A Study on Utility, Privacy, and Fairness in GAN-generated Synthetic Data

Bharathvaj K M, Katyani S, Rudraksh K

CMPUT 622 - Under the guidance of Dr. Nidhi Hegde

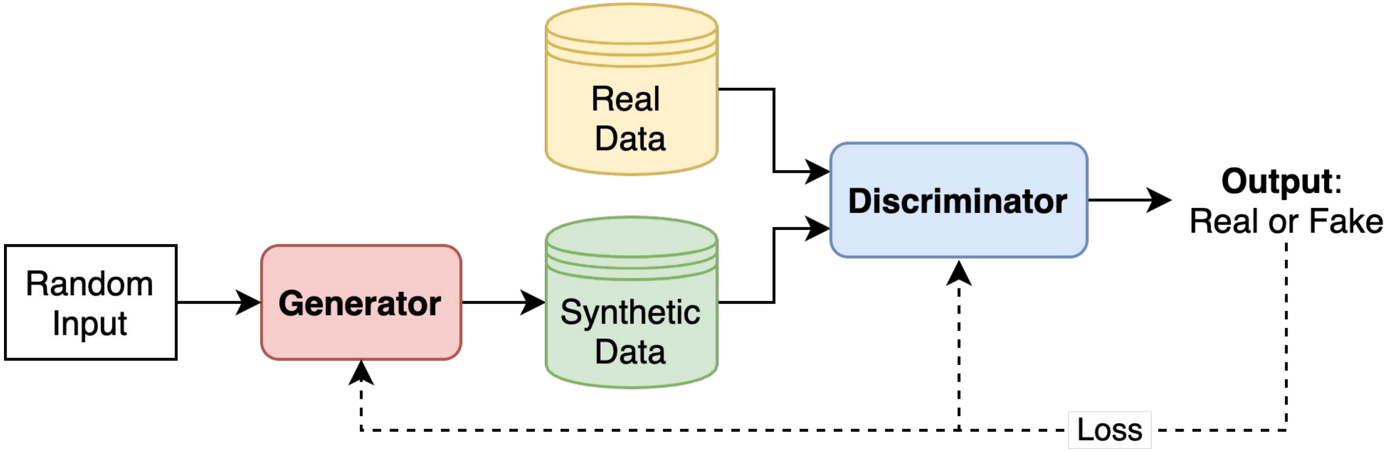
2021-11-24

Outline

- Terminology Review
- Motivation
- Related Work and Research Gap
- Data
- GAN architectures
- Methodology
- Preliminary Results
- Next Steps

Terminology Review

- **Synthetic data**
  - Data created artificially, having the same distribution as real data
- **GAN**
  - A framework for generating synthetic data
  - Adversarial model training



Terminology Review

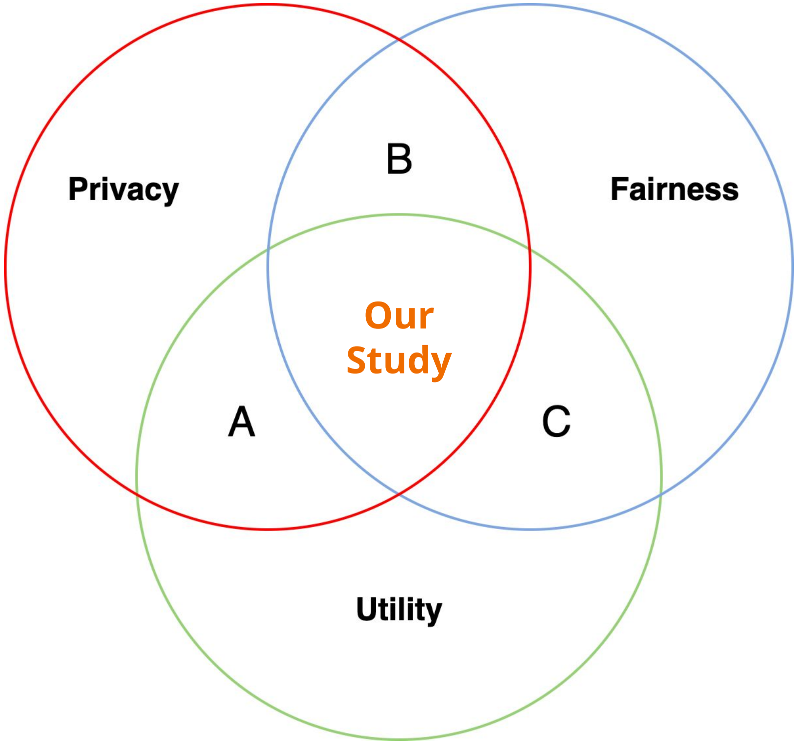
- **Privacy**
  - Protecting personally identifiable information
  - Don’t want synthetic data to exactly match real data
- **Fairness**
  - Unfairness - biased outcome
  - Don’t want synthetic data to reinforce any stereotypes
  - Protected attribute: something that you don’t want to discriminate on the basis of
- **Utility**
  - Performance of the ML model on the task, e.g. classification accuracy
  - Don’t want synthetic data distribution to be too different from real data

