Fairness Impact of Privacy

Milestone #2 Presentation

Blake Bullwinkel, Scarlett Gong, Kristen Grabarz, Lily Ke
October 28, 2021

Problem statement

Original Data	
Age	State
23	NY
47	NE
35	NY
29	CT
...	...
52	CT

Average Age: 44
State: 0.8% NE

+ Noise

Synthesis

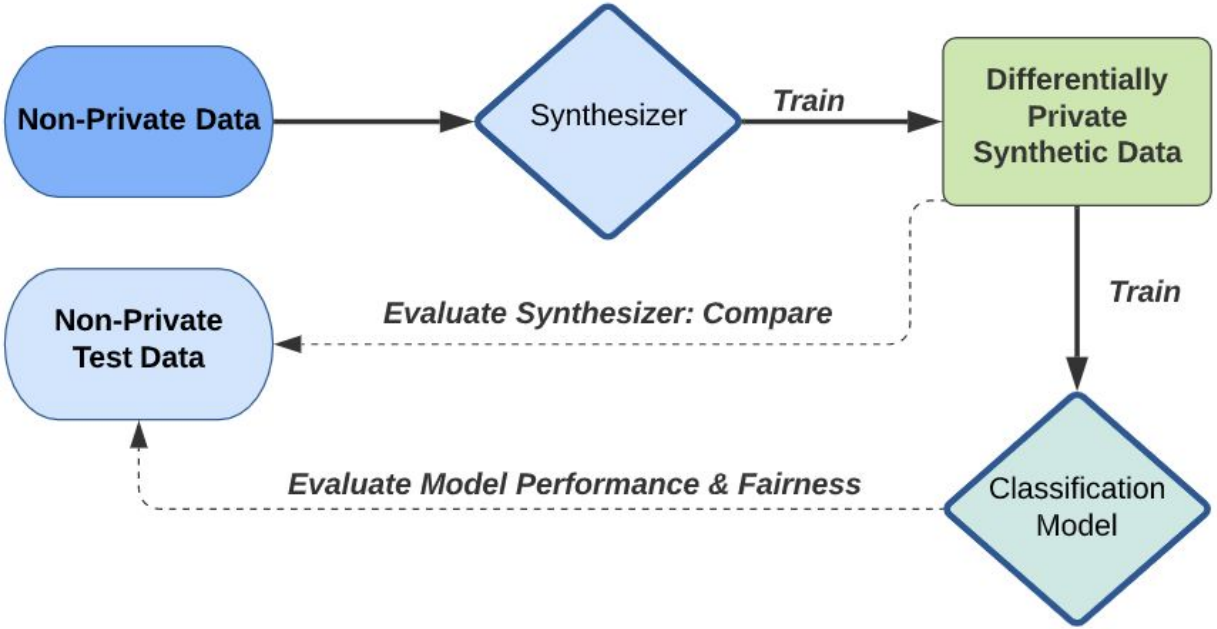
Private Data	
Age	State
24	NY
45	NY
33	NY
31	CT
...	...
51	CT

Average Age: 45
State: 0% NE

Differential privacy protects sensitive information by adding noise to data. However, it can have a disparate impact on model accuracy.

Our goal is to understand how changing ε (privacy loss) across various differentially private synthesizers affects our ability to achieve “fair” outcomes.

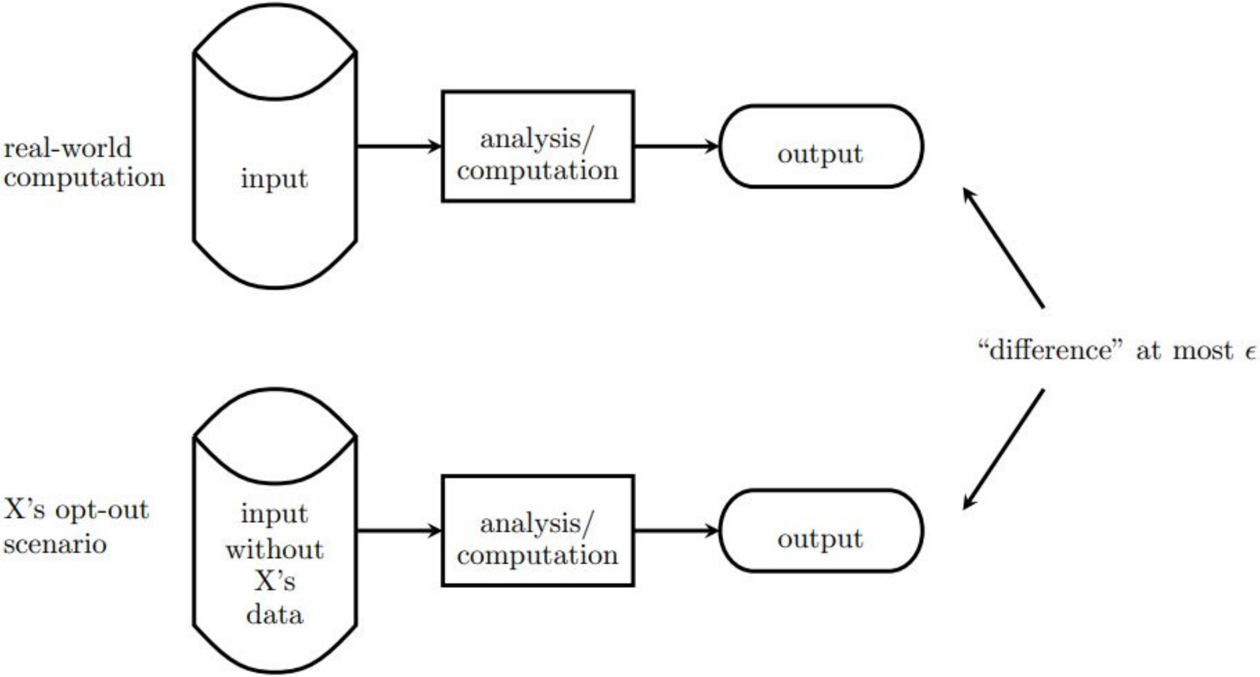
Synthesizers produce differentially private data



