# Research Topics in Data-centric Machine Learning – Synthetic Data and more

Nick Kim

September 2023

# What does "data-centric" mean?

- Two sides to a machine learning problem: model and data
- Fine-tuned model meets fine-tuned data: can we augment or improve the available *data* to boost overall performance?
  - At NeurIPS 2022, 99% model-centric papers vs. 1% data-centric papers

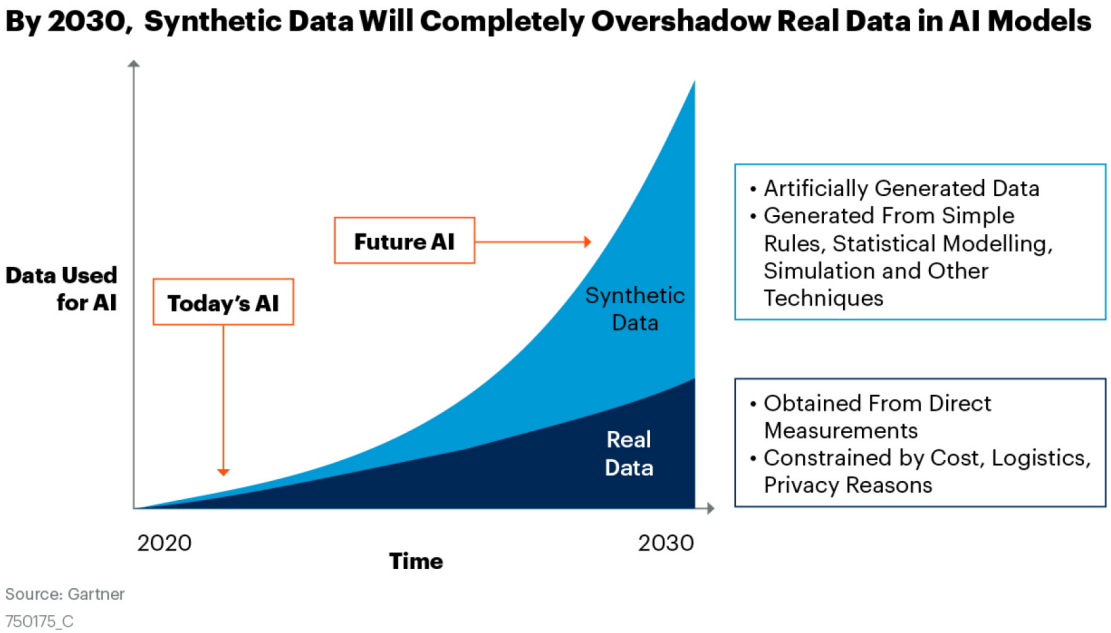
| | Steel defect detection | Solar panel | Surface inspection |
|---|---|---|---|
| Baseline | 76.2% | 75.68% | 85.05% |
| Model-centric | +0% (76.2%) | +0.04% (75.72%) | +0% (85.05%) |
| Data-centric | +16.9% (93.1%) | +3.06% (78.74%) | +0.4% (85.45%) |

