	…	N
M	O	…	Y
F	A	…	N
M	B	…	Y

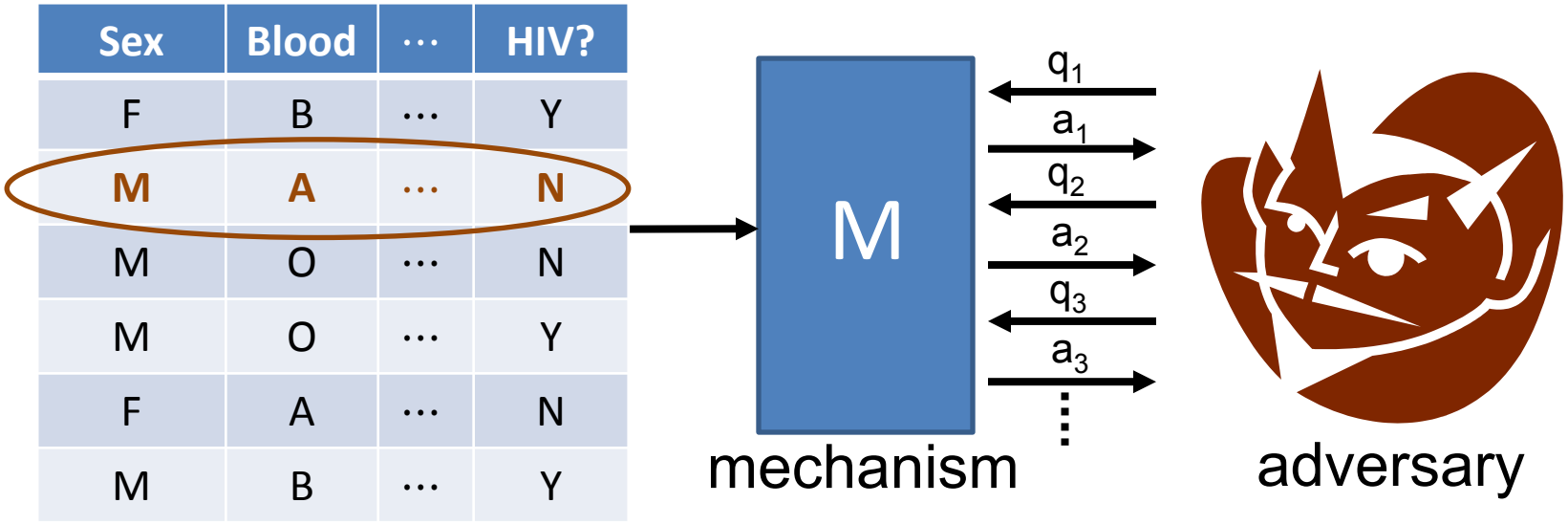


mechanism

adversary

Differential privacy

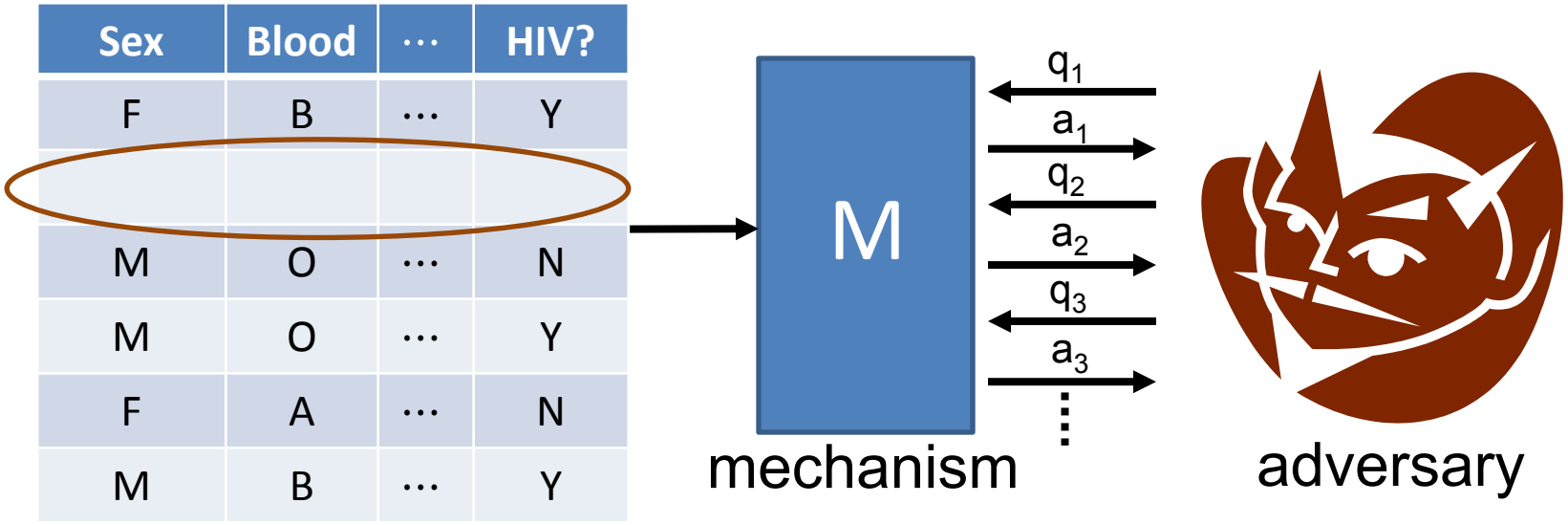
[Dinur-Nissim ’03+Dwork, Dwork-Nissim ’04, Blum-Dwork-McSherry-Nissim ’05, Dwork-McSherry-Nissim-Smith ’06]