Motivating Example

- Sensitive datasets are used to train many machine learning models
  - E.g. Electronic health records
- Models can leak information about their training data - privacy risk
  - Through membership inference attacks (Shokri 2016) [1], for example
- Using a synthetic training dataset helps mitigate privacy risk
- While ideally also maintaining utility
  - I.e. on a task such as binary classification
- This has been studied.
- But what impact does any of this have on fairness?

Related Work and Research Gap

- There is plenty of previous work on
  - Privacy vs Utility (A)
    - Lin 2021 [2]
  - Privacy vs Fairness (B)
    - Gupta 2021 [3]
  - Fairness vs Utility (C)
    - Xu 2018 [4]
- We seek to study all three of these factors and their trade-offs.



Dataset considered

- For our study, we consider the **Census Income** dataset [5]
- Associated task is **binary classification**
    - Input: Green
    - Target: Red (>= $50k or < $50k)
- Protected attributes:
    - E.g. sex, race, age

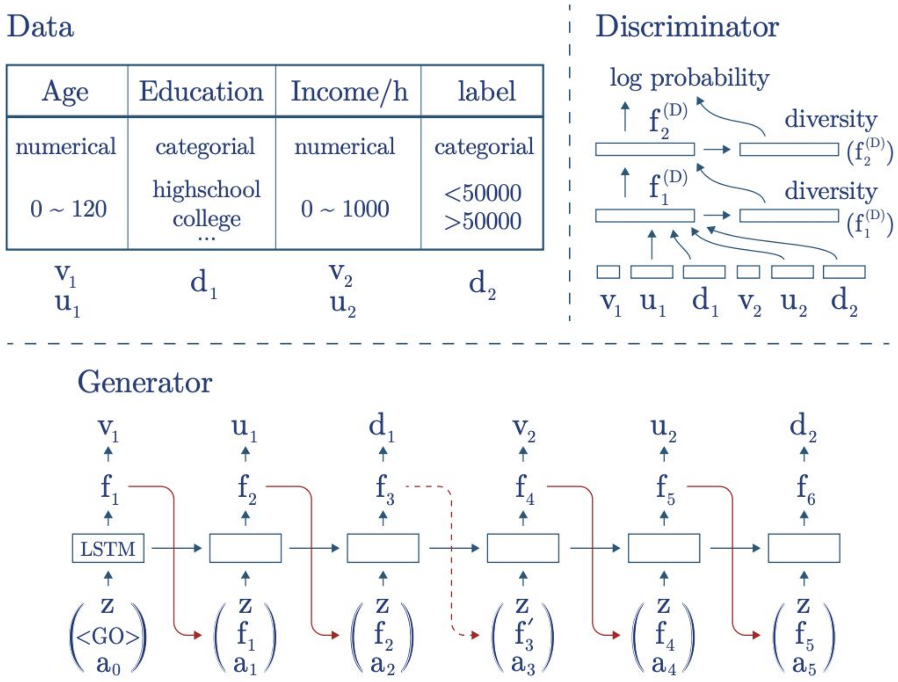
Input Features

```
#    Column
---  ------
0    age
1    workclass
2    education
3    education_num
4    marital_status
5    occupation
6    relationship
7    race
8    sex
9    capital_gain
10   capital_loss
11   hours_per_week
12   native_country
13   income
```