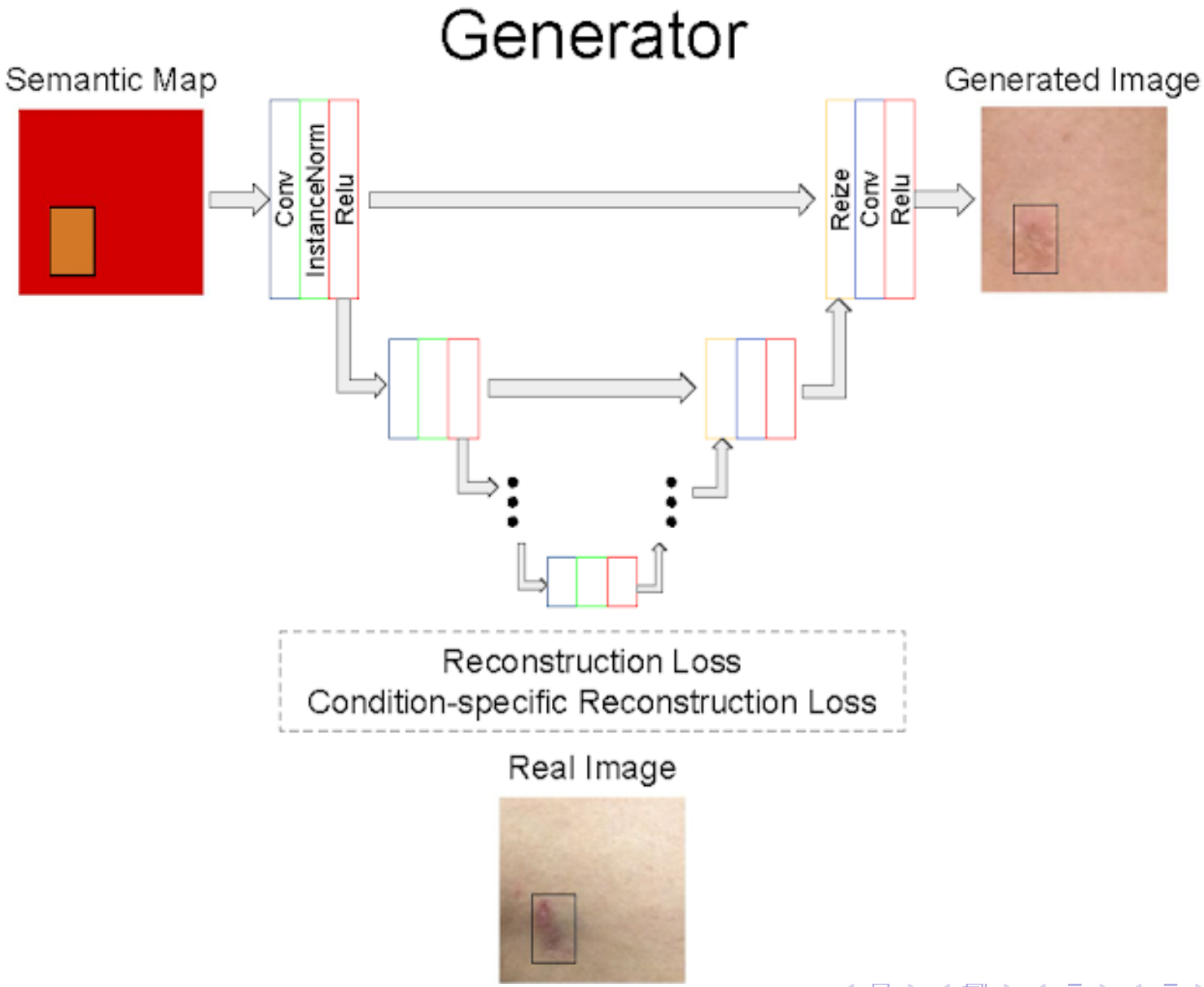
# Data-centric ideas and phases

1. Data collection, labeling, pre-processing, cleaning (1st generation of data company)
2. Data augmentation, pruning, generation – mostly algorithm-based (2nd generation of data company)
   - Increase training data volume, especially when certain class instances are rare, e.g., patients with a rare disease in hospital data
3. Regulation-driven: data governance issues such as bias, fairness, and privacy
   - Identify operational risks, e.g., synthesize loan applicants with diverse backgrounds to ensure fairness of AI loan decisions

# What is synthetic data?

- Artificially generated data meeting designated criteria which can be used alongside (or in place of) real data
- Can be synthesized in many different ways: noise addition, modeling marginal distributions of variables, fully generative models (GAN, VAE, diffusion, etc.)



