Enabling Third-Party Audits of Algorithmic Systems with Privacy Enhancing Technologies

2024 FCSM Research and Policy Conference
10/24/24

Michael Walton
Tomo Lazovich, Ph.D. (they/them)

xD, U.S. Census Bureau

All statements are the author's personal views and do not necessarily reflect Census Bureau policy.

xD
https://www.xd.gov

United States Census Bureau

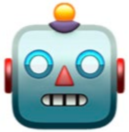
xD is an **emerging technologies group** that's advancing the delivery of data-driven services through new and transformative technologies.

*We do this work by bringing on cohorts of **Emerging Technology Fellows** and by collaborating with others throughout the Census Bureau and beyond!*

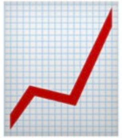
PETs + Responsible AI

PETs
Technologies that **enable** data analysis while "mitigating privacy risks arising from data processing" -EO 14110

PETs-enabled 3rd Party Audits



Responsible AI
Methods for building and **evaluating** AI/ML systems that account for **real-world impacts**



Third-Party Audits of Algorithmic Systems

Fairness

Identify & mitigate potentially harmful biases

Transparency

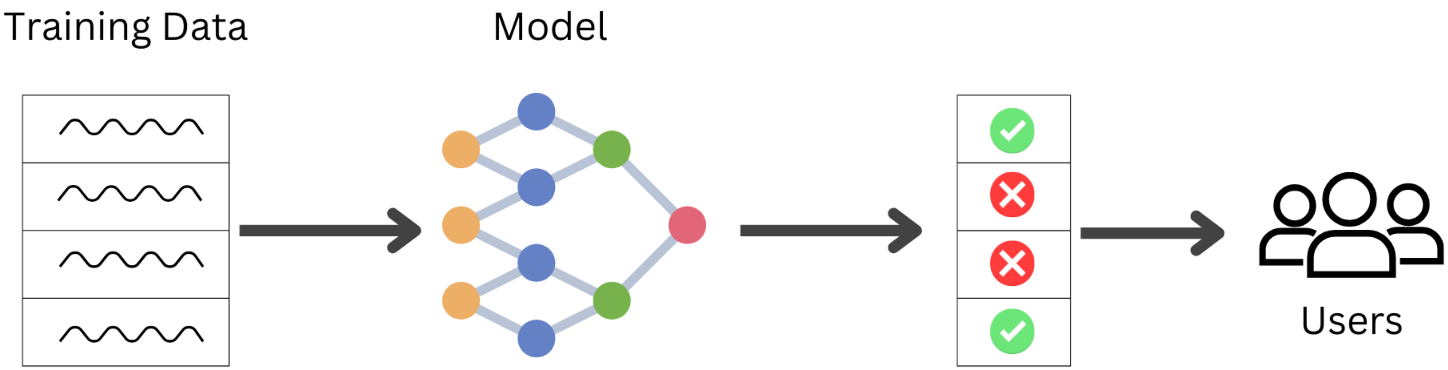
Insights into model decision-making processes