Target

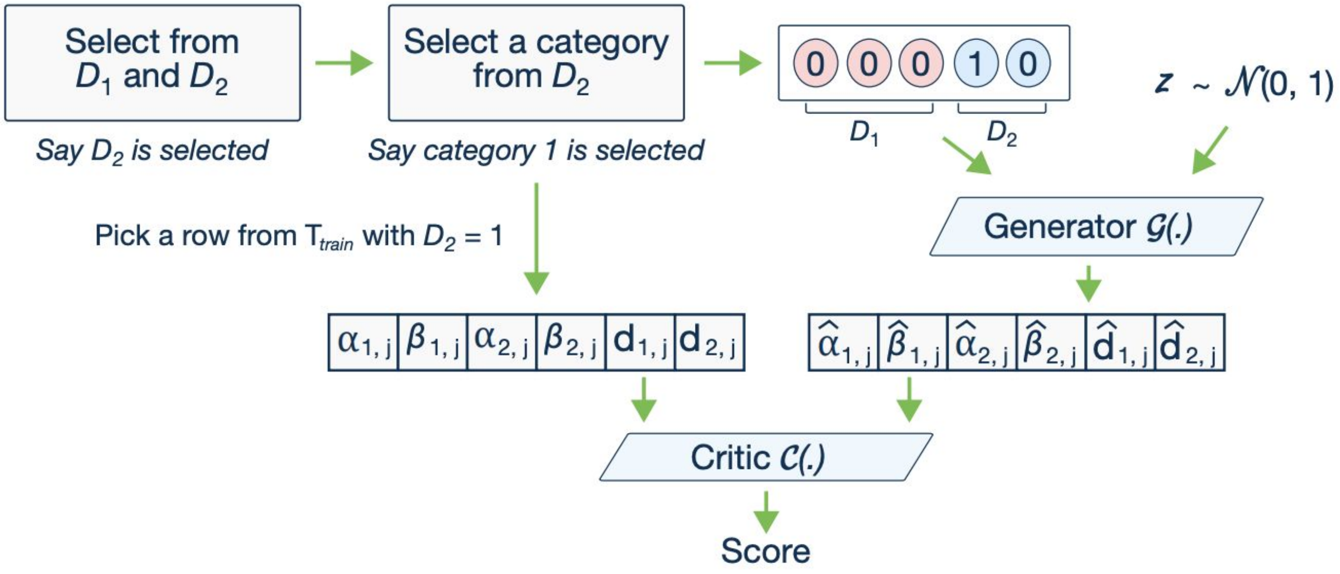
Tabular GAN

- **Key Idea**: Learn marginal distribution of each column by minimizing KL Divergence
- **Gen**: LSTM
  - Continuous: **v, u** -2 steps
  - Discrete: **d** - 1 step
  - Generates each in order
- **Disc**: Multi-layer perceptron
  - Concatenates features together for input.
  - Uses mini-batch discrimination vector.



Credits: Lei Xu and Kalyan Veeramachaneni. Synthesizing tabular data using generative adversarial networks. arXiv preprint arXiv:1811.11264, 2018. https://arxiv.org/pdf/1811.11264.pdf

CTGAN



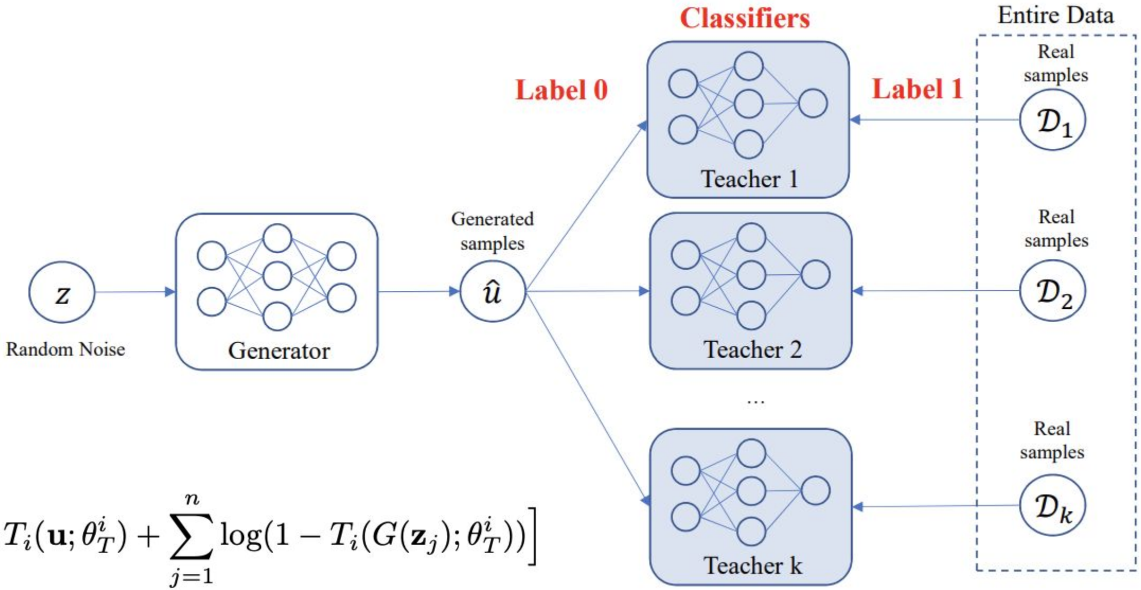
Three key elements

1. Conditional vector
2. Generator loss
3. Training-by-sampling

Credits: Lei Xu, Maria Skoularidou, Alfredo Cuesta-Infante, and Kalyan Veeramachaneni. Modeling tabular data using conditional GAN. arXiv preprint arXiv:1907.00503, 2019. https://arxiv.org/pdf/1907.00503.pdf

