CS208: Applied Privacy for Data Science
One-Shot Releases: Synthetic Data & the 2020 US Census

School of Engineering & Applied Sciences
Harvard University

February 21, 2022

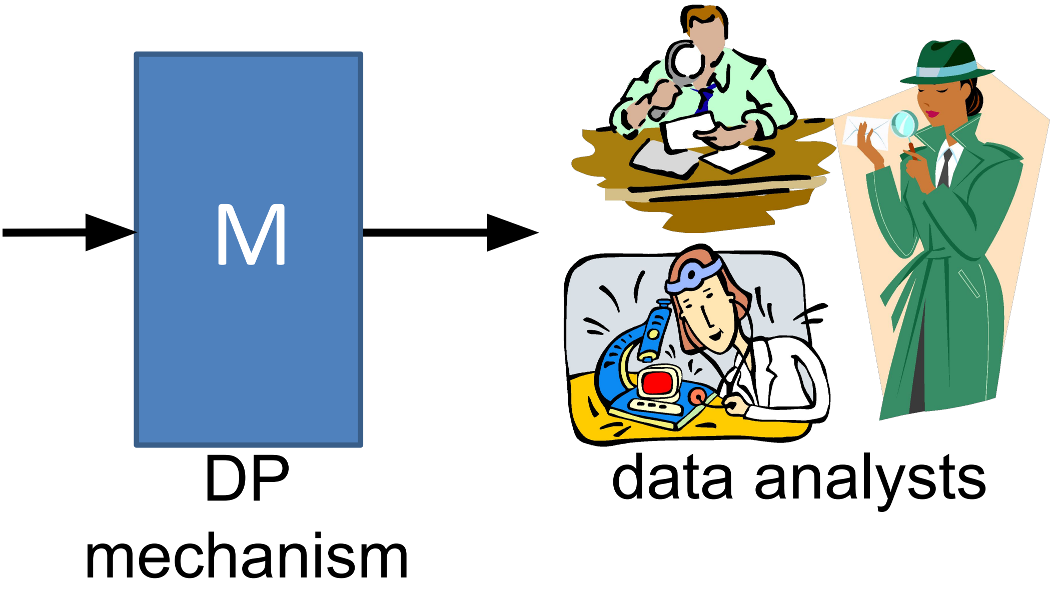
Mid-term Feedback

- Problem Sets have been intense
  - Will be easing up as we move into the project-focused portion of course
- Many students don’t know others, working alone
  - Intros today, class mixer coming
- Confusions about where to find course materials
  - Before class: incomplete version of slides posted on Ed, incomplete notebooks in repo
  - After class: all materials posted on course website
- Want walk-throughs of problems
  - Done in section
- Finding OH, Section times
  - Course Google calendar