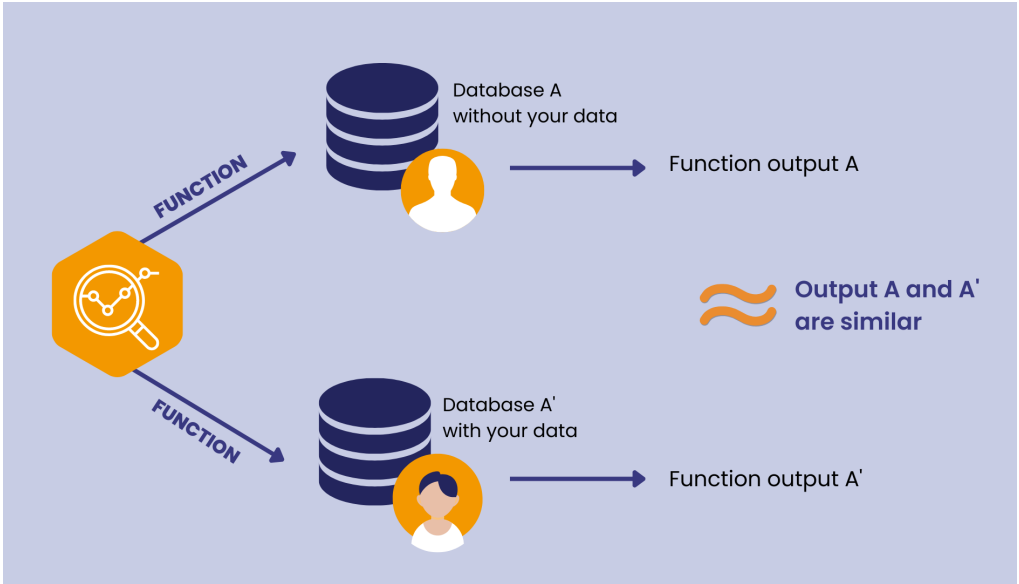

Generator
Semantic Map
Conv
InstanceNorm
Relu
Reize
Conv
Relu
Generated Image
Reconstruction Loss
Condition-specific Reconstruction Loss
Real Image

# Evaluating synthetic data

- Fidelity: how "close" synthetic data are to the real data, quantified using statistical measures such as KL-divergence
- Utility: performance on downstream machine learning tasks
- Privacy: differentially private synthetic data generation as alternative to private ML training, especially relevant in fields like healthcare and finance

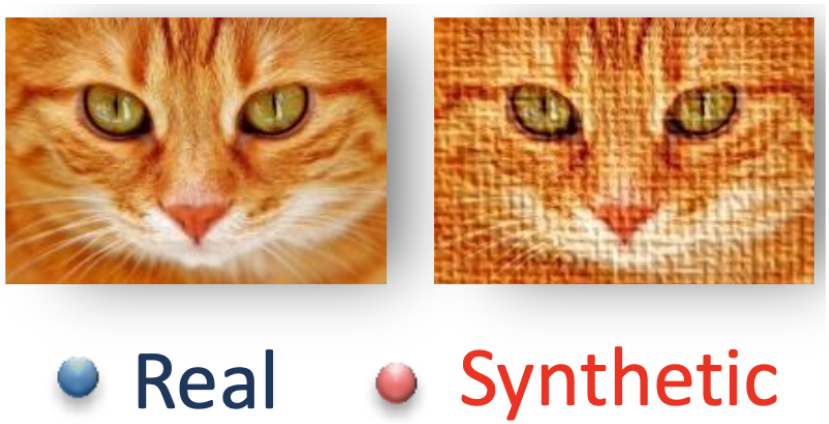
# The need for privacy and how to measure it

- Fake, synthetic data $\Rightarrow$ fully private? Not necessarily...
- The current "gold standard" for measuring privacy is $\varepsilon$-differential privacy. In a nutshell, $\varepsilon$-differential privacy prevents identification of specific individuals' data while still providing meaningful aggregate results.



# Auditing privacy

- Say we have a synthetic data generating model, e.g., a GAN.
- It is possible to fit an adversarial attack model on the data output by the GAN to predict sensitive attributes in the original data or even link synthetic data records back to the original data. (GANs have notorious "data copying" problem)
  - Specifically, given a data record $r$, can determine if $r$ was in the model's training dataset.



# Auditing privacy (cont.)

- Auditing involves applying such adversarial attacks on our model to quantify and investigate its robustness in terms of privacy.
  - Identify instances or circumstances where the privacy promise is violated!