PATE-GAN

Training procedure for the **teacher-discriminator**

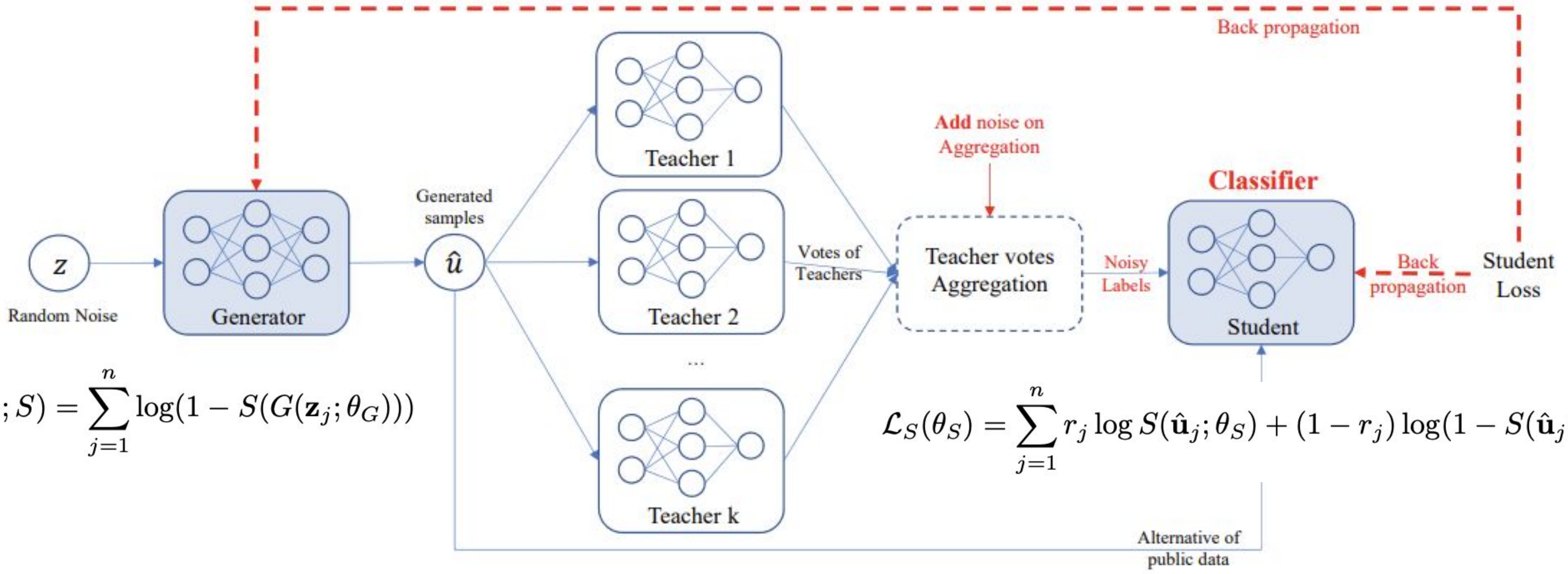


$$\mathcal{L}_T^i(\theta_T^i) = -\Big[\sum_{\mathbf{u}\in\mathcal{D}_i} \log T_i(\mathbf{u};\theta_T^i) + \sum_{j=1}^{n} \log(1 - T_i(G(\mathbf{z}_j);\theta_T^i))\Big]$$

Credits: James Jordon, Jinsung Yoon, and Mihaela Van Der Schaar. PATE-GAN: Generating synthetic data with differential privacy guarantees. In International Conference on Learning Representations, 2018. https://openreview.net/pdf?id=S1zk9iRqF7

PATE-GAN

Training procedure for the **student-discriminator** and the **generator**



$$\mathcal{L}_G(\theta_G; S) = \sum_{j=1}^{n} \log(1 - S(G(\mathbf{z}_j; \theta_G)))$$

$$\mathcal{L}_S(\theta_S) = \sum_{j=1}^{n} r_j \log S(\hat{\mathbf{u}}_j; \theta_S) + (1 - r_j) \log(1 - S(\hat{\mathbf{u}}_j; \theta_S))$$

Credits: James Jordon, Jinsung Yoon, and Mihaela Van Der Schaar. PATE-GAN: Generating synthetic data with differential privacy guarantees. In International Conference on Learning Representations, 2018. https://openreview.net/pdf?id=S1zk9iRqF7