- **Synthesizers** are trained on original non-private dataset
- **Models** are then trained on the resulting differentially private synthetic data

Privacy parameter (ε)

Smaller values of epsilon indicate that more privacy is preserved

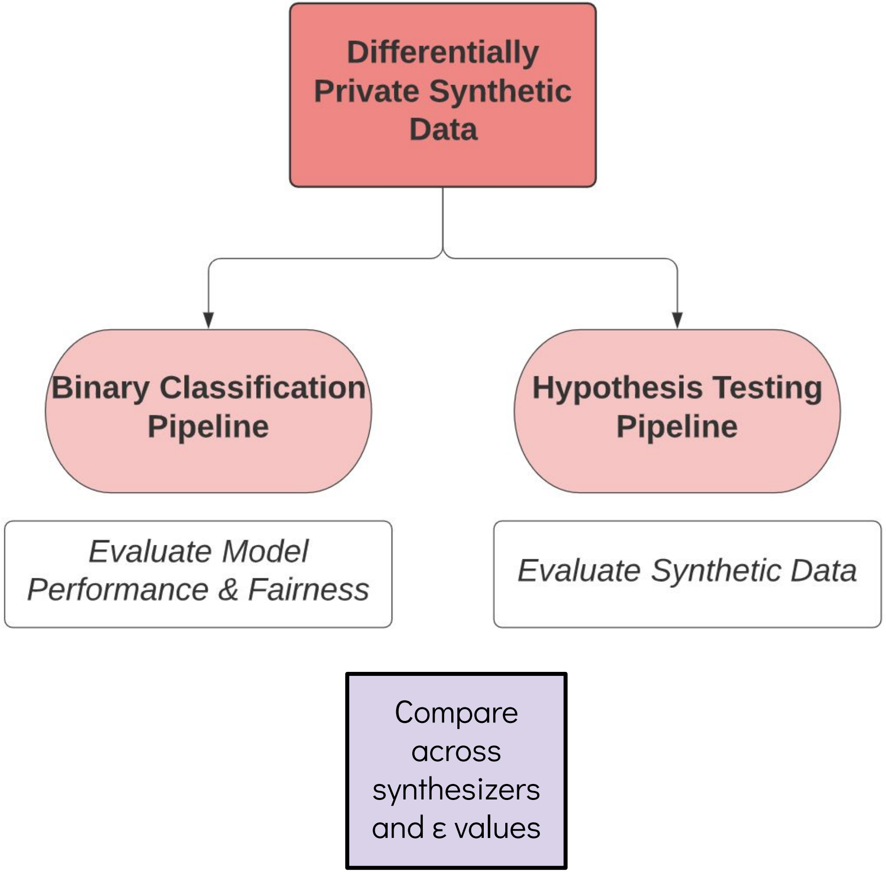


"Differential privacy: A primer for a non-technical audience." Vanderbilt Journal of Entertainment & Technology Law 21, no. 1 (2018): 209-275.

Our approach

Using popular datasets in the fairness literature, we will:

- **Generate** synthetic datasets using synthesizers and 8 ε values.
- **Perform** tasks including binary classification and hypothesis testing on the differentially private synthetic data
- **Measure and compare** fairness outcomes across these variants to understand the tradeoff between privacy and fairness.



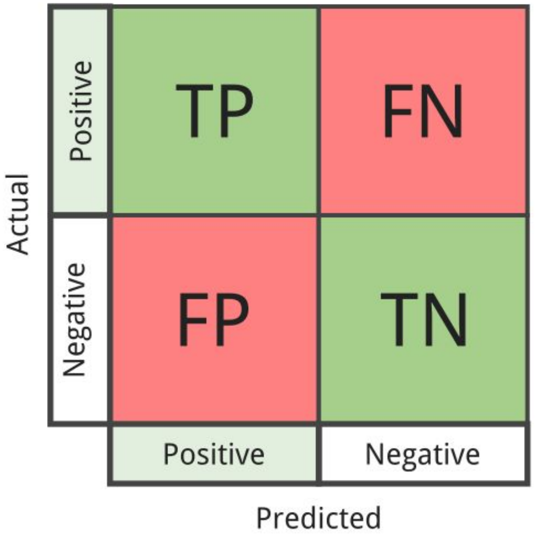
Metrics & Synthesizers

Fairness metrics

Core Fairness Metrics:

- **Binary Classification:** Understand model performance
  - True positive, False positive rate
  - Equalized Odds Distance:

    δ_y = Pr(ŷ=1|A=0,Y=y) - Pr(ŷ=1|A=1,Y=y), y ∈ {0,1}



- **Hypothesis Testing:** Understand how synthetic data compares to non-private data
  - Method: Difference in proportions hypothesis testing
  - Target outcomes across protected / unprotected groups
  - Target outcomes across original versus synthetic data for protected and unprotected groups

Evaluating fairness using permutation tests. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (pp. 1467-1477).

Differentially private synthesizers

MWEM

- Earliest and simplest synthesizer (2012)
- Fewer computational resources

QUAIL

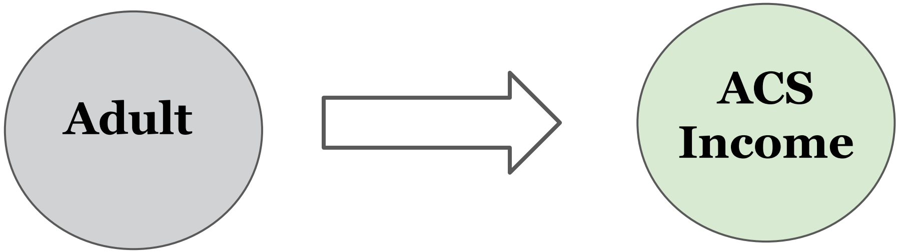
- Ensemble-based
- Helps reallocate the ε budget, for the ML task
- Recent (2020)

