Statistics 414 From Predictive AI to Generative AI
Introduction to the Evaluation of Generative Data - Fidelity, Utility and Privacy

Lan Tao
lantao@ucla.edu

Department of Statistics & Data Science, UCLA

February 27, 2025

1. Synthetic Data Generator (SDG) Evaluation
2. Fidelity, Utility, and Privacy
3. Auditing Synthetic Data - Fidelity Evaluation
4. Auditing Synthetic Data - Utility Evaluation
5. Auditing Synthetic Data - Privacy Evaluation
6. Potential Research Direction

1 Synthetic Data Generator (SDG) Evaluation

2 Fidelity, Utility, and Privacy

3 Auditing Synthetic Data - Fidelity Evaluation

4 Auditing Synthetic Data - Utility Evaluation

5 Auditing Synthetic Data - Privacy Evaluation

6 Potential Research Direction

Synthetic Data Generator (SDG) Evaluation

.

A good SDG should satisfy:

- **Syntactical accuracy:** certain structural properties of the data are preserved
- **Privacy:** precisely quantify how much information about the original data is revealed
- **Statistical accuracy:** precisely quantify the statistical similarity between the synthetic and the original data
- **Efficiency:** the algorithm should scale well with the dimension of the data space

Synthetic Data Generator (SDG) Evaluation

.

- There is currently no systematic framework for developing SDGs for which all properties are satisfied simultaneously.
- Ullman et al. demonstrated that a computationally efficient algorithm (i.e. runs in polynomial time) that generates synthetic data that both:
  1. satisfies differential privacy
  2. preserves the correlations between pairs of features

  does not exist.
- No “one-size-fits-all“ differentially private synthetic data generation method. Synthetic data should be generated with a use case in mind.

1. Synthetic Data Generator (SDG) Evaluation
2. Fidelity, Utility, and Privacy
3. Auditing Synthetic Data - Fidelity Evaluation
4. Auditing Synthetic Data - Utility Evaluation
5. Auditing Synthetic Data - Privacy Evaluation
6. Potential Research Direction

Fidelity, Utility, and Privacy

Fidelity:

- Directly compare the synthetic dataset with the real dataset
- Measures how well the synthetic data statistically matches the real data

Fidelity, Utility, and Privacy

Fidelity:

- Directly compare the synthetic dataset with the real dataset
- Measures how well the synthetic data statistically matches the real data

Utility:

- Determined by its effectiveness in facilitating various downstream machine learning tasks
- Contrasting the performance of models on real vs synthetic data, inspecting concrete metrics (e.g. accuracy, mse, model fairness properties)
- Often requires train on synthetic, test on real (TSTR) paradigm

Fidelity, Utility, and Privacy

Fidelity:

- Directly compare the synthetic dataset with the real dataset
- Measures how well the synthetic data statistically matches the real data

Utility:

- Determined by its effectiveness in facilitating various downstream machine learning tasks
- Contrasting the performance of models on real vs synthetic data, inspecting concrete metrics (e.g. accuracy, mse, model fairness properties)
- Often requires train on synthetic, test on real (TSTR) paradigm

Privacy:

- Determined by the amount of information that it reveals about the real data used to produce it
- Differential privacy is required depending on the use case

Fidelity, Utility, and Privacy

Fidelity:

- Directly compare the synthetic dataset with the real dataset
- Measures how well the synthetic data statistically matches the real data

Utility:

- Determined by its effectiveness in facilitating various downstream machine learning tasks
- Contrasting the performance of models on real vs synthetic data, inspecting concrete metrics (e.g. accuracy, mse, model fairness properties)
- Often requires train on synthetic, test on real (TSTR) paradigm

Privacy:

- Determined by the amount of information that it reveals about the real data used to produce it
- Differential privacy is required depending on the use case

Fidelity v.s. Utility: They are not synonymous nor perfectly correlated, fidelity can be reduced while leaving utility unaltered in some scenarios.
Privacy v.s. Fidelity: When fidelity increases, the privacy of synthetic data decreases. Multiple synthetic datasets might need to be generated, each with user specified privacy guarantees.