Utility

Independently validate performance claims

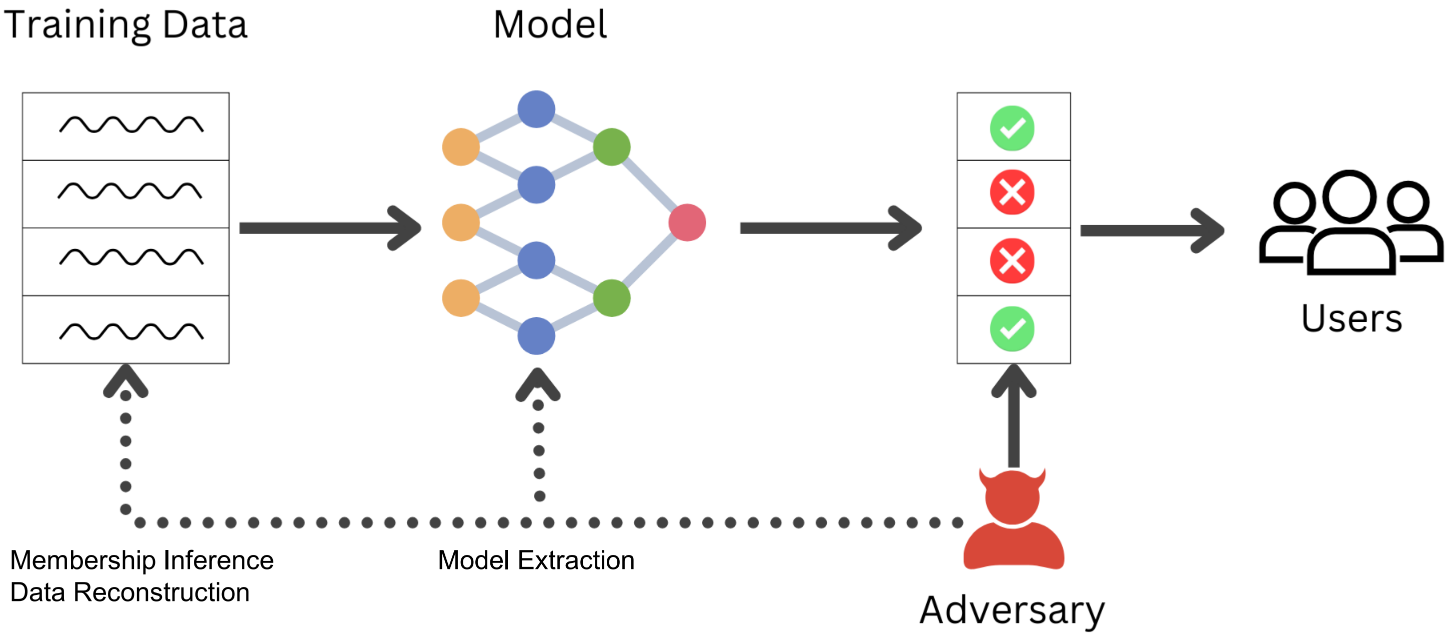
Crucial for public trust in AI/ML systems and promoting responsible deployment