DPCTGAN

- More recent (2018-2019)
- GAN-based
- More computationally expensive

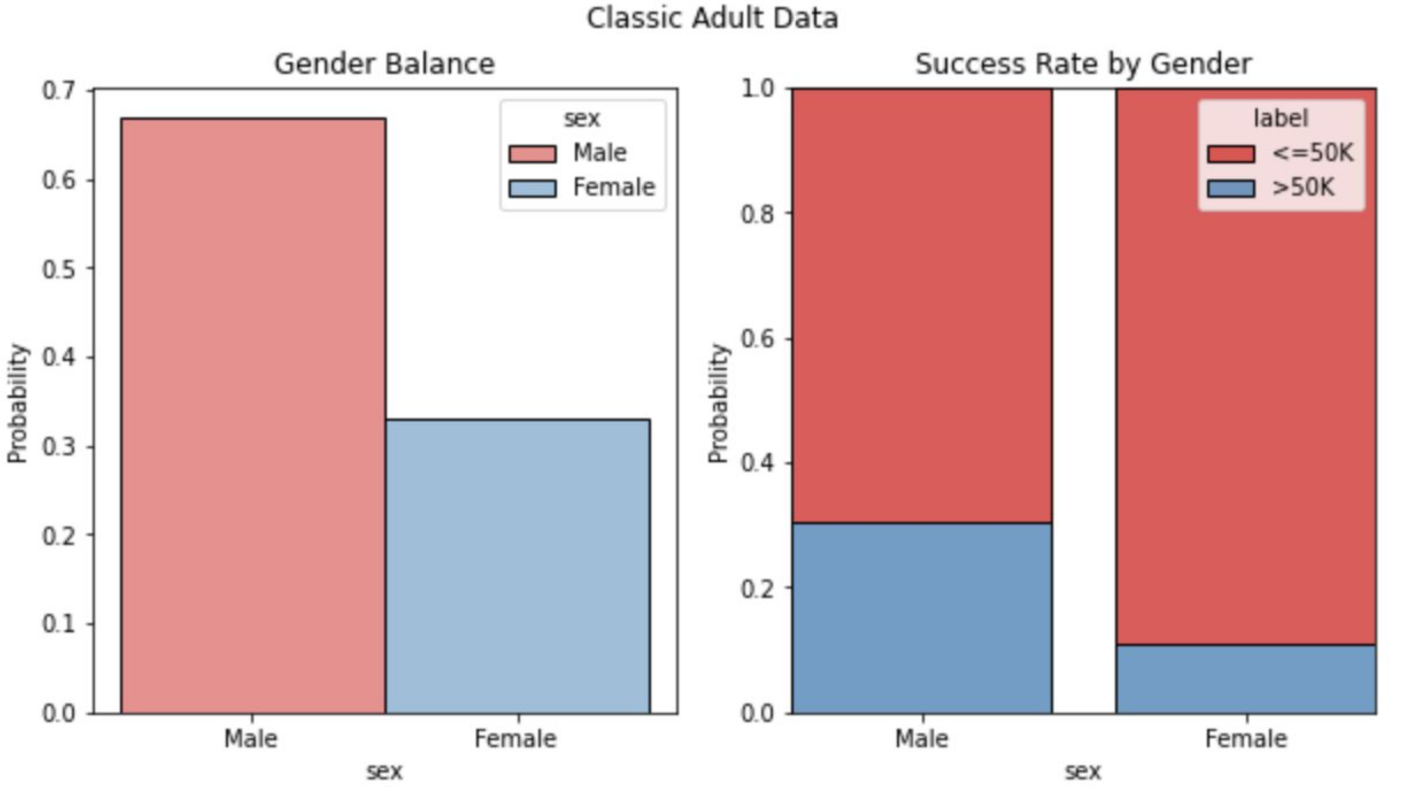
Ensuring our data is relevant to latest literature

- **August 2021:** Ding et al. publish paper noting limitations and idiosyncrasies with classic Adult dataset (taken from 1994 Census data) and recommend substitutes
  - Age
  - Documentation
  - Outdated feature encodings
  - Fairness criteria and trade-offs are sensitive to income threshold ($50k default)