One-Shot Releases

n

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



M

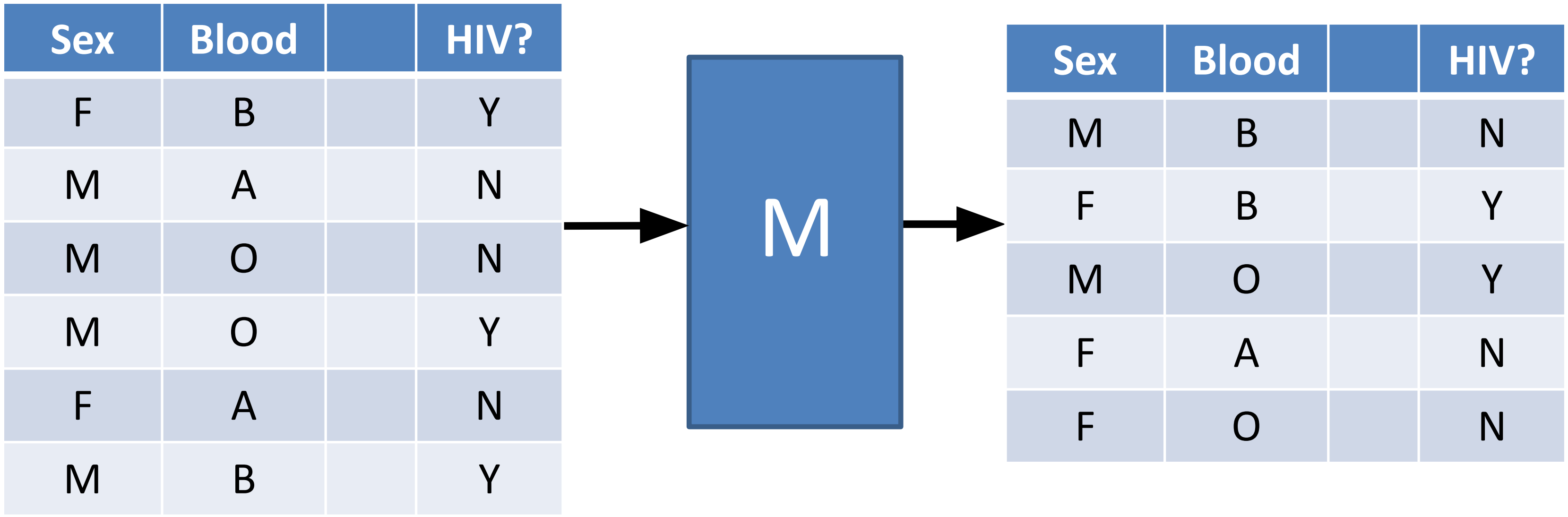
DP mechanism

data analysts

Goal: release as much useful info as possible given privacy budget

- Ideally support unforeseen analyses
- Summary statistics
- ML model
- Synthetic data

Differentially Private Synthetic Data



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

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

- $M\colon \mathcal{X}^n \to \mathcal{X}^m$ such that:
  - $M(x)$ has the same syntax as a real dataset.
  - $M(x)$ reflects many statistical properties of x.
  - M is differentially private.

Synthetic Data via DP Histograms

- Use singleton bins $B_y = \{y\}$ for each $y \in \mathcal{Y}$.
- Construct a DP histogram $\left(a_1, \ldots, a_{|\mathcal{X}|}\right) \leftarrow M_{\text{hist}}(x)$, where $a_y \approx \#\{i : x_i = y\}$.
- Output synthetic dataset $\hat{x}$ with a_y copies of each element y.

Difficulties?

Synthetic Data via DP Histograms

- Use singleton bins $B_y = \{y\}$ for each $y \in \mathcal{Y}$.
- Construct a DP histogram $\left(a_1, \ldots, a_{|\mathcal{X}|}\right) \leftarrow M_{\text{hist}}(x)$, where $a_y \approx \#\{i : x_i = y\}$.
- Output synthetic dataset $\hat{x}$ with a_y copies of each element y.

Difficulties?

- a_y's may not be nonnegative integers.
  - Soln 1: use Geometric Mechanism and clamp at 0.
  - Soln 2: use Exponential Mechanism with range $\{0, \ldots, n\}$.
- Poor utility & efficiency when $\mathcal{X}$ is large.

Stability-Based Histogram

1. Let $B_1, \dots, B_k \subseteq \mathcal{X}$ be disjoint bins.
2. Define $q_j : \mathcal{X}^n \to \{0,1\}$ by $q_j(x) = \#\{i : x_i \in B_j\}$.
3. For each j s.t. $q_j(x) > 0$:
   a) Let $a_j = q_j(x) + Z_j$ for $Z_j \sim \mathrm{Geo}(2/\varepsilon)$.
   b) If $a_j > \left\lceil \frac{2}{\varepsilon} \cdot \ln \frac{1}{\delta} \right\rceil$, output (j, a_j).
4. Treat all other bins as having a zero count.