Differential Privacy

- **Adjacency between two tabular datasets:** one can be obtained from the other by either the removal, addition, or replacement of a row
- Differential privacy requires that an algorithm's output not differ too much between adjacent datasets

Differential Privacy

- **Adjacency between two tabular datasets:** one can be obtained from the other by either the removal, addition, or replacement of a row
- Differential privacy requires that an algorithm's output not differ too much between adjacent datasets

Definition

A randomized algorithm $\mathcal{M}$ is (ϵ, δ)-differentially private if for all $S \in \text{Im}(\mathcal{M})$ and for all neighboring datasets $\mathcal{D}$, $\mathcal{D}'$:

$$\mathbb{P}(\mathcal{M}(\mathcal{D}) \in S) \leq e^{\epsilon}\mathbb{P}(\mathcal{M}(\mathcal{D}') \in S) + \delta.$$

Differential Privacy

- **Adjacency between two tabular datasets:** one can be obtained from the other by either the removal, addition, or replacement of a row
- Differential privacy requires that an algorithm's output not differ too much between adjacent datasets

Definition

A randomized algorithm $\mathcal{M}$ is (ϵ, δ)-differentially private if for all $S \in \mathrm{Im}(\mathcal{M})$ and for all neighboring datasets $\mathcal{D}$, $\mathcal{D}'$:

$$\mathbb{P}(\mathcal{M}(\mathcal{D}) \in S) \leq e^{\epsilon}\mathbb{P}(\mathcal{M}(\mathcal{D}') \in S) + \delta.$$

Remark:

- $|\log(\mathbb{P}(\mathcal{M}(\mathcal{D}) \in S)) - \log(\mathbb{P}(\mathcal{M}(\mathcal{D}') \in S))| \leq \epsilon$
- Regard Log likelihood as information, the difference of information, which is called **privacy loss**, should be within a given range.
- δ: **privacy leakage probability**
- With probability δ, information change is larger than ϵ; with probability $1 - \delta$, information change is less than ϵ.

1. Synthetic Data Generator (SDG) Evaluation
2. Fidelity, Utility, and Privacy
3. **Auditing Synthetic Data - Fidelity Evaluation**
4. Auditing Synthetic Data - Utility Evaluation
5. Auditing Synthetic Data - Privacy Evaluation
6. Potential Research Direction

Fidelity-driven Evaluation

- **Goal:** the **distribution** used to generate synthetic data is close to the (unknown) real data distribution
- Choosing a distance with which to compare distributions, then evaluating this distance empirically from samples of the real and synthetic datasets
- Marginals distribution can be efficiently estimated with total variational distance, correlations and Cramer’s V etc.
- Assess whether the basic properties of real data are captured, e.g. histograms of individual attributes, relations between pairs of attributes

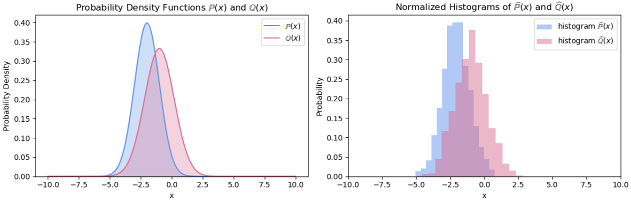


Figure: Example of distribution comparison

Fidelity-driven Evaluation

- Estimating distributional distances **in higher dimensions**, capturing relations between several attributes at a time, is challenging
- Bowen et al. have proposed that comparing the density of the synthetic and empirical distributions over random subsets of X
- Another solution is the **propensity score**, which captures the accuracy of a classifier trained to differentiate real from synthetic data points
- Sajjadi et al. proposed metrics of **precision** (the quality of synthetic samples) and **recall** (the diversity of synthetic samples), inspired by common failure modes of GANs.

Example of Fidelity Evaluation (1)

Real Data

$\mathbb{P}$

Generative Data

$\mathbb{Q}$

$$\mathbb{Q} = \alpha\mu + (1-\alpha)\nu_{\mathbb{Q}}$$

$$\mathbb{P} = \beta\mu + (1-\beta)\nu_{\mathbb{P}}$$

α Precision

β Recall

- The key intuition is that **precision** should measure how much of $\mathbb{Q}$ can be generated by a "part" of $\mathbb{P}$, while **recall** should measure how much of $\mathbb{P}$ can be generated by a "part" of $\mathbb{Q}$.
- Step 1: find the common support of $\mathbb{P}$ and $\mathbb{Q}$
- Step 2: Precision: how much Generative data $\mathbb{Q}$ in common support
  Recall: how much Real data $\mathbb{P}$ in common support
- *PRD*$(\mathbb{Q}, \mathbb{P})$: the set of attainable pairs of precision and recall of a distribution $\mathbb{Q}$ w.r.t. a distribution $\mathbb{P}$. It consists of all (α, β) satisfying the above equations.