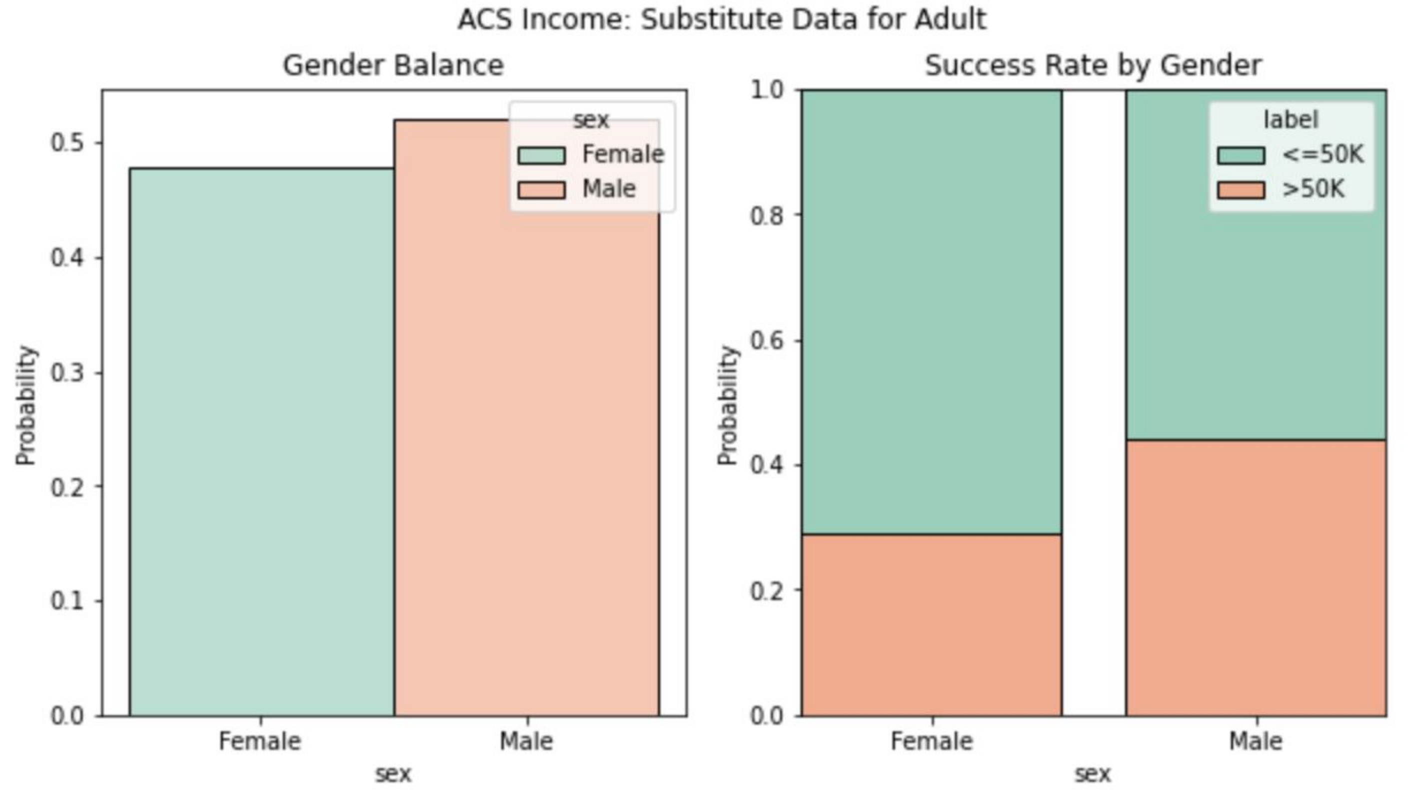


Ding, F., Hardt, M., Miller, J., & Schmidt, L. (2021). Retiring Adult: New Datasets for Fair Machine Learning. arXiv preprint arXiv:2108.04884.

Classic adult data distribution



ACS income data distribution: more fair



Key Takeaways so Far

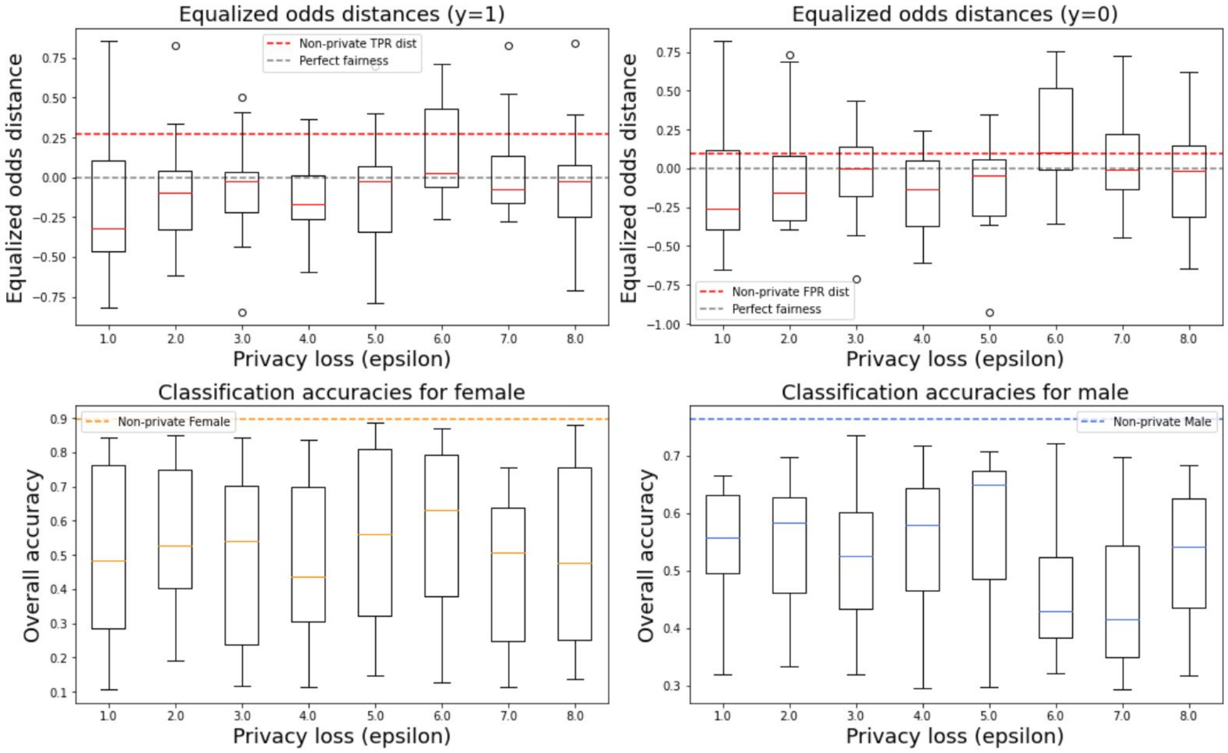
Based on experiments and comparisons to non-private baseline data

Baseline Performance: Non-Private Data

	Adult	ACS Income	COMPAS
Accuracy (Unprivileged Group)	0.897	0.712	0.638
Accuracy (Privileged Group)	0.764	0.715	0.608
Equalized Odds (y=1)	0.268	0.419	0.339
Equalized Odds (y=0)	0.095	0.175	0.171