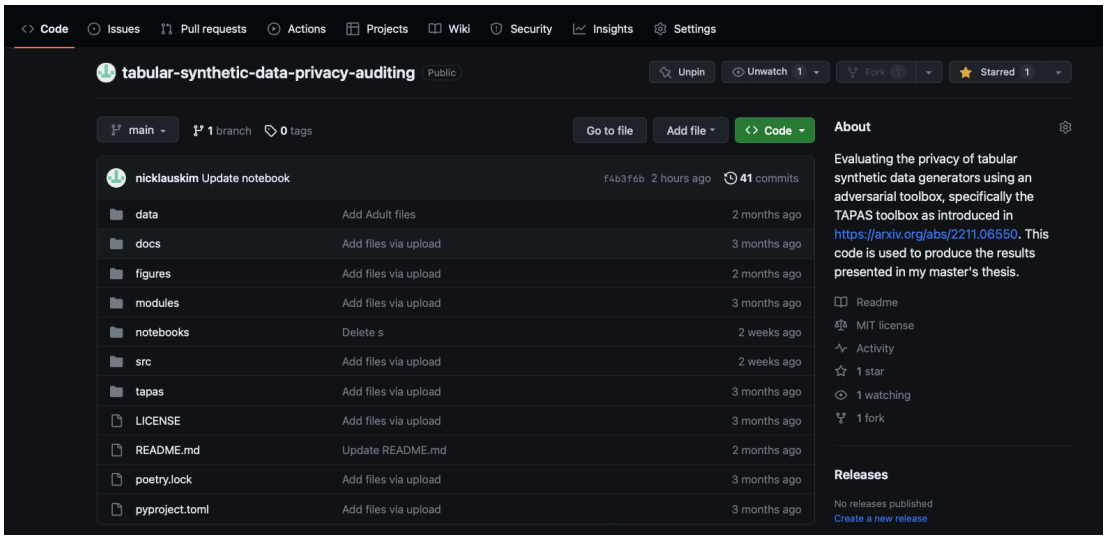



Figure: Python Package for Performing Privacy Auditing

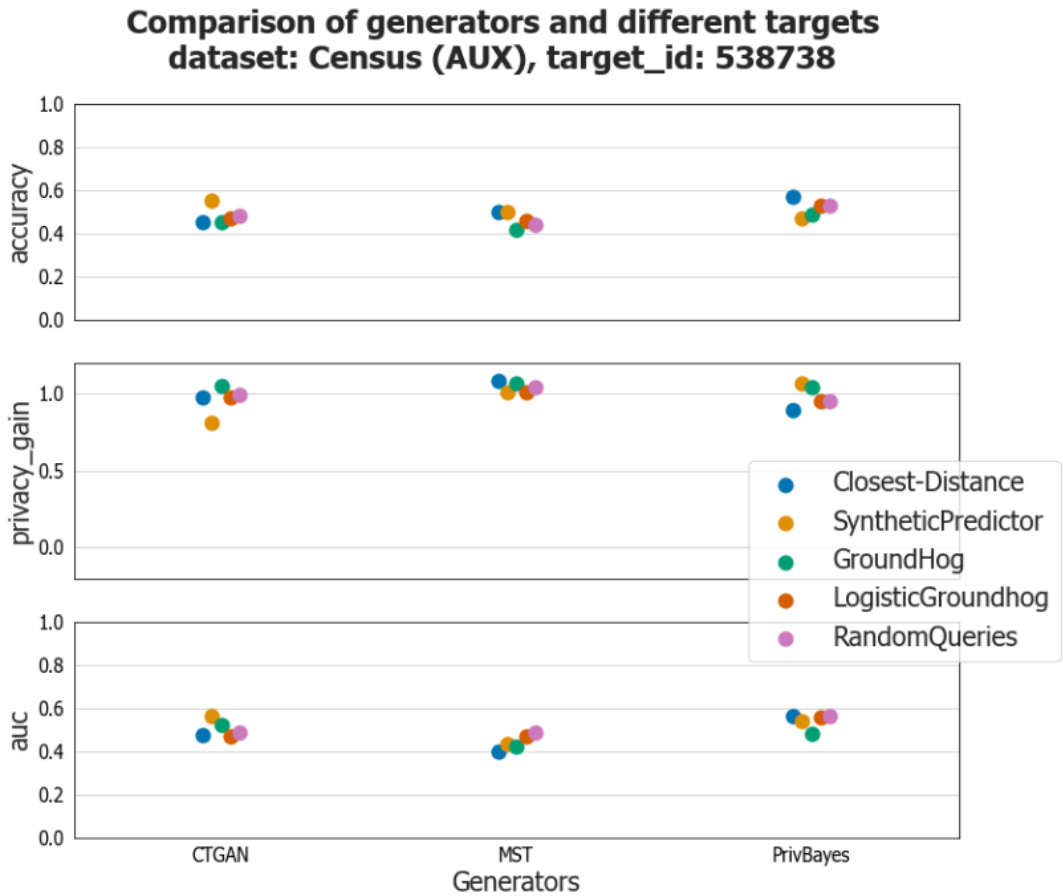



Figure: Attack Accuracy Metrics for Different Attacks, Generators

# A quick auditing application

- Consider dataset of hospital stays and discharges for thousands of patients in state of Texas
- Natural to use synthetic data as privacy workaround – train ML models on synthetic data instead of real data, or share synthetic datasets between data provider and analyst
- Can audit candidate generative models & answer such questions as:
  - Which model best preserves privacy under different attacks?
  - Do the audit results reveal anything about model differences?
  - Are any individual records at particularly high risk of privacy breach?