MODEL OWNER VIEW



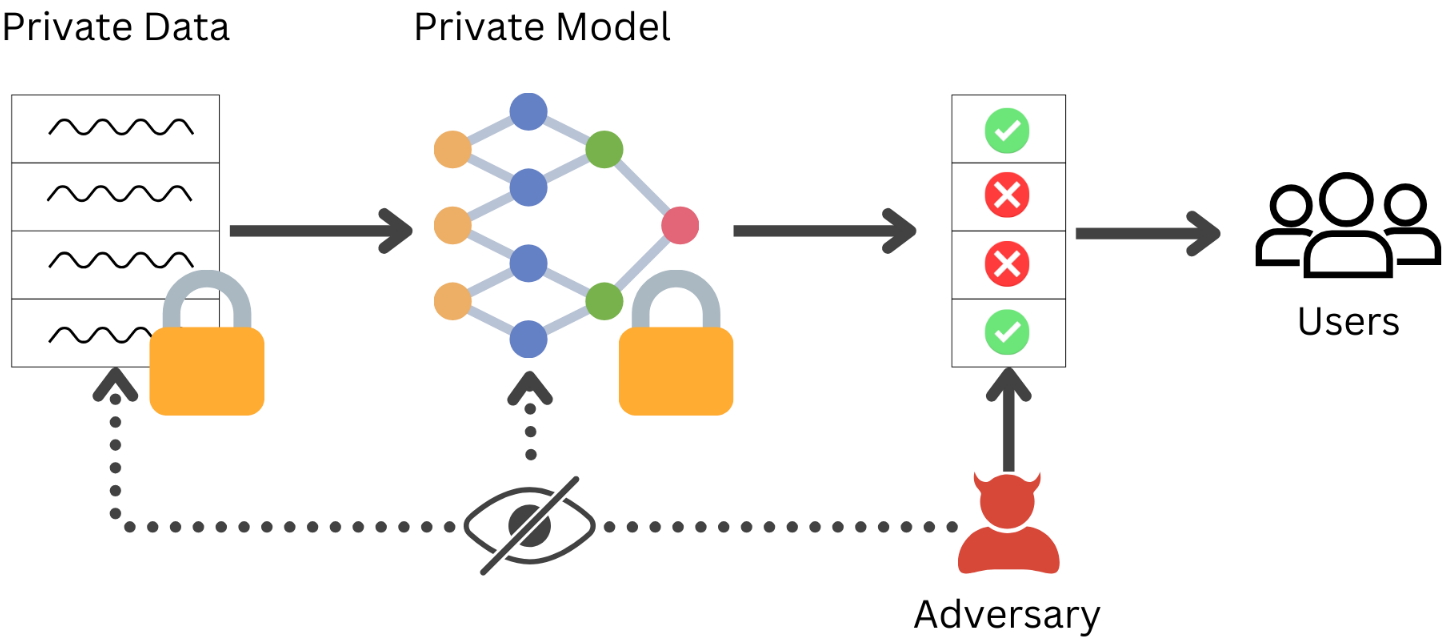
GOAL: Provide a model API for users

MODEL OWNER VIEW

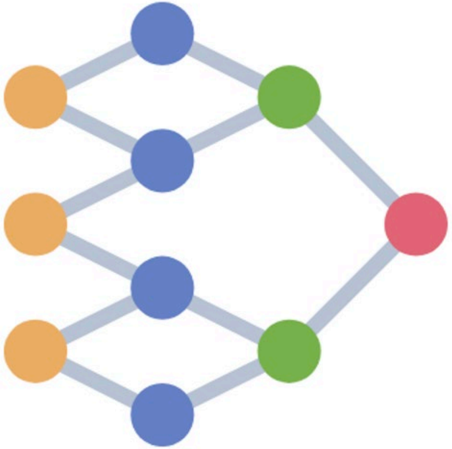


Challenges...