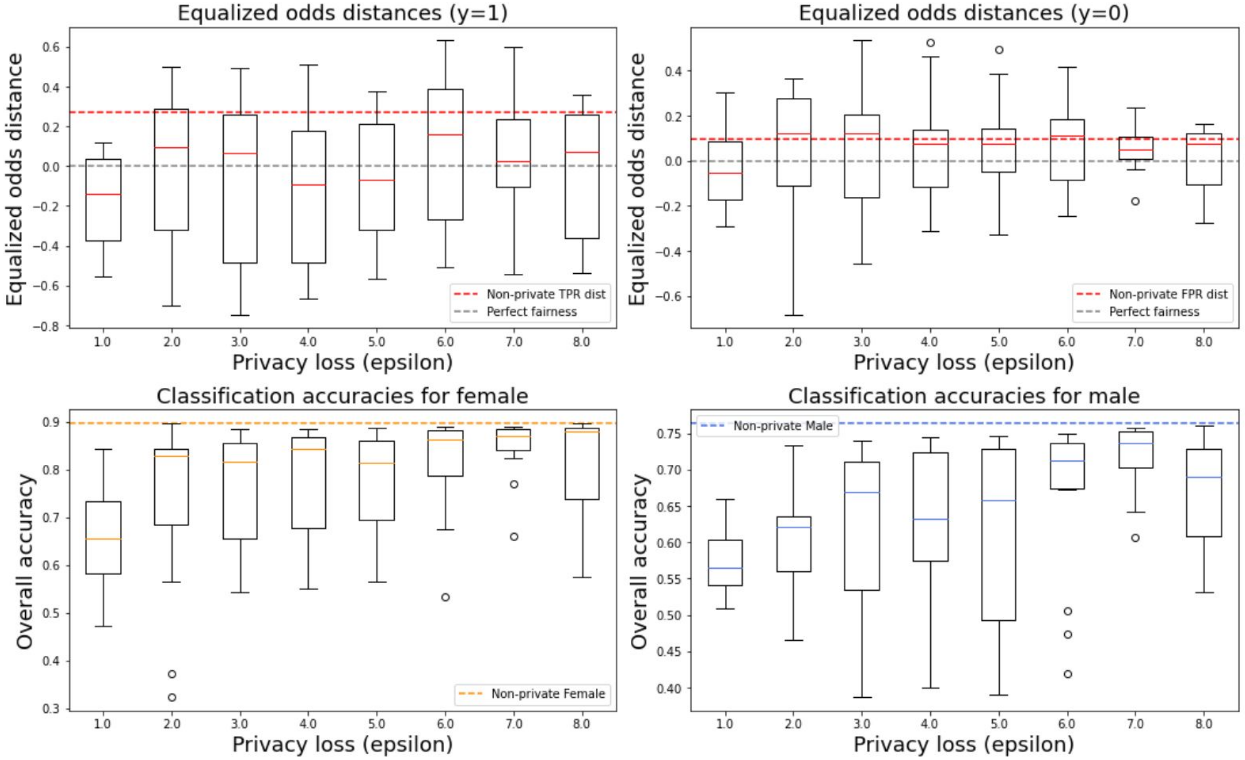
Key takeaway #1

MWEM creates synthetic data with more balanced classes across all values of epsilon considered, thereby improving fairness metrics but decreasing accuracy.



Key takeaway #2

Wrapping MWEM in QUAIL also creates more balanced classes, but increasing epsilon may have a disparate impact on the success rates of the classes, thereby illustrating a tradeoff between privacy and accuracy



Learned Lessons

opendp/
smartnoise-core

Different Smartnoise Synthesizers

Deepnote

Google colab

.npy

Real-time Collaboration

GPU resources

Store Results Locally

Lessons Learned and Upcoming Plans

Learned Lessons

- Data pre-processing
- Usages of different synthesizers
- Colab GPU on GAN models
- Store model results locally in .npy format for efficiency

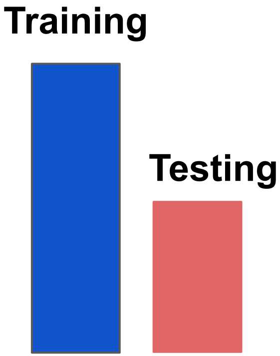
Issues encountered

Bug in MWEM model

Thank you, Lucas!

Get stuck in Python Panda dataframe loop for COMPAS MWEM

Long run-time in GAN-based method without GPU



Overfitting in COMPAS

Upcoming plans

Synthesizers

Train datasets on PateGAN and PrivBayes

Dataset

Synthesize German Credit dataset

Computation

Migrate from Colab & Deepnote to AWS Sagemaker

Models

Add Pre/Post-Processing techniques

Institute for
Applied Computational Science
HARVARD SCHOOL OF ENGINEERING AND APPLIED SCIENCES

Microsoft

Thank you!