Requirement: an adversary shouldn’t be able to tell if any one person’s data were changed arbitrarily

Differential privacy

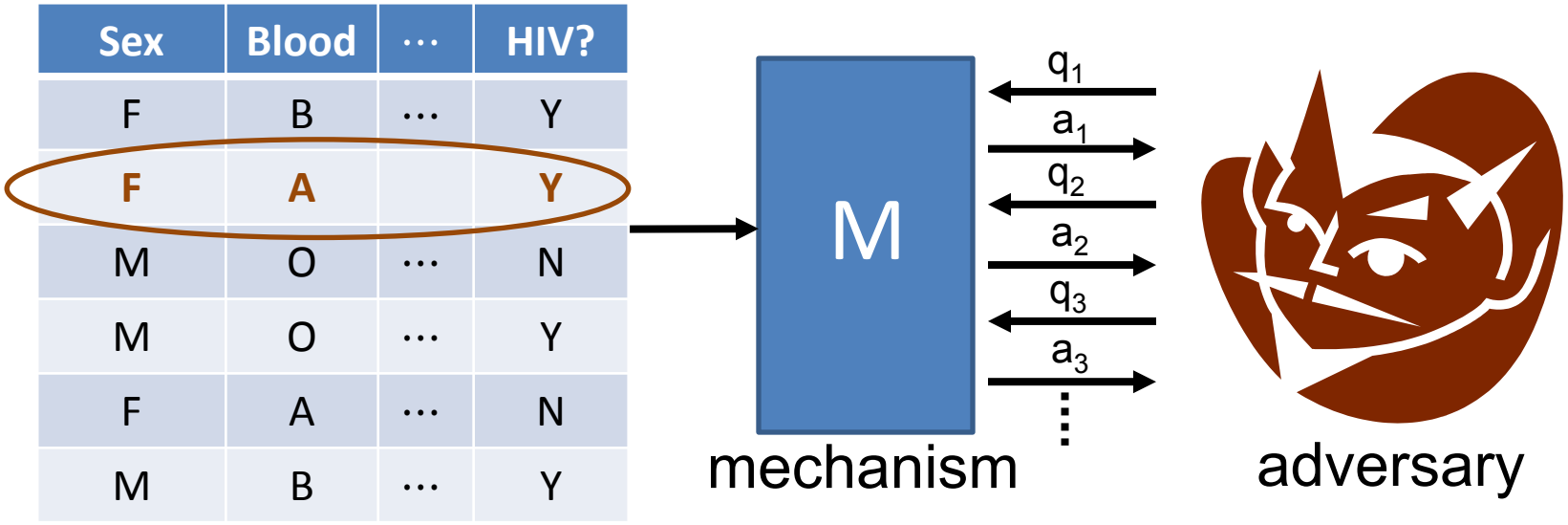
[Dinur-Nissim ’03+Dwork, Dwork-Nissim ’04, Blum-Dwork-McSherry-Nissim ’05, Dwork-McSherry-Nissim-Smith ’06]