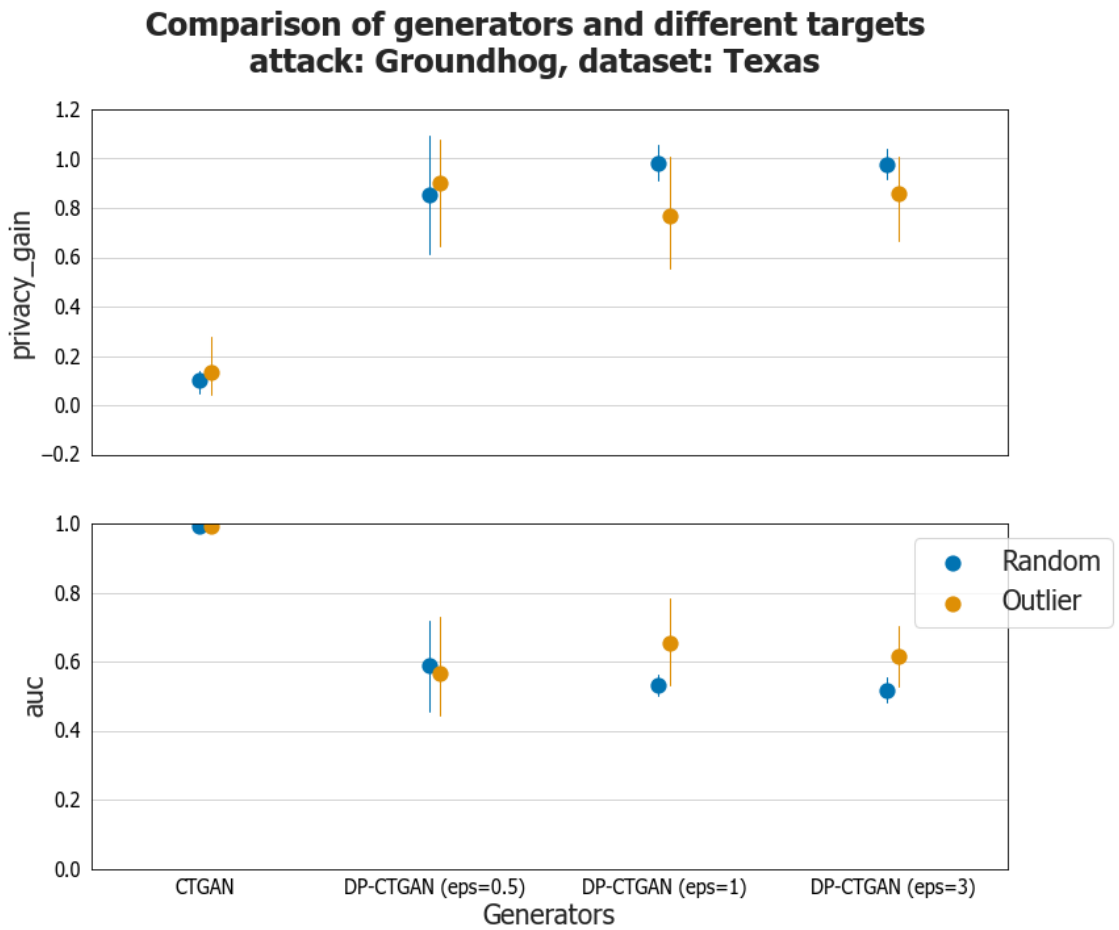



Figure: Random vs. Outlier Targets for Texas Dataset

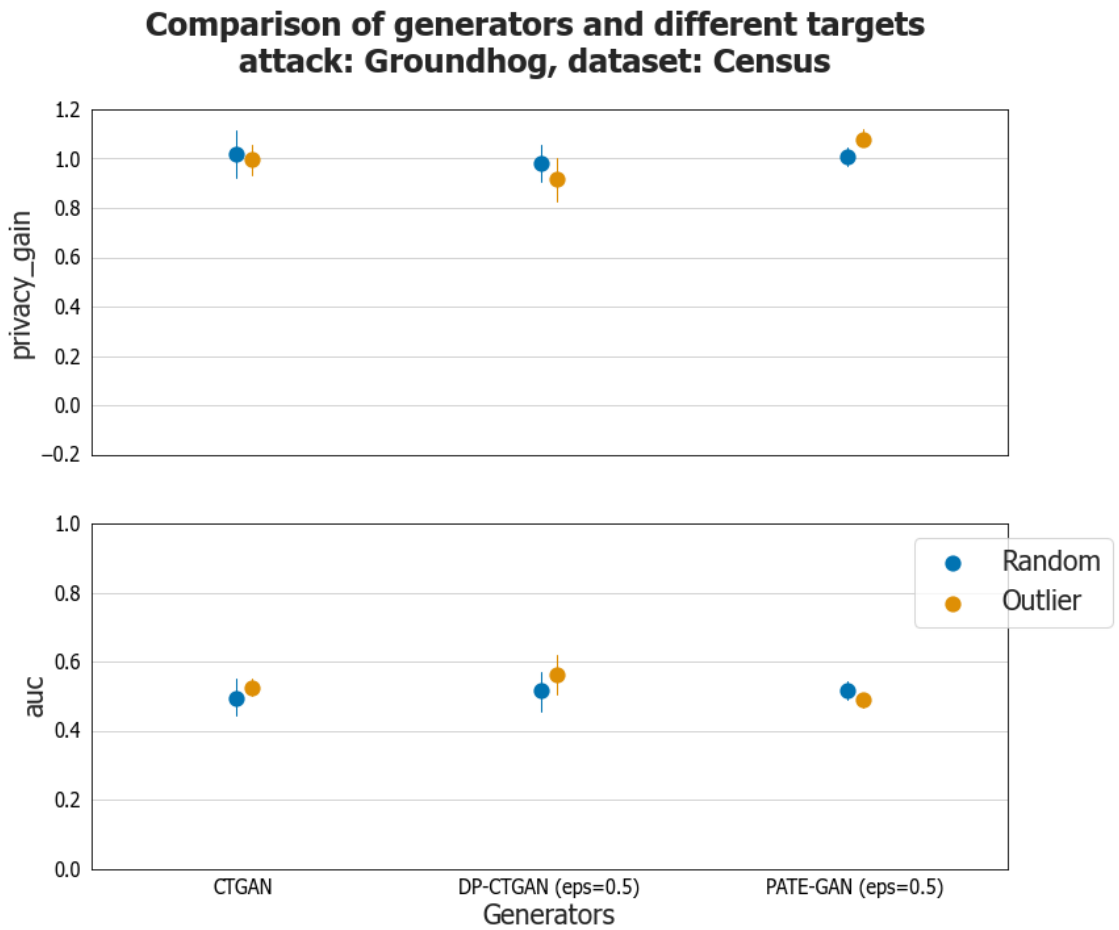



Figure: Random vs. Outlier Targets for Census Dataset

- Data-centric approaches are promising for increasing performance and robustness of machine learning systems.
- Synthetic data is one popular data-centric methodology with many benefits to data quality such as improving data volume, fairness, privacy, etc.
- Privacy in particular is a big (open) problem when it comes to synthetic data methodology.
- Using privacy auditing, we can assess the potential of different synthetic data generation models for truly private machine learning and safe, secure data sharing.

# References


- Houssiau, F., Jordon, J., Cohen, S. N., Daniel, O., Elliott, A., Geddes, J., Mole, C., Rangel-Smith, C., & Szpruch, L. (2022). TAPAS: A Toolbox for Adversarial Privacy Auditing of Synthetic Data (arXiv:2211.06550). arXiv. https://doi.org/10.48550/arXiv.2211.06550
- Jagielski, M., Ullman, J., & Oprea, A. (2020). Auditing Differentially Private Machine Learning: How Private is Private SGD? (arXiv:2006.07709). arXiv. https://doi.org/10.48550/arXiv.2006.07709
- Shokri, R., Stronati, M., Song, C., & Shmatikov, V. (2017). Membership Inference Attacks against Machine Learning Models (arXiv:1610.05820). arXiv. https://doi.org/10.48550/arXiv.1610.05820

# References (cont.)


- Stadler, T., Oprisanu, B., & Troncoso, C. (2022). Synthetic Data—Anonymisation Groundhog Day (arXiv:2011.07018). arXiv. https://doi.org/10.48550/arXiv.2011.07018
- Xu, L., Skoularidou, M., Cuesta-Infante, A., & Veeramachaneni, K. (2019). Modeling Tabular data using Conditional GAN (arXiv:1907.00503). arXiv. https://doi.org/10.48550/arXiv.1907.00503
- Yeom, S., Giacomelli, I., Fredrikson, M., & Jha, S. (2018). Privacy Risk in Machine Learning: Analyzing the Connection to Overfitting (arXiv:1709.01604). arXiv. https://doi.org/10.48550/arXiv.1709.01604

Thank you!