I think we can end it here

Metrics for success

Fairness metrics

Baseline models and its results

Non-private data

- MWEM
- QUAIL
- CTGAN

Comparison with baseline model

Synthetic data with different synthesizers

Synthetic Adult data

- MWEM
- MWEL + QUAIL

Synthetic New Adult data

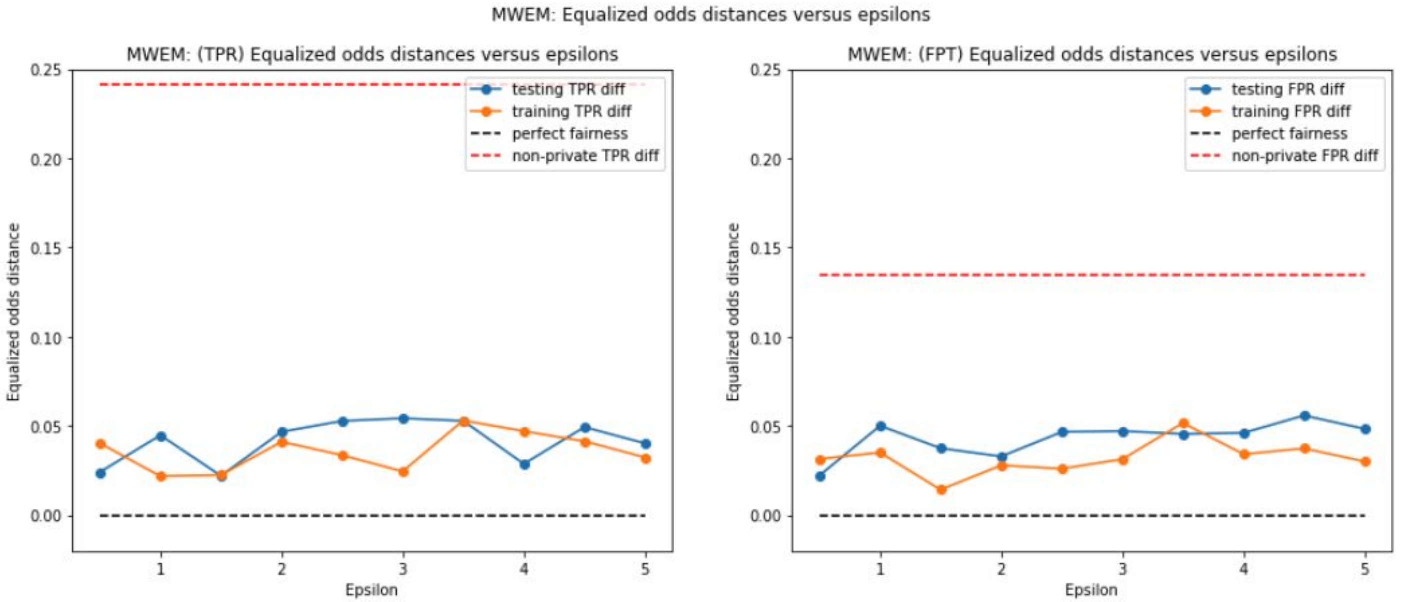
- MWEM
- MWEL + QUAIL

Synthetic COMPAS data

- MWEM
- MWEL + QUAIL
- DPCTGAN
- PATEGAN

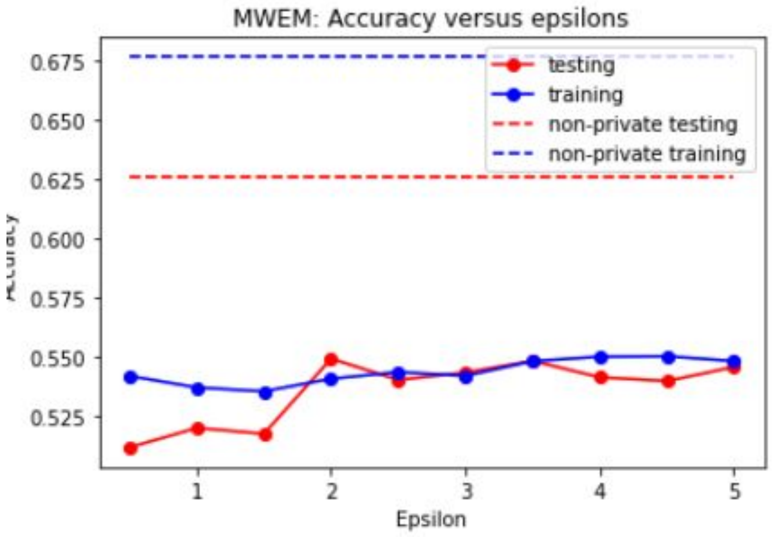
MWEM

- Equalized odds distance:
  - distances for both TPR and FPR are smaller compared to original data
  - no clear trend across different epsilon values



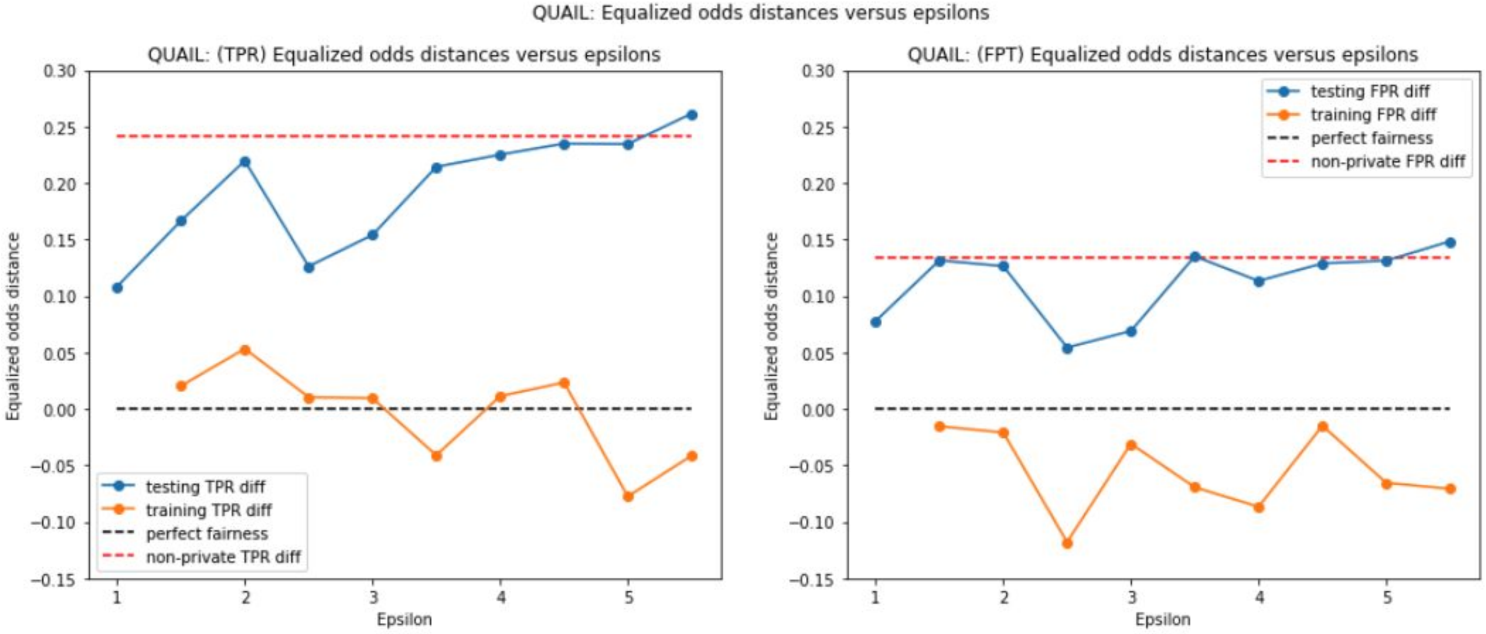
MWEM

- Accuracy:
  - accuracy is lower compared to original data
  - seems bigger the epsilon values higher the accuracy