Requirement: an adversary shouldn’t be able to tell if any one person’s data were changed arbitrarily

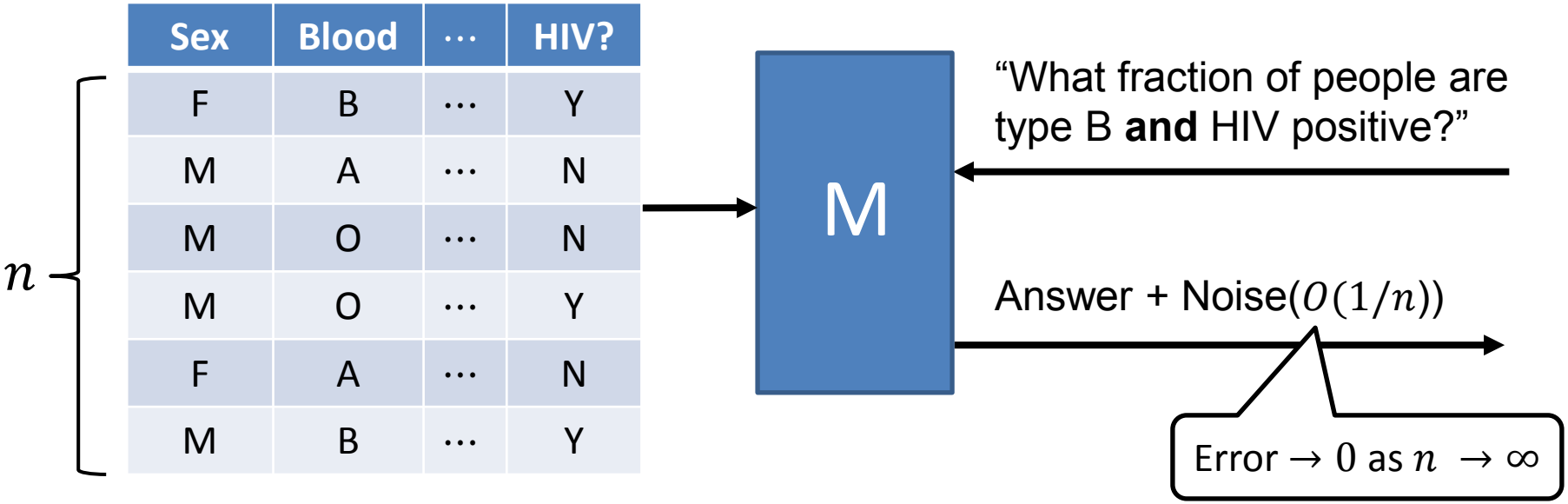
Differential privacy

[Dinur-Nissim ’03+Dwork, Dwork-Nissim ’04, Blum-Dwork-McSherry-Nissim ’05, Dwork-McSherry-Nissim-Smith ’06]



Requirement: an adversary shouldn’t be able to tell if any one person’s data were changed arbitrarily