Intuition for (ε, δ)-DP:

- Only difference from pure DP is treatment of zero bins.
- If $q_j(x) = 0$, then $q_j(x') \leq 1$ for any $x' \sim x$, and

$$\Pr\left[1 + Z_j > \left\lceil \frac{2}{\varepsilon} \cdot \ln \frac{1}{\delta} \right\rceil\right] < \delta.$$

Stability-Based Histogram

1. Let $B_1, \ldots, B_k \subseteq \mathcal{X}$ be disjoint bins.
2. Define $q_j : \mathcal{X}^n \to \{0,1\}$ by $q_j(x) = \#\{i : x_i \in B_j\}$.
3. For each j s.t. $q_j(x) > 0$:
    a) Let $a_j = q_j(x) + Z_j$ for $Z_j \sim \mathrm{Geo}(2/\varepsilon)$.
    b) If $a_j > \left\lceil \frac{2}{\varepsilon} \cdot \ln \frac{1}{\delta} \right\rceil$, output (j, a_j).
4. Treat all other bins as having a zero count.

Benefits:

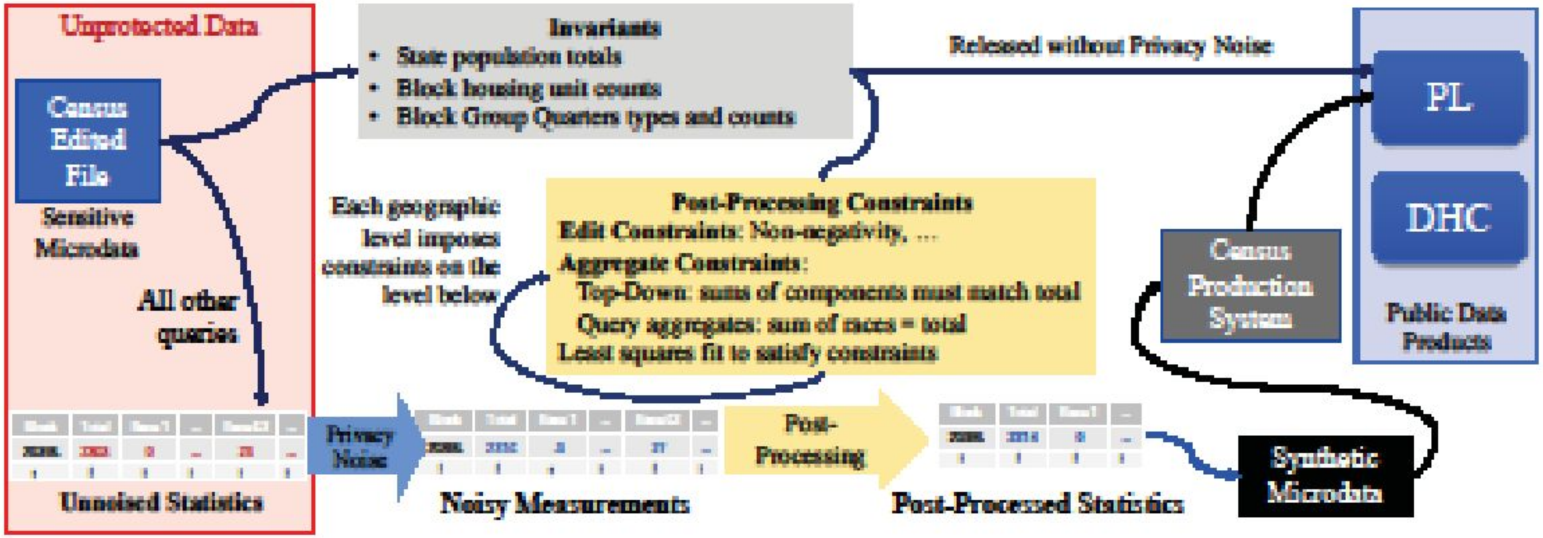
- Computation and output size linear in n rather than $|\mathcal{X}|$.
- Max error $O((1/\varepsilon) \cdot \ln(1/\delta))$ whp, independent of $|\mathcal{X}|$.
- But still can have poor utility when $|\mathcal{X}|$ large. (Why?)

Class Introductions

Please share your:

- Name
- Program
- Interests (academic and non-academic, if you wish)

Census DAS Process



Unprotected Data

Census Edited File

Sensitive Microdata

All other queries

Unnoised Statistics

Invariants
- State population totals
- Block housing unit counts
- Block Group Quarters types and counts

Released without Privacy Noise

Each geographic level imposes constraints on the level below

Post-Processing Constraints
Edit Constraints: Non-negativity, …
Aggregate Constraints:
Top-Down: sums of components must match total
Query aggregates: sum of races = total
Least squares fit to satisfy constraints

Privacy Noise

Noisy Measurements

Post-Processing

Post-Processed Statistics

Synthetic Microdata

Census Production System

PL

DHC

Public Data Products

Figure 5-1: Process used to produce privacy-protected data products.