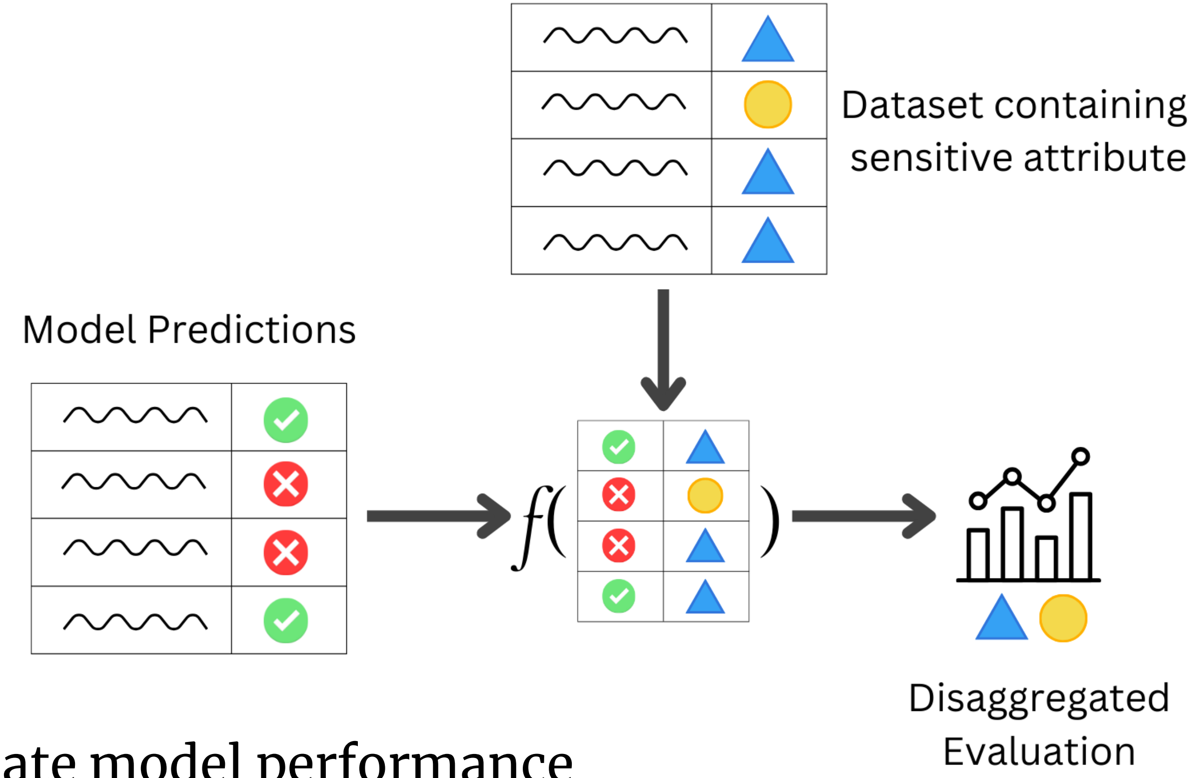
MODEL OWNER VIEW



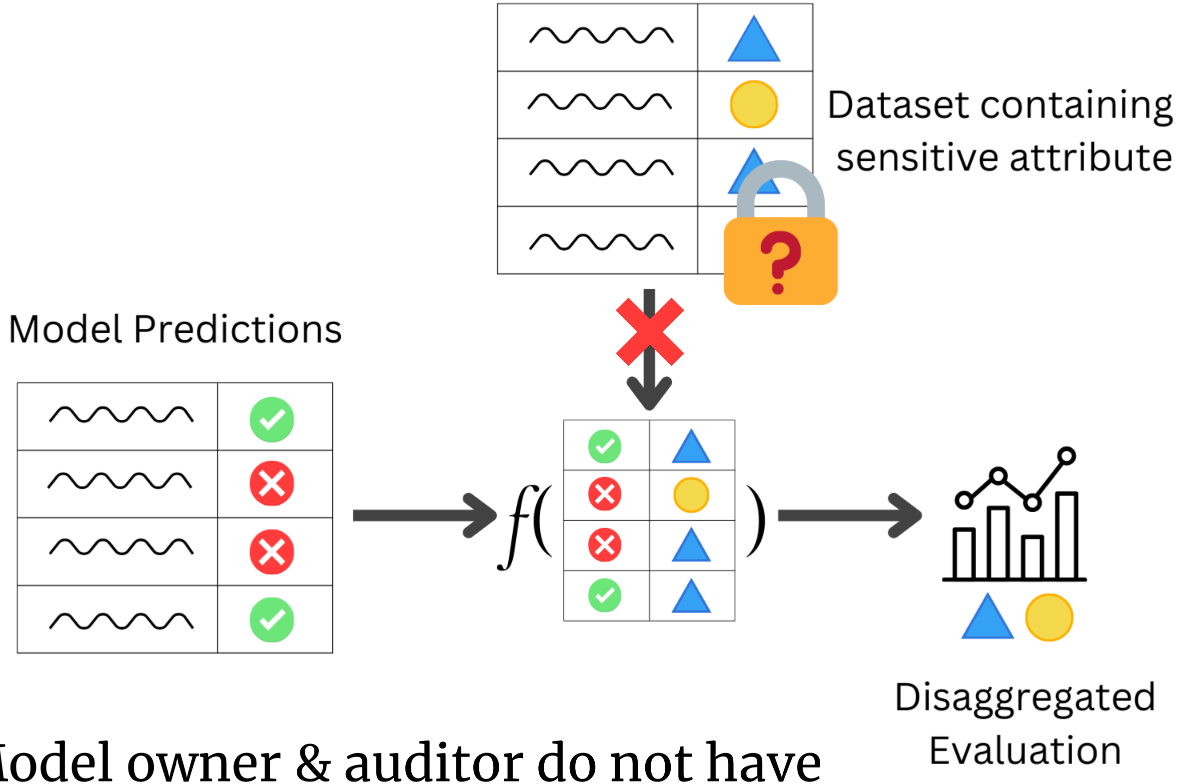
Mitigate privacy risks (eg DP, DP-SGD, distillation, regularization etc.)



How does the model perform across disaggregated subpopulations?



GOAL: Evaluate model performance conditional on a sensitive attribute



Challenge: Model owner & auditor do not have / cannot access sensitive attribute dataset

How can we protect privacy of the sensitive attribute dataset while preserving metric fidelity?