Key Elements - GAN Models

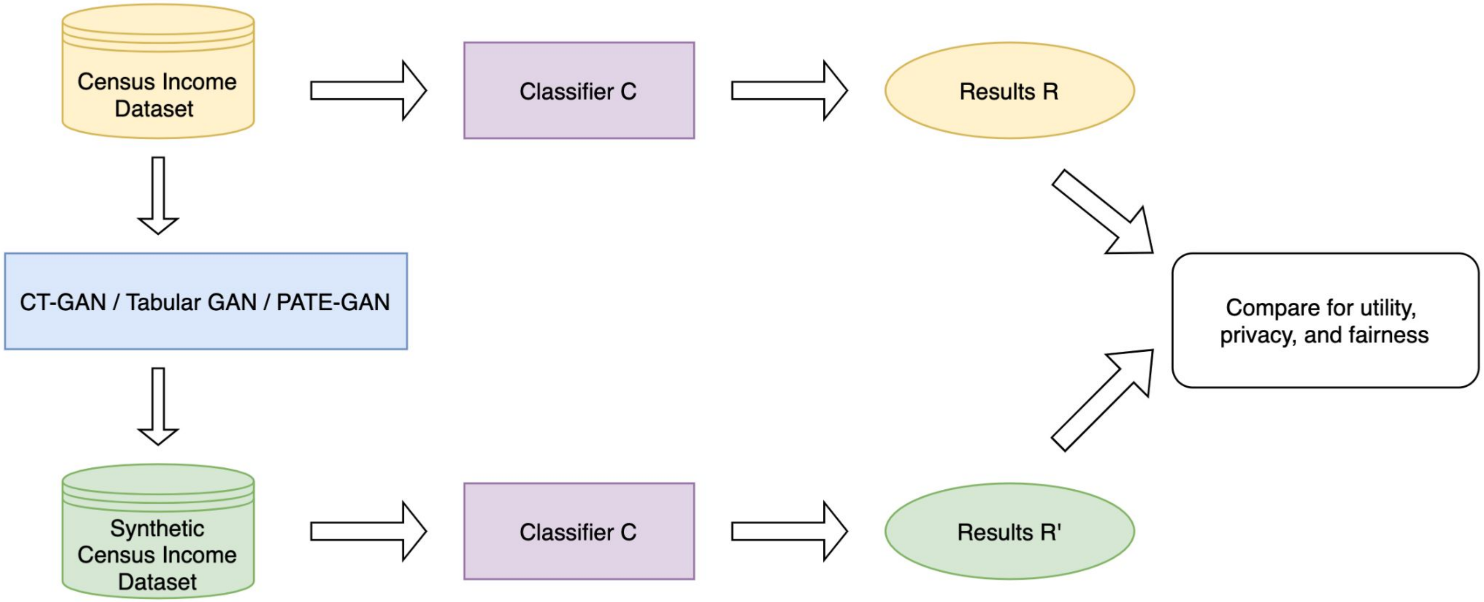
Normal GANs

- In TabularGAN, LSTM architecture is used in the generator, whereas CTGAN uses conditional generator
- In TabularGAN, Column order of original dataset matters
- TabularGAN minimises the KL divergence between real and synthetic columns

Differentially private GAN

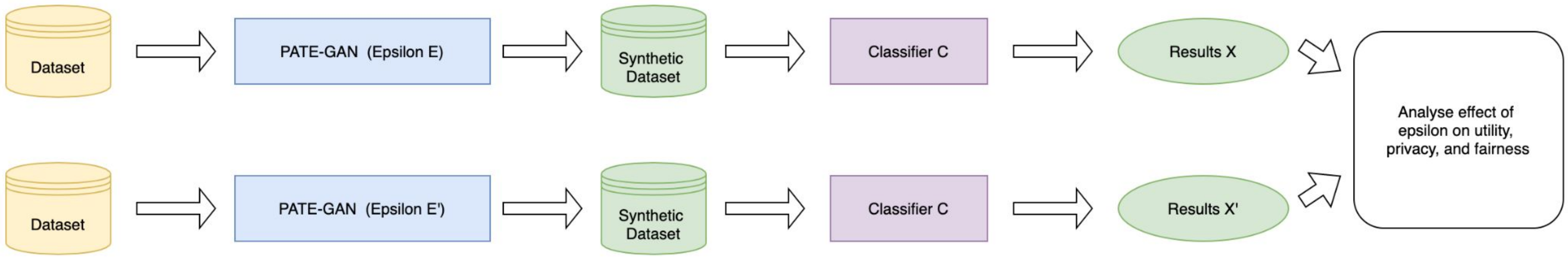
- Ensures differential privacy of the generator in GANs
- Tightly bounds the influence of any individual sample on the model
- Based on Private Aggregation of Teacher Ensembles (PATE) framework

Methodology for Study



Flow for studying utility, privacy and fairness of model trained on synthetic data.