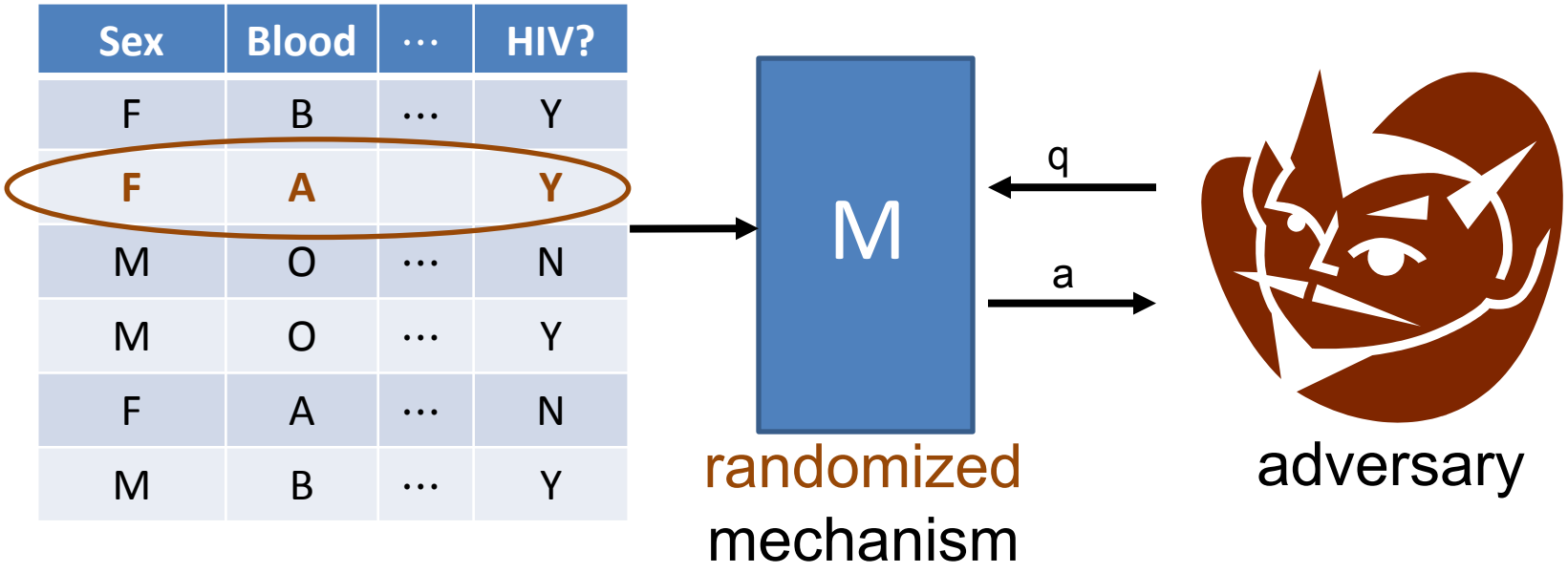
Simple approach: random noise



- Very little noise needed to hide each person as $n \to \infty$.
- **Note:** this is just for one query

DP for one query/release

[Dinur-Nissim ’03+Dwork, Dwork-Nissim ’04, Blum-Dwork-McSherry-Nissim ’05, Dwork-McSherry-Nissim-Smith ’06]