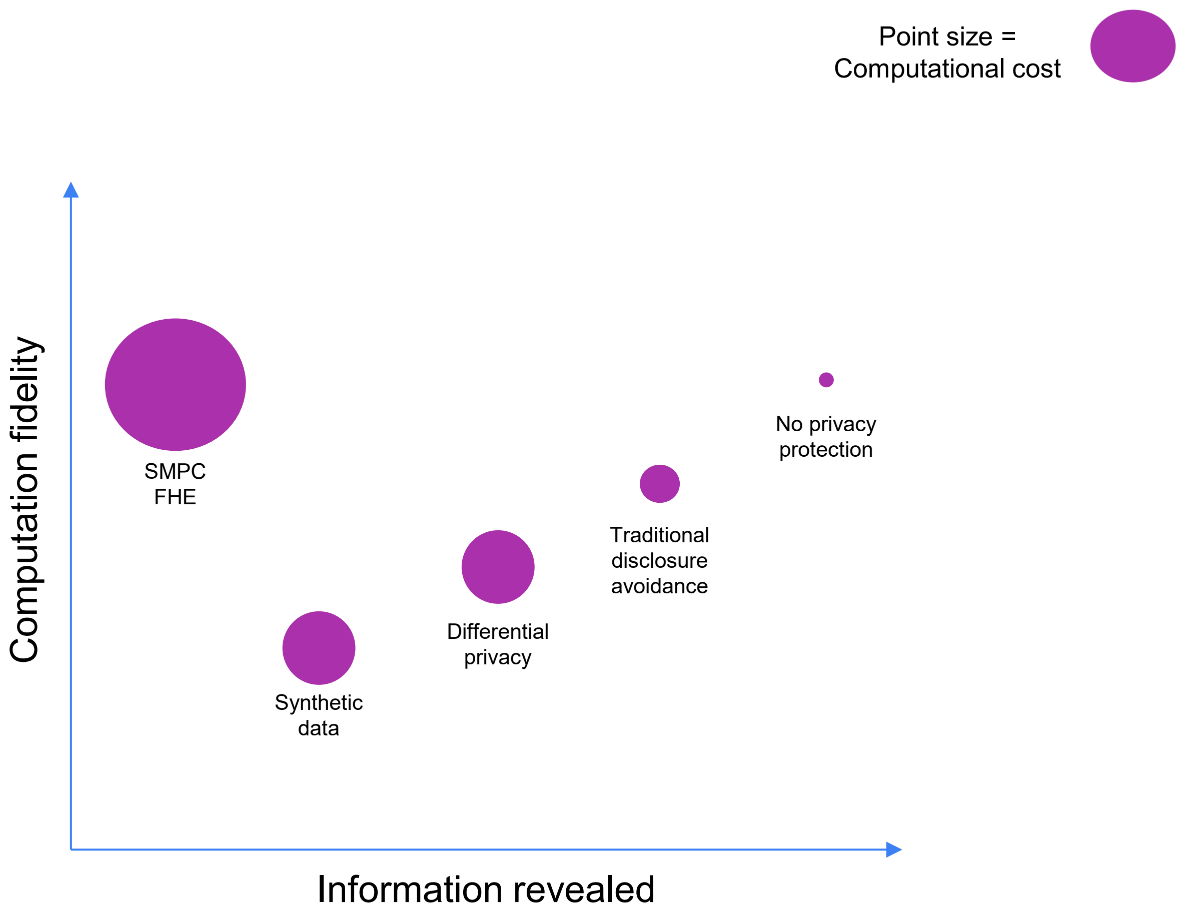
MODEL AUDITING WITH PETs

Add noise: ***differential privacy***, synthetic data generation

Encrypt: ***secure multi-party computation***, fully homomorphic encryption, zero knowledge proofs, secure enclaves

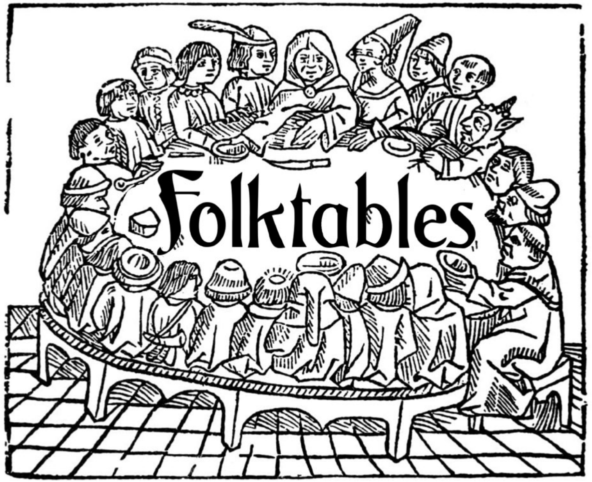
PETs TRADEOFFS

Techniques trade off between **fidelity**, **privacy**, and **computation cost**



EXPERIMENTAL SETUP

- Model Owner: Fit a (toy) logistic regression model on folktables ACS Employment task
- Sensitive Attribute Owner: Demographic features w/ common UID
- Auditor: Evaluate PETs in combination with common fairness metrics using fairlearn