Consistency and Privacy for the 2020 Census 51 JSR-21-02 January 11, 2022

Consistency & Optimization

- Structural Zeroes: Enforced by edit and imputation, DP can't reintroduce it
  - Householder and spouse/partner must be at least 15 yrs old
  - Every household must have exactly one householder
  - At least one of the binary race flags must be 1
  - Etc.

- Invariants: public statistics with exact values
  - State population totals
  - Linear constraints: sum of county populations equals state population
  - Single-gender group quarters (dorms, prisons)

- Optimizing accuracy: for a set Q of queries
  - Obtain DP answers to a set Q' of "measurement" queries, then use optimization tools to reconstruct synthetic data to optimize answers on Q.

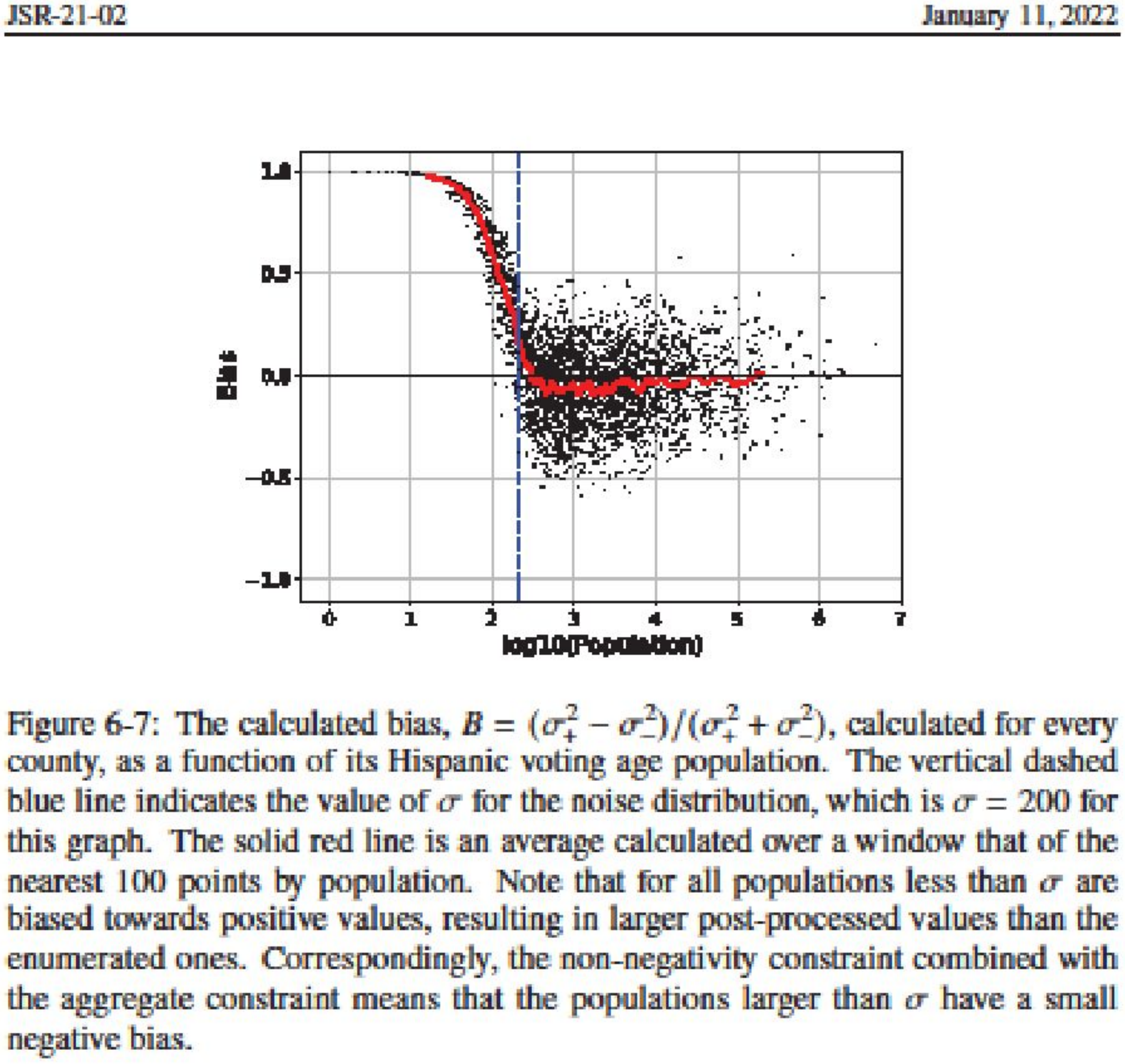
Census Bureau’s Use of DP

Excerpts from:

- Michael Hawes and Michael Ratcliffe. “Understanding the 2020 Census Disclosure Avoidance System: Differential Privacy 201 and the TopDown Algorithm,” Census Webinar, May 13, 2021.

- Michael Hawes. “Understanding the 2020 Census Disclosure Avoidance System: Analysis of Production Settings for Redistricting and Voting Rights Act Use Cases,” Census Webinar, August 10, 2021.