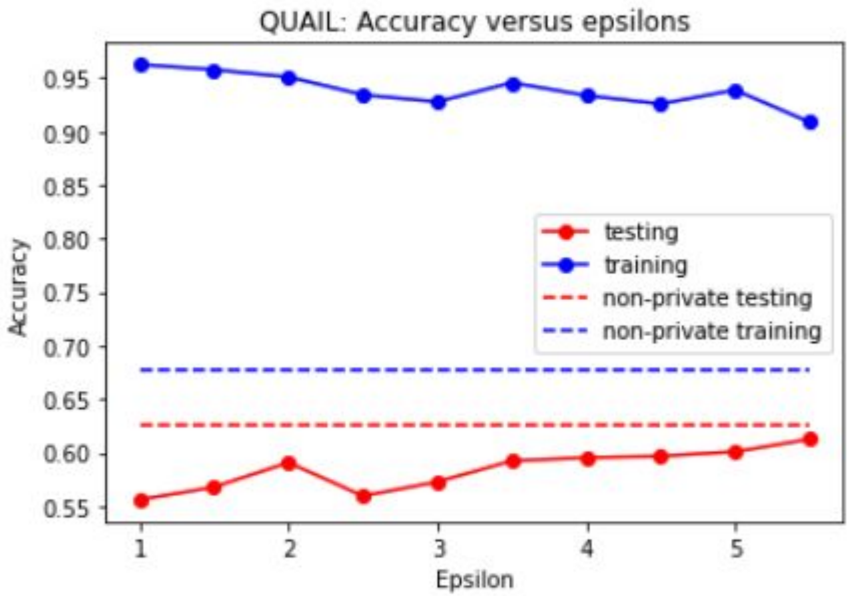
MWEM + QUAIL

- Equalized odds distance:
  - distances for both TPR and FPR are smaller(more fair) compared to original data
  - seems smaller the epsilon values smaller the distances



MWEM + QUAIL

- Accuracy:
  - accuracy is lower compared to original data,
  - seems bigger the epsilon values higher the accuracy
  - observe over-fitting (the accuracy for training is much higher than the accuracy for testing, (90-60)/60 = ~ 50%

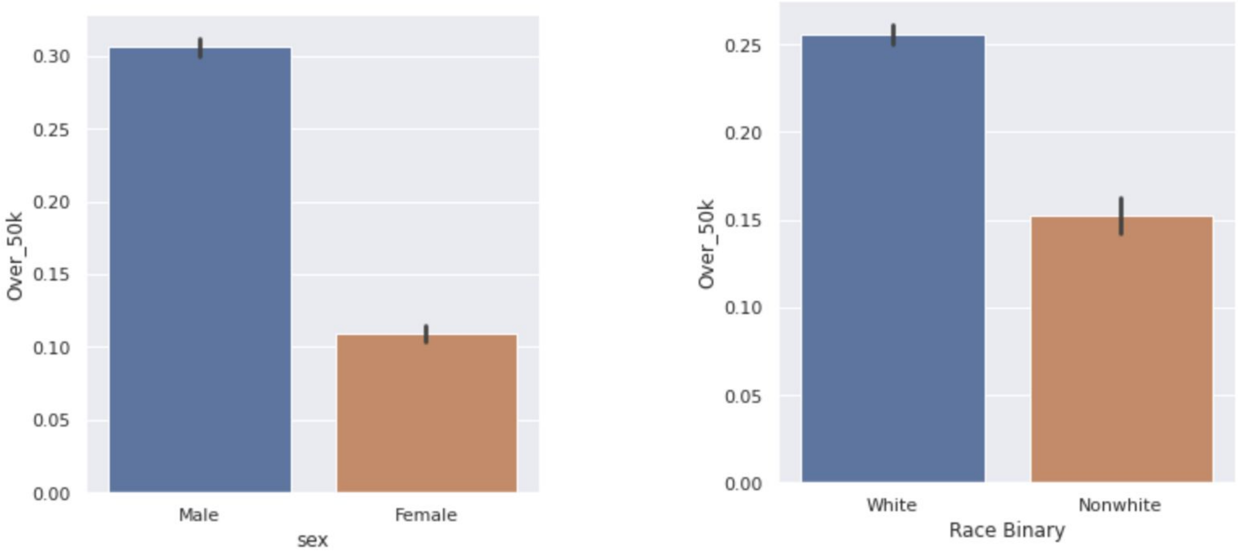


OLD

Lit review: Differentially private synthesizers

1. MWEM (2012): simple but effective with shorter runtime
2. PrivBayes (2014): developed by dataResponsibily
3. GAN-based (2018-2019): based on GAN architecture, privatized by DPSGD
    a. PATE-GAN
    b. DPGAN
    c. DP-CTPGAN
4. FFPDG (2021): “native fair” synthesizer developed by Amazon

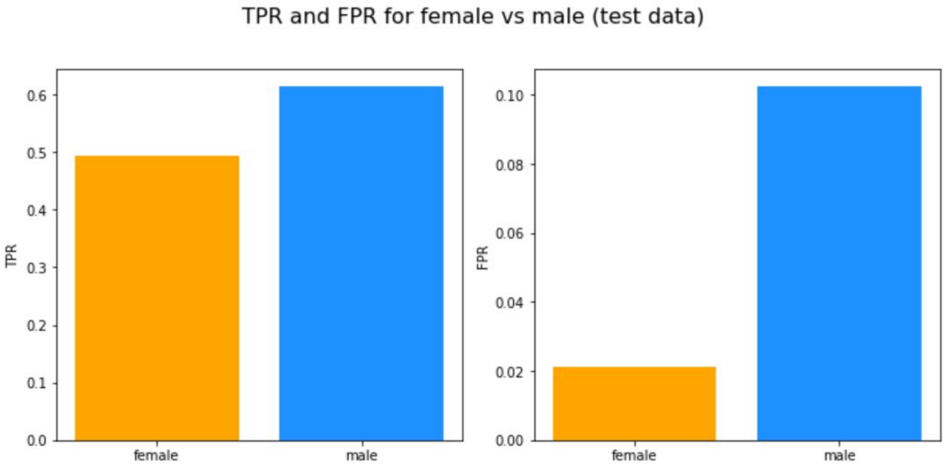
Adult data set: EDA



Both unprotected groups (Women and Nonwhite individuals) are less likely to make an income of at least $50k

Adult: Binary Classification Pipeline & Results

We confirmed that men is the privileged class and has higher TPR and FPR than women, both of which are associated with favorable outcomes in the Adult data set.