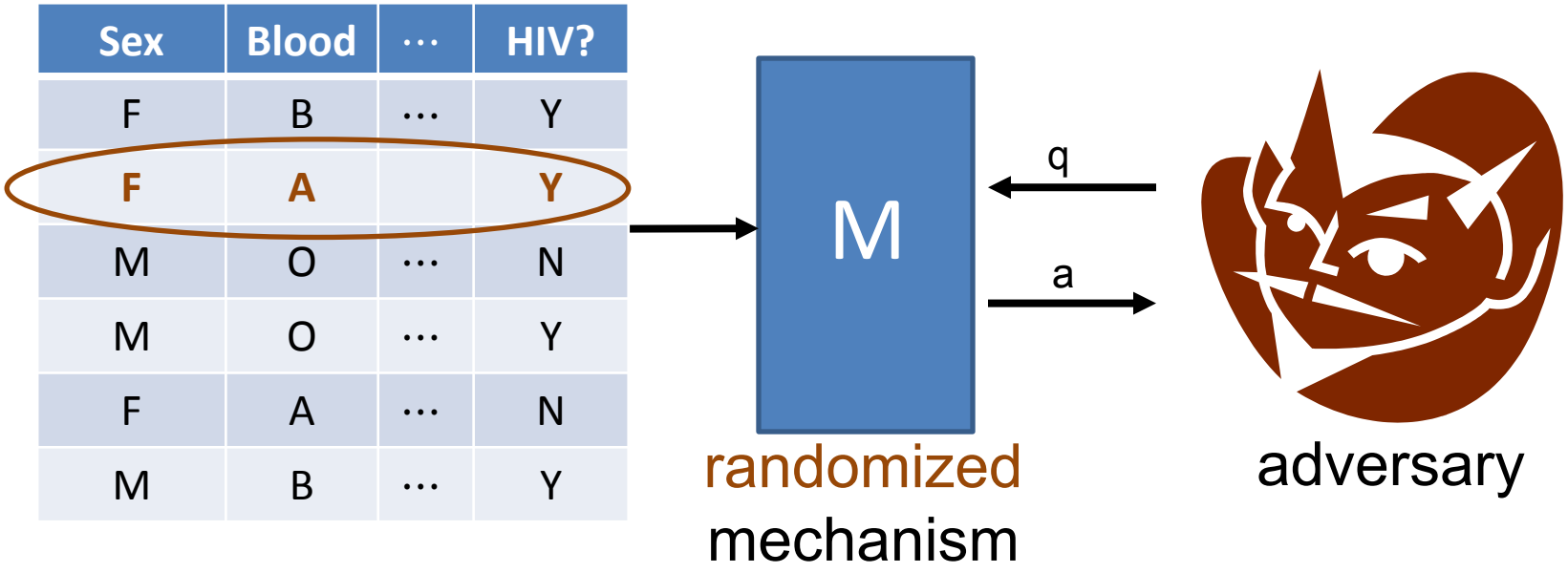


Requirement: for all D, D’ differing on one row, and all q

Distribution of M(D,q) $\approx_\varepsilon$ Distribution of M(D’,q)

DP for one query/release

[Dinur-Nissim ’03+Dwork, Dwork-Nissim ’04, Blum-Dwork-McSherry-Nissim ’05, Dwork-McSherry-Nissim-Smith ’06]