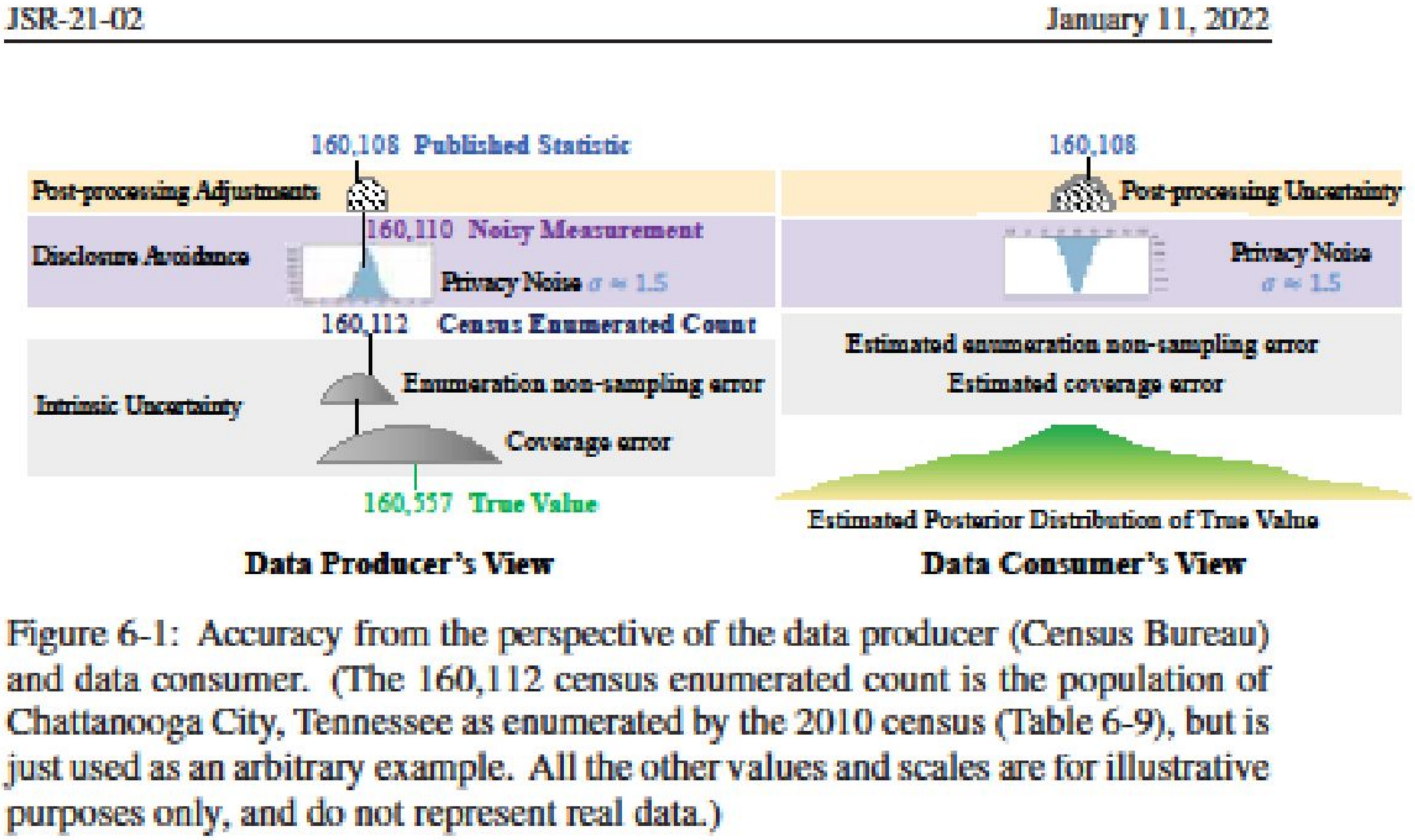
Bias



JSR-21-02 January 11, 2022

Figure 6-7: The calculated bias, $B = (\sigma_+^2 - \sigma_-^2)/(\sigma_+^2 + \sigma_-^2)$, calculated for every county, as a function of its Hispanic voting age population. The vertical dashed blue line indicates the value of σ for the noise distribution, which is $\sigma = 200$ for this graph. The solid red line is an average calculated over a window that of the nearest 100 points by population. Note that for all populations less than σ are biased towards positive values, resulting in larger post-processed values than the enumerated ones. Correspondingly, the non-negativity constraint combined with the aggregate constraint means that the populations larger than σ have a small negative bias.

Data Producer vs. Consumer



JSR-21-02 January 11, 2022

160,108 Published Statistic
Post-processing Adjustments
160,110 Noisy Measurement
Disclosure Avoidance
Privacy Noise σ ≈ 1.5
160,112 Census Enumerated Count
Enumeration non-sampling error
Intrinsic Uncertainty
Coverage error
160,557 True Value
Data Producer's View
160,108
Post-processing Uncertainty
Privacy Noise
σ ≈ 1.5
Estimated enumeration non-sampling error
Estimated coverage error
Estimated Posterior Distribution of True Value
Data Consumer's View

Figure 6-1: Accuracy from the perspective of the data producer (Census Bureau) and data consumer. (The 160,112 census enumerated count is the population of Chattanooga City, Tennessee as enumerated by the 2010 census (Table 6-9), but is just used as an arbitrary example. All the other values and scales are for illustrative purposes only, and do not represent real data.)