Studying privacy budget effect on fairness and utility (Our Contribution)



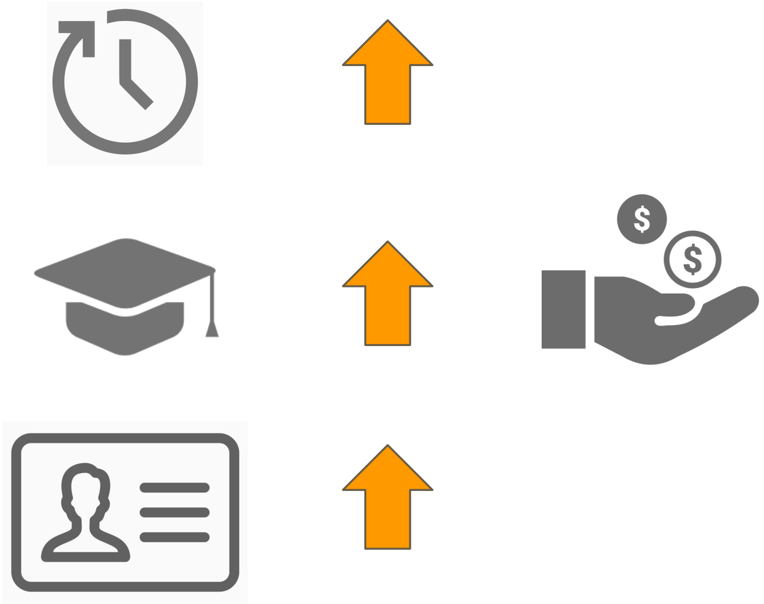
Flow for studying the effect of privacy budget on the model's fairness and utility.