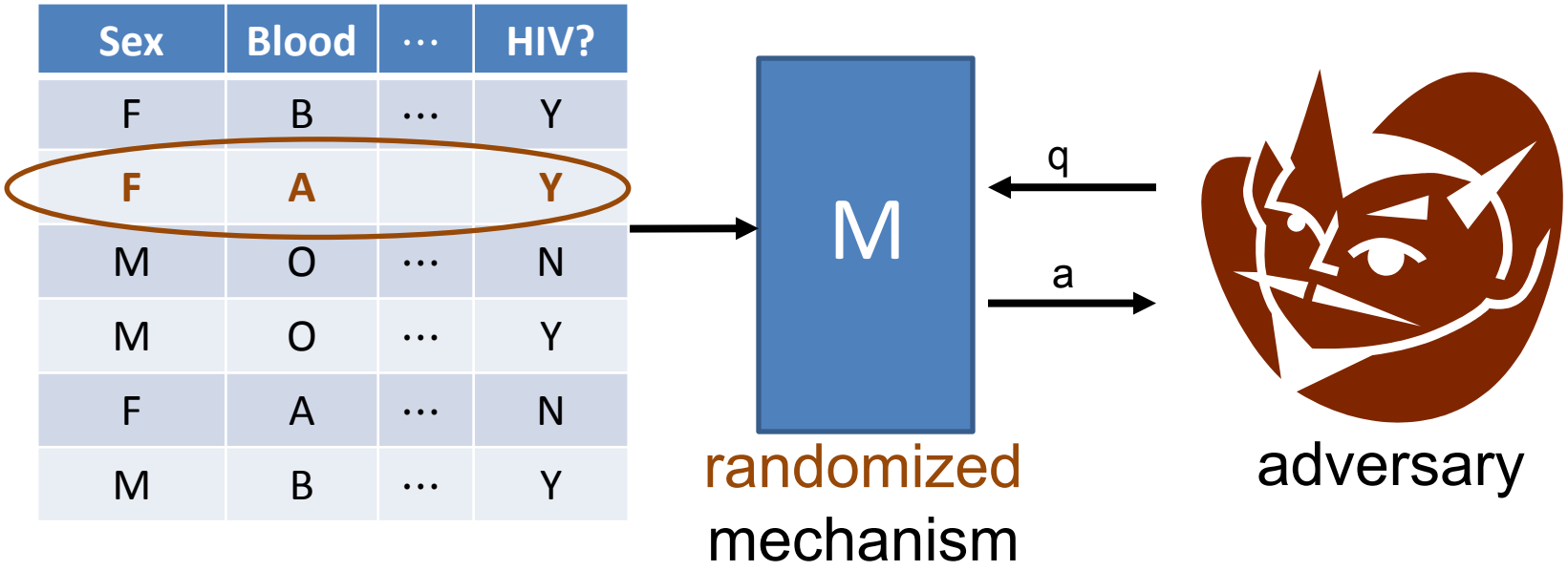


Requirement: for all D, D’ differing on one row, and all q

$\forall$ sets T, $\Pr[M(D,q) \in T] \lesssim (1+\varepsilon) \cdot \Pr[M(D',q) \in T]$

DP for one query/release

[Dwork-McSherry-Nissim-Smith '06]



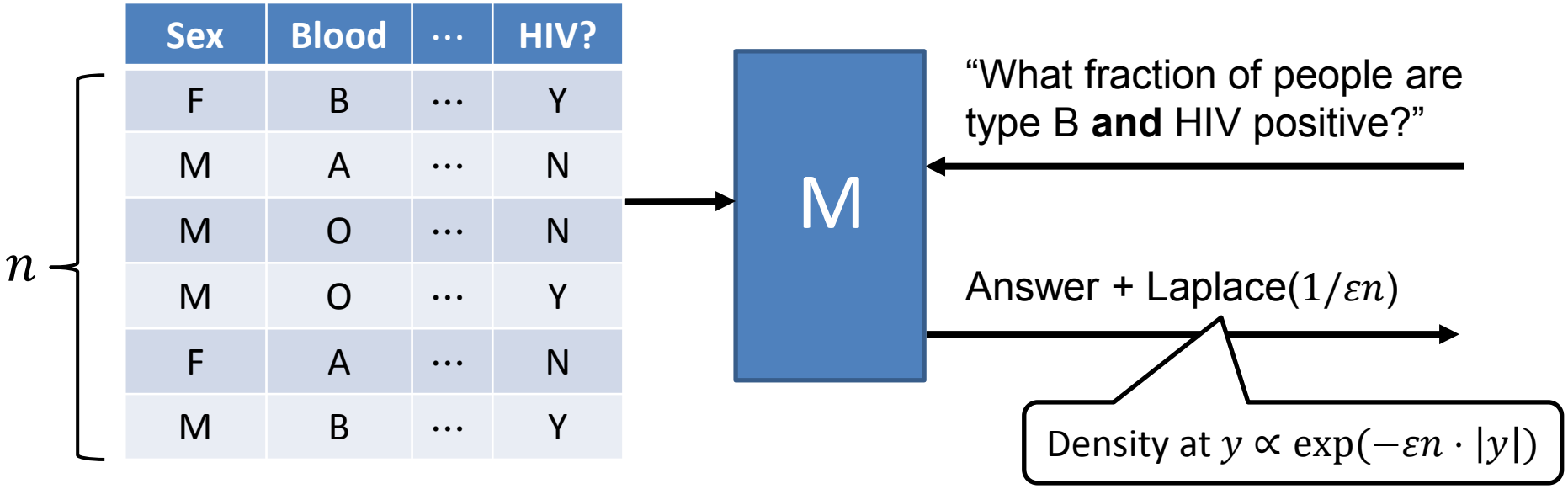
Def: M is ε-DP if for all D, D' differing on one row, and all q

$\forall$ sets T, $\Pr[M(D,q) \in T] \leq e^{\varepsilon} \cdot \Pr[M(D',q) \in T]$

(Probabilities are (only) over the randomness of M.)