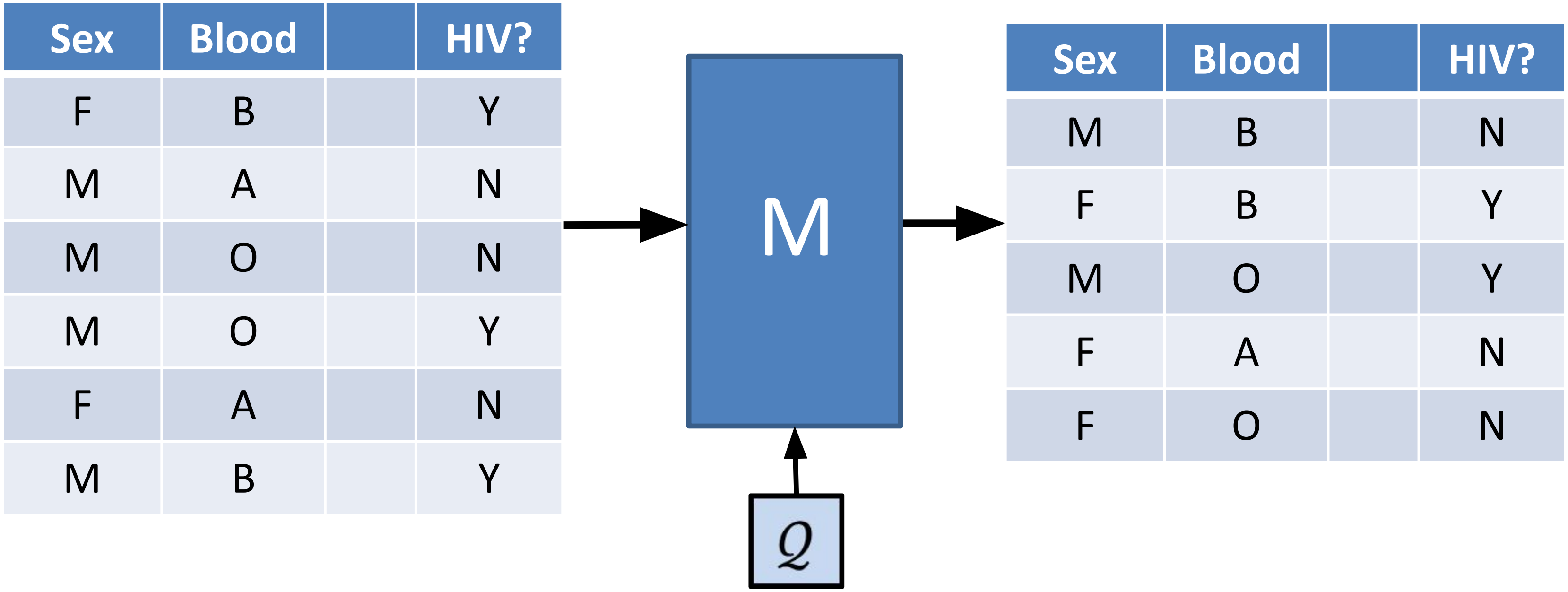
Some challenges raised by JASON

- Detailed queries not used directly in published statistics but consuming privacy-loss budget
- Release of noisy measurements
- Consistency between and within data products
- Threats, risks, and protections not sufficiently quantified
- Better communication with data users
- Re-interpretation of Title 13

STATE OF ART SYNTHETIC DATA GENERATION

Private Multiplicative Weights

[Blum-Ligett-Roth `08,…,Hardt-Rothblum `10]



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

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

- (ε,δ)-DP $M\colon\ \mathcal{X}^n \to \mathcal{X}^m$ such that $\forall q \in \mathcal{Q},\ \ q\colon \mathcal{X} \to [0{,}1]$

$$\left|\frac{1}{n}\sum_{i=1}^{n} q(x_i) - \frac{1}{m}\sum_{i=1}^{m} q(M(x)_i)\right| \le O\left(\frac{\sqrt{\log|\mathcal{X}|\cdot\log(1/\delta)}\cdot\log|\mathcal{Q}|}{\varepsilon n}\right)^{1/2}$$

Private Multiplicative Weights

[Hardt-Rothblum `10]

- (ε,δ)-DP $M\colon \mathcal{X}^n \to \mathcal{X}^m$ such that $\forall q \in \mathcal{Q},\ q\colon \mathcal{X} \to [0,1]$

$$\left|\frac{1}{n}\sum_{i=1}^{n} q(x_i) - \frac{1}{m}\sum_{i=1}^{m} q(M(x)_i)\right| \le O\left(\frac{\sqrt{\log|\mathcal{X}|\cdot\log(1/\delta)}\cdot\log|\mathcal{Q}|}{\varepsilon n}\right)^{1/2}$$

Approach:

- DP online learning of a synthetic data distribution, playing against a "query" player trying to distinguish it from dataset

Problem: computation time $\mathrm{poly}(n,|\mathcal{X}|,|\mathcal{Q}|)$.

- Exponential in dimensionality of data and query family.
- Inherent in the worst case (cf. "Complexity of DP").

Practical Approaches

- Use DP queries to learn a model of the data distribution, and use the model to generate synthetic data

- Models (from more structured to less):
  - Multivariate Gaussian (↔ means and (co)variances)
  - Graphical models/Bayes Nets
  - Generative Adversarial Networks

- See [Liu-Vietri-Wu NeurIPS `21] and references therein.