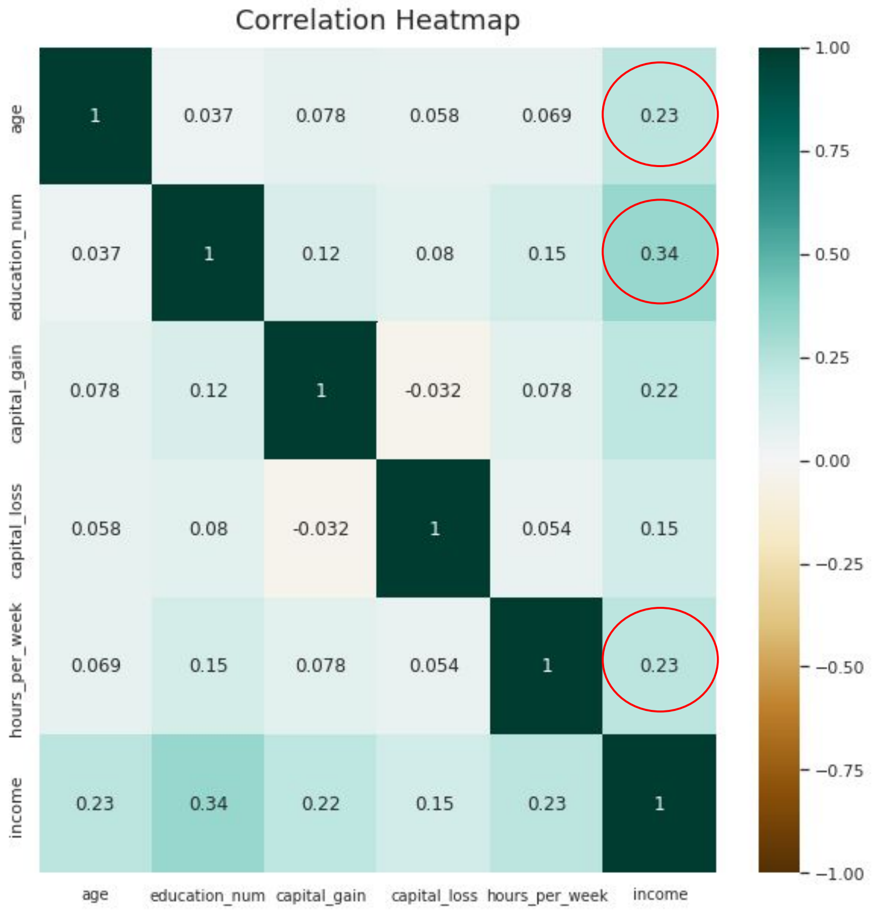
Preliminary Results

- Data Insights
- Preliminary Results
  - Utility
  - Privacy
  - Fairness

Data Insights

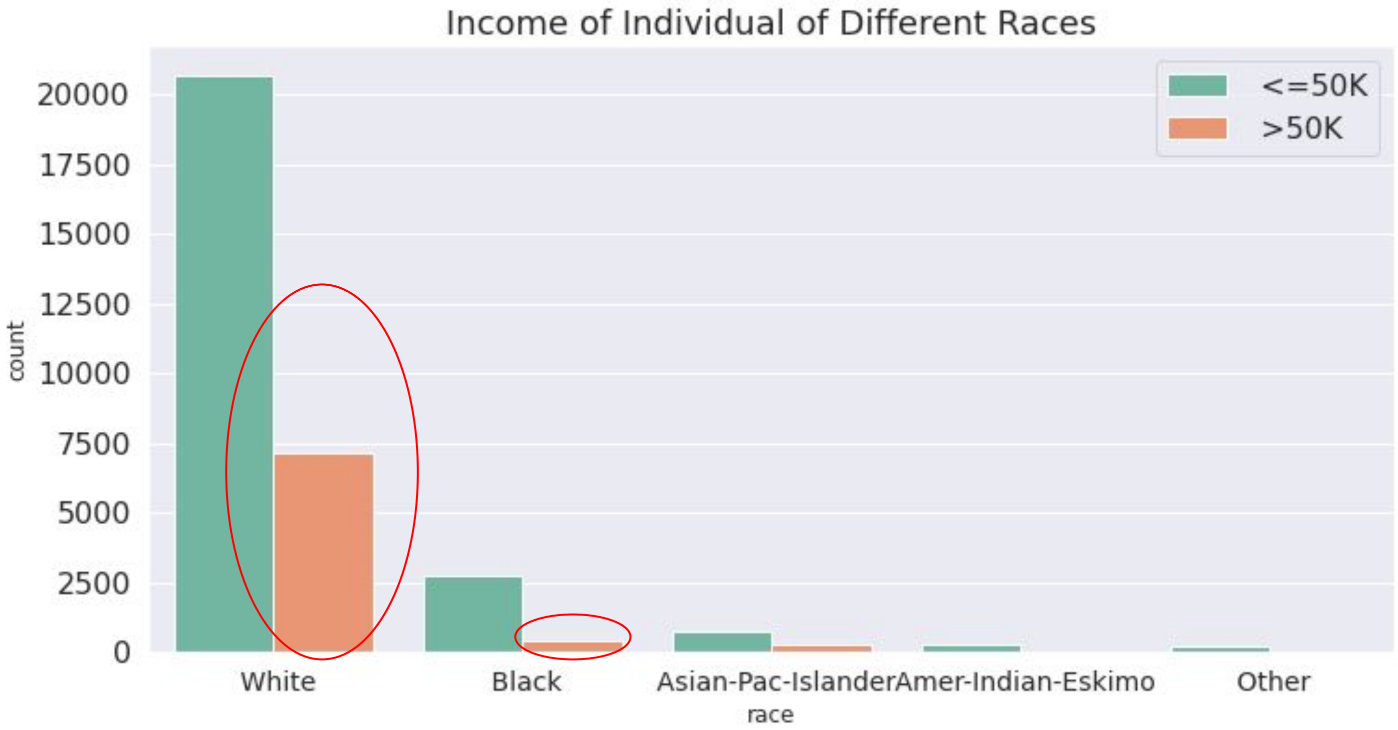
- **Correlations for original data:**





Data Insights

- **Race and Income**

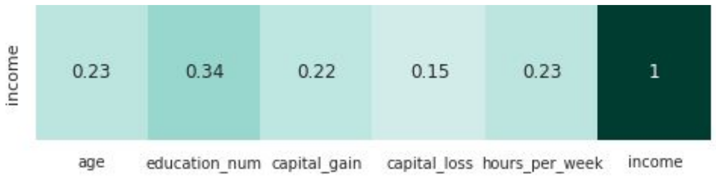


Data Insights

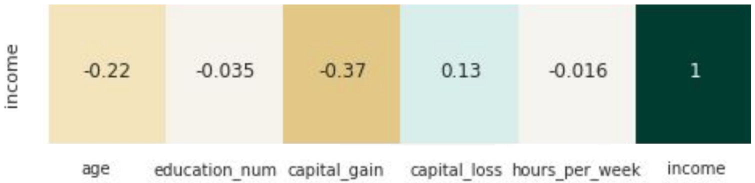
- **Gender and Income**

Preliminary Results: Utility

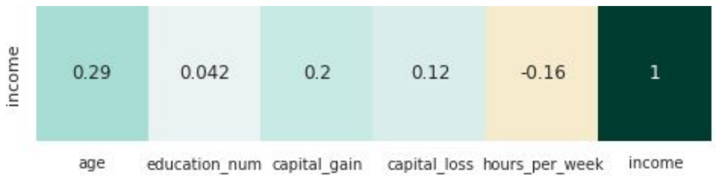
- **Preservation of original correlations**