Example of Fidelity Evaluation (1)

Real Data

$\mathbb{P}$

Generative Data

$\mathbb{Q}$

$$\mathbb{Q} = \alpha\mu + (1-\alpha)\nu_{\mathbb{Q}}$$

$$\mathbb{P} = \beta\mu + (1-\beta)\nu_{\mathbb{P}}$$

α Precision

β Recall

- The key intuition is that **precision** should measure how much of $\mathbb{Q}$ can be generated by a "part" of $\mathbb{P}$, while **recall** should measure how much of $\mathbb{P}$ can be generated by a "part" of $\mathbb{Q}$.
- Step 1: find the common support of $\mathbb{P}$ and $\mathbb{Q}$
- Step 2: Precision: how much Generative data $\mathbb{Q}$ in common support
  Recall: how much Real data $\mathbb{P}$ in common support
- *PRD*$(\mathbb{Q}, \mathbb{P})$: the set of attainable pairs of precision and recall of a distribution $\mathbb{Q}$ w.r.t. a distribution $\mathbb{P}$. It consists of all (α, β) satisfying the above equations.

The key question that these metrics seek to answer: how do the empirical and synthetic distributions overlap: precision (resp. recall) captures how much of the synthetic (resp. real) data falls within the support of real (resp. synthetic) data.

Example of Fidelity Evaluation (1)

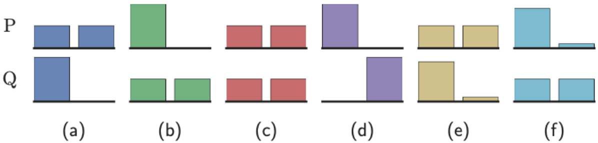


Figure 2: Intuitive examples of P and Q.

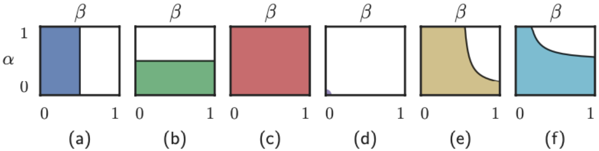


Figure 3: $\mathrm{PRD}(Q, P)$ for the examples above.

Example of Fidelity Evaluation (1)

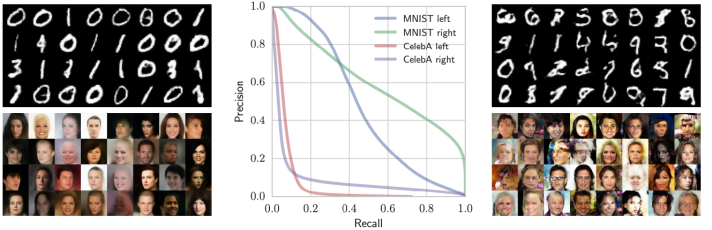


Figure 1: Comparison of GANs trained on MNIST and CelebA. Although the models obtain a similar FID on each data set (32/29 for MNIST and 65/62 for CelebA), their samples look very different. For example, the model on the left produces reasonably looking faces on CelebA, but too many dark images. In contrast, the model on the right produces more artifacts, but more varied images. By the proposed metric (middle), the models on the left achieve higher precision and lower recall than the models on the right, which suffices to successfully distinguishing between the failure cases.

Example of Fidelity Evaluation (1)

Theorem (Algorithm)

Let P and Q be two probability distributions defined on a finite state space Ω. For $\lambda > 0$ define the functions

$$\alpha(\lambda) = \sum_{\omega\in\Omega} min(\lambda P(\omega), Q(\omega)), \text{ and } \beta(\lambda) = \sum_{\omega\in\Omega} min(P(\omega), \frac{Q(\omega)}{\lambda}).$$

Then, it holds that

$$PRD(Q, P) = \{(\theta\alpha(\lambda), \theta\beta(\lambda)) | \lambda \in (0, \infty), \theta \in [0, 1]\}.$$

Remark:

- The set of $PRD(Q, P)$: a union of segments of the lines $\alpha = \lambda\beta$ over all $\lambda \in (0, \infty)$. Each segment starts at the origin $(0, 0)$ and ends at the maximal achievable value $(\alpha(\lambda), \beta(\lambda))$.
- This provides a surprisingly simple algorithm to compute PRD(Q, P) in practice.

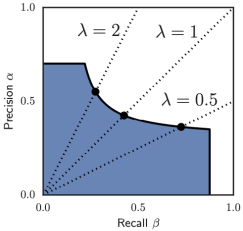


Figure 4: Illustration of the algorithm.