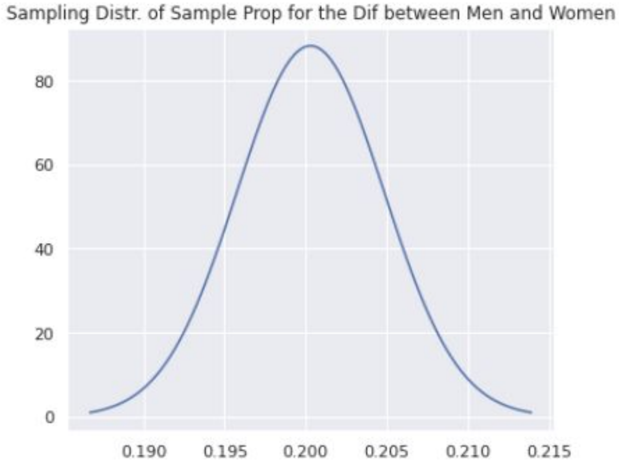


Test Set Fairness Metrics	**Gender**
Equalized Opportunity (TP rate difference)	0.120
Equalized Odds (FP rate difference)	0.081
Demographic Parity (FP+TP rate difference)	0.201

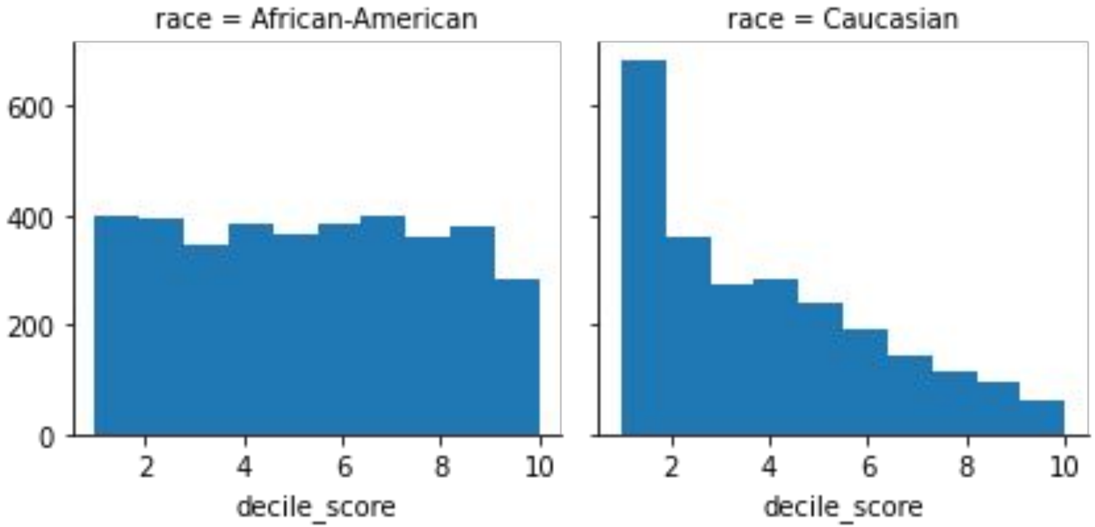
Adult: Hypothesis Test Pipeline & Results

- Compared rate of favorable outcome (>$50k) across protected versus unprotected group: Men versus Women

- **Reject the null hypothesis of no difference**:
  - Men significantly more likely than women to yield a positive outcome
  - Plan to expand comparison to original versus synthetic data



COMPAS dataset: EDA

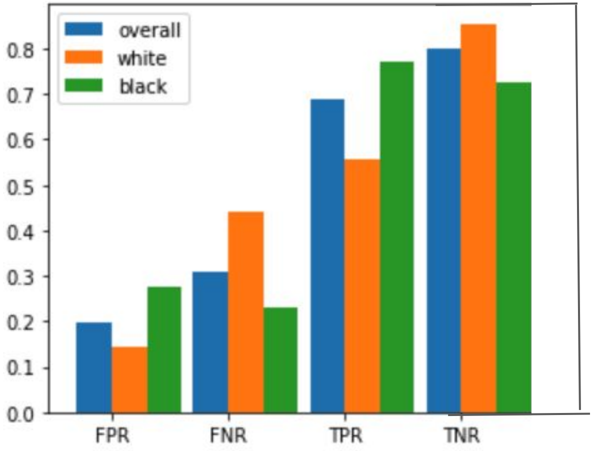


Histogram of decile_score provided by COMPAS tool

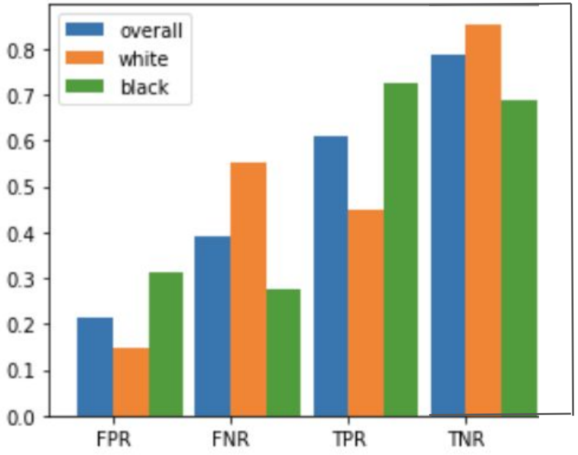
Plotting the decile scores produced by COMPAS tool as a prediction score, the distribution for white individuals is right-skewed

COMPAS: Binary Classification Pipeline & Results

This shows the classifications appeared to favor white defendants over black defendants by underpredicting recidivism for white and over predicting recidivism for black defendants.



Logistic Regression



Decision Tree

COMPAS: Hypothesis Test Pipeline & Results

Compared recidivism rate across protected versus unprotected group: African American versus Caucasian individuals.

Reject the null hypothesis of no difference:

- Mean of the African American predicted recidivism rate > the mean of the Caucasian predicted recidivism rate
- Mean of the African American predicted recidivism rate > the mean of the African American real recidivism rate