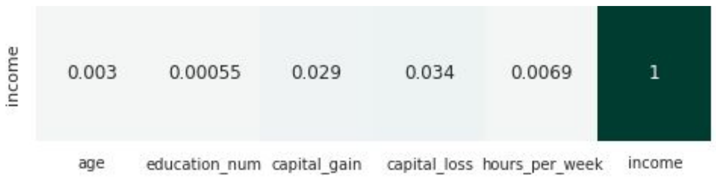
Original



PATE-GAN



TabularGAN



CTGAN

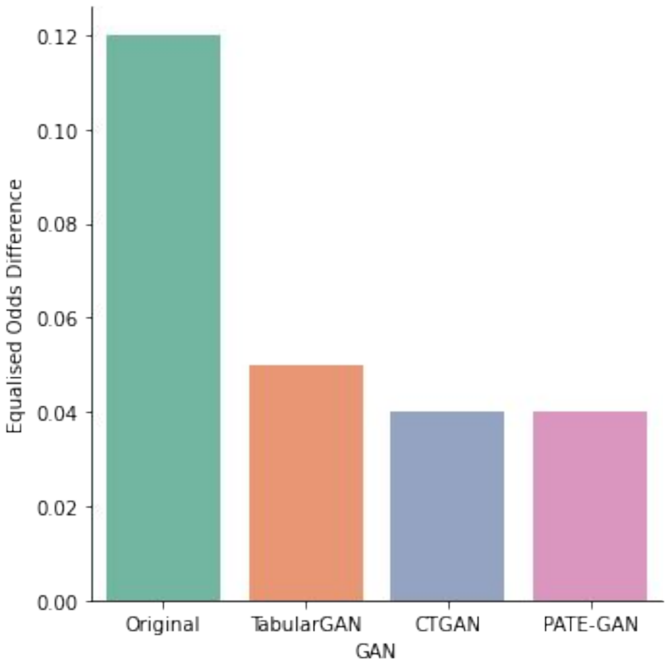
Preliminary Results: Privacy

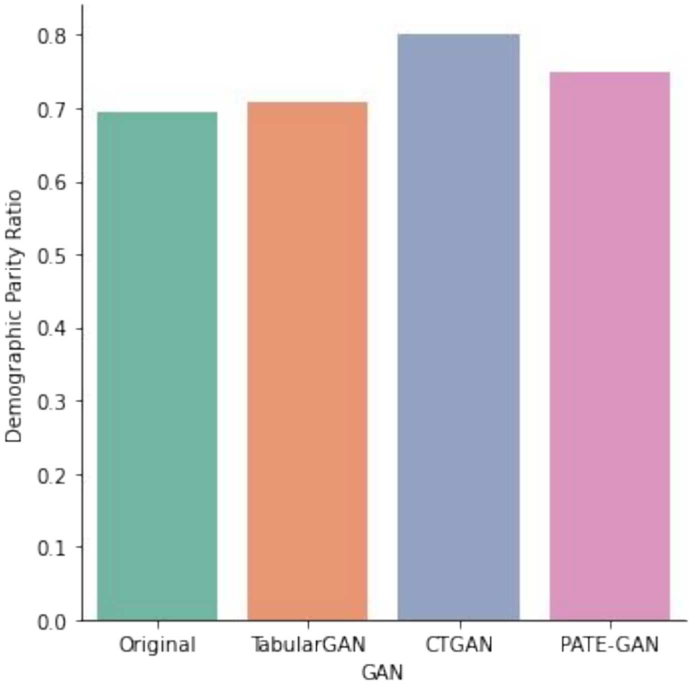
	TabularGAN	CTGAN	PATE-GAN
Distance-to-Closest Record (DCR)	✓	✓	✓

	TabularGAN	CTGAN	PATE-GAN
Nearest Neighbour Distance Ratio (NNDR)	✓	✓	✓

Preliminary Results: Fairness

- **Privileged Group** - Male
- **Unprivileged Group** - Female





Next Steps

- Work towards in depth analysis of results obtained.
- Vary the privacy budget (ε) in PATE-GAN and generate synthetic datasets for different values of ε.
- Run experiments to analyze the effect of ε on the fairness as well as the utility metrics.

Thank you, questions?

If you want to contact us later for any other questions or suggestions, feel free to email us at:

Bharathvaj Kumba Mothilal - kumbamot@ualberta.ca

Katyani Singh - katyani@ualberta.ca

Rudraksh Kapil - rkapil@ualberta.ca

References

1. Shokri, R., Stronati, M., Song, C., & Shmatikov, V. (2017, May). Membership inference attacks against machine learning models. In *2017 IEEE Symposium on Security and Privacy (SP)* (pp. 3-18). IEEE.

2. Lin, Z., Sekar, V., & Fanti, G. (2021, March). On the Privacy Properties of GAN-generated Samples. In *International Conference on Artificial Intelligence and Statistics* (pp. 1522-1530). PMLR.

3. Gupta, A., Bhatt, D., & Pandey, A. (2021). Transitioning from Real to Synthetic data: Quantifying the bias in model. *arXiv preprint arXiv:2105.04144*.

4. Xu, D., Yuan, S., Zhang, L., & Wu, X. (2018, December). Fairgan: Fairness-aware generative adversarial networks. In *2018 IEEE International Conference on Big Data (Big Data)* (pp. 570-575). IEEE.

5. Dua, D., & Graff, C. (2019). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.