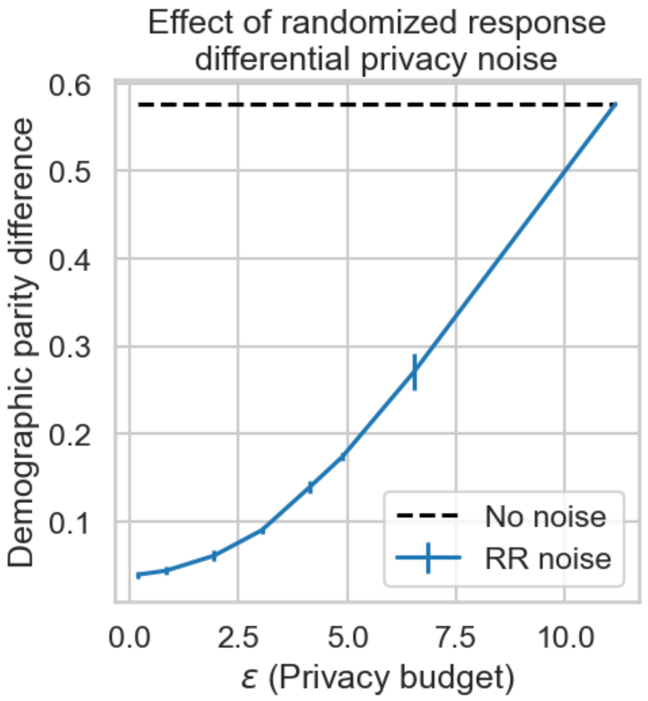
Fairlearn

OpenDP

- **Advantage:** Privacy guarantees

- **Disadvantage:** uncertainty increases with the magnitude of noise added

(ongoing work exploring corrections!)