The Laplace Mechanism

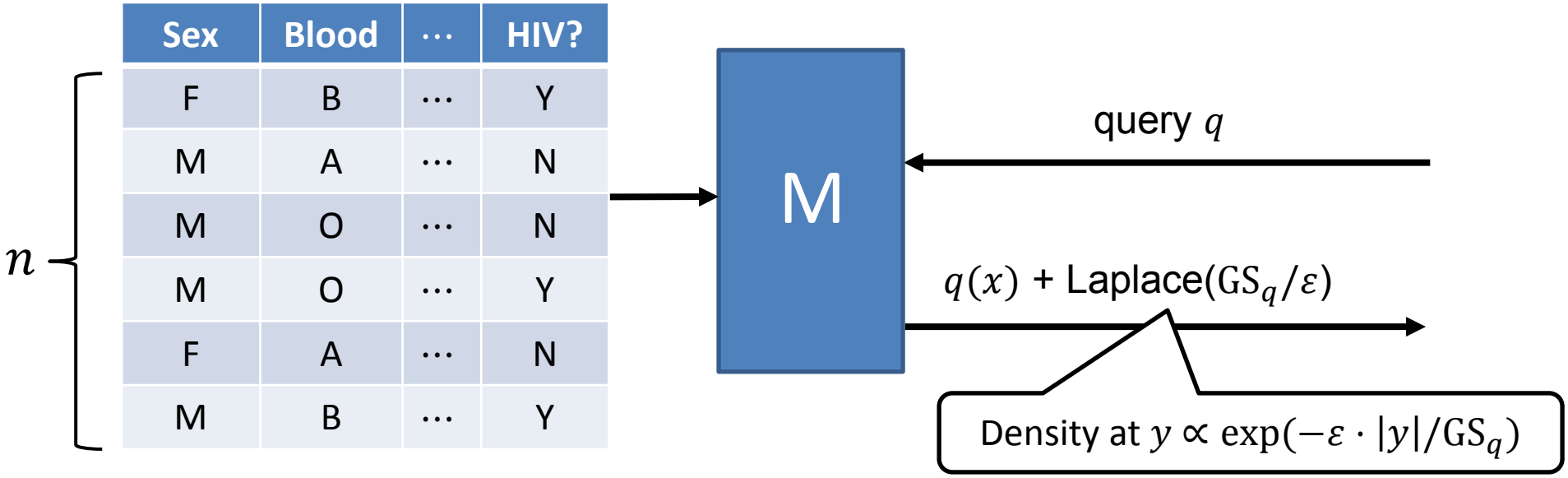
[Dwork-McSherry-Nissim-Smith '06]



- Very little noise needed to hide each person as $n \to \infty$.

The Laplace Mechanism

[Dwork-McSherry-Nissim-Smith '06]



- Very little noise needed to hide each person as $n \to \infty$.

The Laplace Mechanism

[Dwork-McSherry-Nissim-Smith '06]

- Let $\mathcal{X}$ be a data universe, and $\mathcal{X}^n$ a space of datasets. (For now, we are treating n as known and public.)
- For $x, x' \in \mathcal{X}^n$, write $x \sim x'$ if x and x' differ on at one row.
- For a query $q : \mathcal{X}^n \to \mathbb{R}$, the global sensitivity is
$$\mathrm{GS}_q = \max_{x \sim x'} |q(x) - q(x')| .$$
- The Laplace distribution with scale s, $\mathrm{Lap}(s)$:
  - Has density function $f(y) = e^{-|y|/s}/2s$.
  - Mean 0, standard deviation $\sqrt{2} \cdot s$.

Theorem: $M(x, q) = q(x) + \mathrm{Lap}(\mathrm{GS}_q/\varepsilon)$ is ε-DP.

Calculating Global Sensitivity

1. $\mathcal{X} = \{0,1\}$, $q(x) = \sum_{i=1}^{n} x_i$, $\mathrm{GS}_q =$
2. $\mathcal{X} = \mathbb{R}$, $q(x) = \sum_{i=1}^{n} x_i$, $\mathrm{GS}_q =$
3. $\mathcal{X} = [0,1]$, $q(x) = \mathrm{mean}(x_1, x_2, \ldots, x_n)$, $\mathrm{GS}_q =$
4. $\mathcal{X} = [0,1]$, $q(x) = \mathrm{median}(x_1, x_2, \ldots, x_n)$, $\mathrm{GS}_q =$
5. $\mathcal{X} = [0,1]$, $q(x) = \mathrm{variance}(x_1, x_2, \ldots, x_n)$, $\mathrm{GS}_q =$

Q: for which of these queries is the Laplace Mechanism “useful”?

Proof that the Laplace Mechanism is Differentially Private