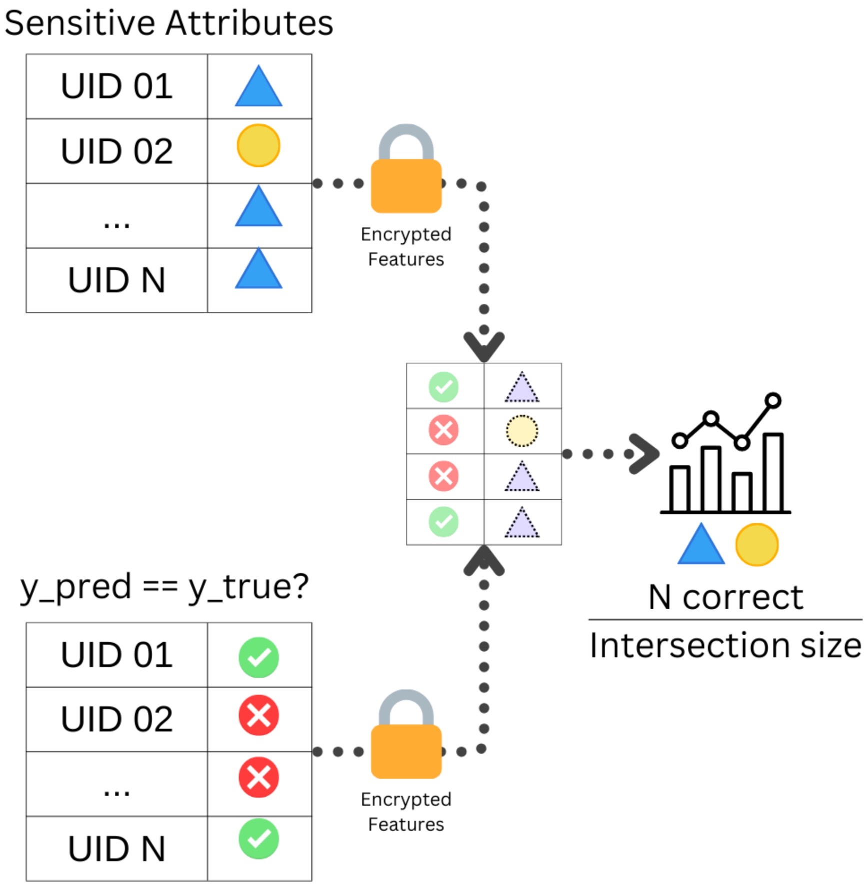


PRIVATE JOIN AND COMPUTE

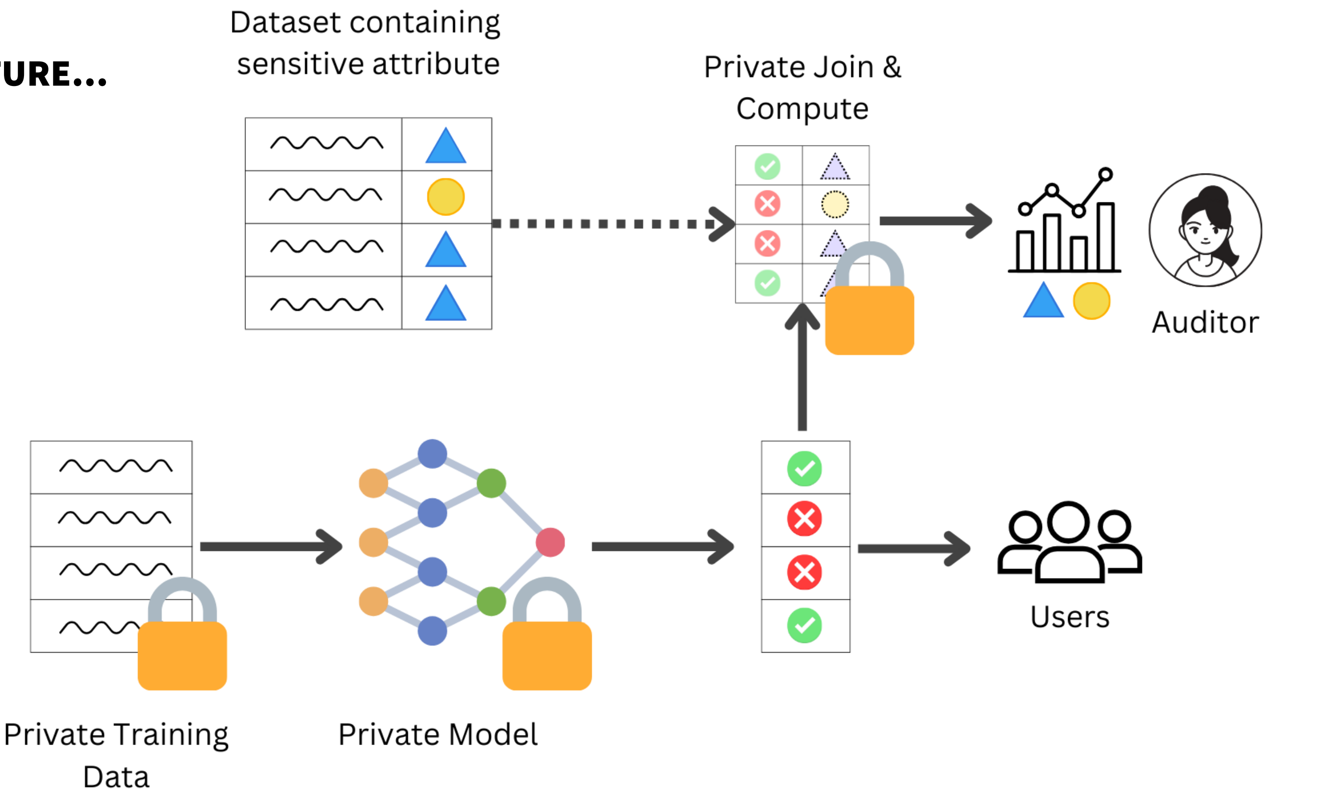
Both parties encrypt common identifier, join & sum encrypted attributes

- **Advantage:** Exact calculation
- **Disadvantage:** computational cost, privacy attacks

(ongoing work exploring mitigations!)



THE FULL PICTURE...



No single silver-bullet PET, complex tradeoffs between privacy, utility, fairness & compute

We’d love to hear from you!

—

inquiries@xd.gov

Mike Walton michael.w.walton@census.gov

Backup

DIFFERENTIAL PRIVACY + FAIRNESS METRICS

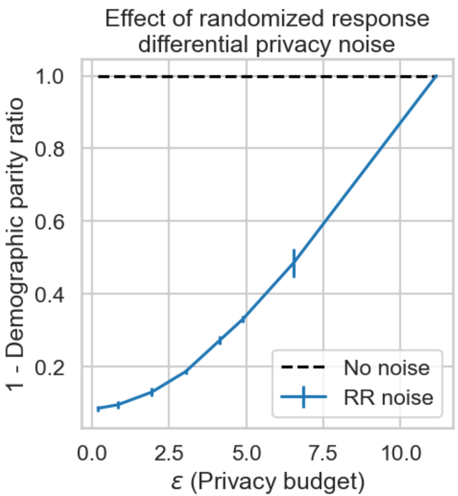
Evaluate fairness metrics with no noise

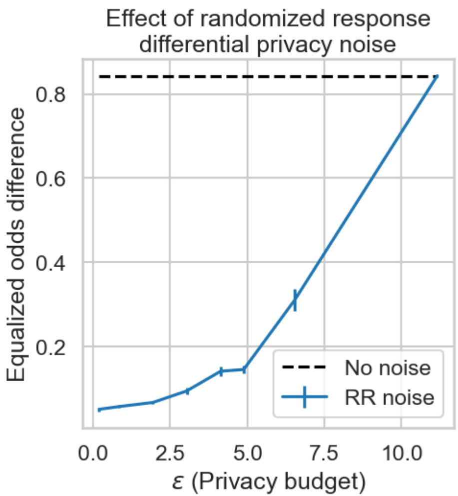
```python
def get_fairness_metrics(y_true, y_pred, sensitive_features):
    dpd = demographic_parity_difference(y_true, y_pred, sensitive_features=sensitive_features)
    dpr = demographic_parity_ratio(y_true, y_pred, sensitive_features=sensitive_features)
    eod = equalized_odds_difference(y_true, y_pred, sensitive_features=sensitive_features)
    eor = equalized_odds_ratio(y_true, y_pred, sensitive_features=sensitive_features)

    return np.array([dpd, dpr, eod, eor])
```

Evaluate fairness metrics with randomized response noise

```python
def run_trials(num_trials, p, y_true, y_pred, group_vals, num_metrics=4):
    results = np.zeros((num_trials, num_metrics))

    categories = np.unique(group_vals)
    rr_meas = make_randomized_response(categories, prob=p)
    v_func = np.vectorize(rr_meas)

    for i in range(num_trials):
        noisy_groups = v_func(group_vals)
        results[i, :] = get_fairness_metrics(y_true, y_pred, noisy_groups)

    return results
```

DIFFERENTIAL PRIVACY + FAIRNESS METRICS





The demographic parity ratio is defined as the ratio between the smallest and the largest group-level selection rate, $E[h(X)|A=a]$, across all values a of the sensitive feature(s). The demographic parity ratio of 1 means that all groups have the same selection rate.

$\mathbb{E}[h(X) \mid A=a, Y=y] = \mathbb{E}[h(X) \mid Y=y] \quad \forall a, y$. Equalized odds requires that the true positive rate, $\mathbb{P}(h(X)=1|Y=1)$, and the false positive rate, $\mathbb{P}(h(X)=1|Y=0)$, be equal across groups.

https://fairlearn.org/v0.10/api_reference/index.html

PRIVATE JOIN AND COMPUTE PROTOCOL, MORE FORMALLY

Private Intersection Sum with Cardinality

Inputs:

P_1 : Set $V = \{v_i\}_{i=1}^{m_1}$ P_2 : Set of pairs $W = \{(w_i, t_i)\}_{i=1}^{m_2}$

Outputs:

$\mathsf{P}_1 : C = |\{i : w_i \in V\}|$ $\mathsf{P}_2 : C = |\{i : w_i \in V\}|, S = \sum_{i:w_i \in V} t_i$

Figure 1: F_{PIS-C} : The Private Intersection-Sum with Cardinality functionality.

For details of the protocol, see Ion et al., *On Deploying Secure Computing: Private Intersection-Sum-With-Cardinality*, *https*://eprint.iacr.org/2019/723.pdf