Example of Fidelity Evaluation (1)

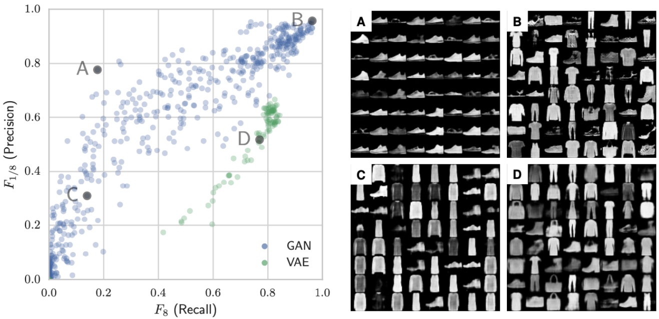


Figure 7: $F_{1/8}$ vs F_8 scores for a large number of GANs and VAEs on the Fashion-MNIST data set. For each model, we plot the maximum $F_{1/8}$ and F_8 scores to show the trade-off between precision and recall. VAEs generally achieve lower precision and/or higher recall than GANs which matches the folklore that VAEs often produce samples of lower quality while being less prone to mode collapse. On the right we show samples from four models which correspond to various success/failure modes: (A) high precision, low recall, (B) high precision, high recall, (C) low precision, low recall, and (D) low precision, high recall.

Example of Fidelity Evaluation (2)

- **Fidelity Auditor:** a synthetic data fidelity evaluation framework for comparing real data and synthetic data
- To frame synthetic data fidelity as real-synthetic data classification problem

Example of Fidelity Evaluation (2)

- **Fidelity Auditor:** a synthetic data fidelity evaluation framework for comparing real data and synthetic data
- To frame synthetic data fidelity as real-synthetic data classification problem
- $\mathbb{P}(x)$: density functions of real data; $\mathbb{Q}(x)$: density functions of synthetic data
- We use the total variation (TV) distance between $\mathbb{P}(x)$ and $\mathbb{Q}(x)$ to quantify the disparity between real and synthetic data

$$\mathrm{TV}(\mathbb{P}, \mathbb{Q}) = \frac{1}{2} \int_{\mathbb{R}^p} [\mathbb{P}(x) - \mathbb{Q}(x)]\, dx$$

Example of Fidelity Evaluation (2)

- **Fidelity Auditor:** a synthetic data fidelity evaluation framework for comparing real data and synthetic data
- To frame synthetic data fidelity as real-synthetic data classification problem
- $\mathbb{P}(x)$: density functions of real data; $\mathbb{Q}(x)$: density functions of synthetic data
- We use the total variation (TV) distance between $\mathbb{P}(x)$ and $\mathbb{Q}(x)$ to quantify the disparity between real and synthetic data

$$\mathrm{TV}(\mathbb{P}, \mathbb{Q}) = \frac{1}{2} \int_{\mathbb{R}^p} [\mathbb{P}(x) - \mathbb{Q}(x)]\, dx$$

- For any sample x from the mixed dataset $\mathcal{D}$, whose density function is $\mathbb{D}(x) = \frac{\mathbb{P}(x)+\mathbb{Q}(x)}{2}$, the probability of x being real is $\eta(x) = \frac{\mathbb{P}(x)}{\mathbb{P}(x)+\mathbb{Q}(x)}$.

Example of Fidelity Evaluation (2)

- $f : \mathbb{R}^p \to \{0, 1\}$ is a classifier used to discriminate real and synthetic samples, where 1 is the label for real and 0 is the label for synthetic.
- The expected classification error can be written as

$$R(f) = \mathbb{E}_X \left[I(f(X) = 1) \frac{\mathbb{Q}(x)}{\mathbb{P}(x) + \mathbb{Q}(x)} + I(f(X) = 0) \frac{\mathbb{P}(x)}{\mathbb{P}(x) + \mathbb{Q}(x)} \right],$$

where $X \sim \mathbb{D}$.

Example of Fidelity Evaluation (2)

- $f : \mathbb{R}^p \to \{0, 1\}$ is a classifier used to discriminate real and synthetic samples, where 1 is the label for real and 0 is the label for synthetic.
- The expected classification error can be written as

$$R(f) = \mathbb{E}_X \left[I(f(X) = 1) \frac{\mathbb{Q}(x)}{\mathbb{P}(x) + \mathbb{Q}(x)} + I(f(X) = 0) \frac{\mathbb{P}(x)}{\mathbb{P}(x) + \mathbb{Q}(x)} \right],$$

where $X \sim \mathbb{D}$.

- The minimal risk $R(f^*)$ is given as

$$R(f^*) = \frac{1}{2} \int_{\mathbb{R}^p} \min\{\mathbb{P}(x), \mathbb{Q}(x)\} dx = \frac{1}{2} - \frac{1}{2} \mathrm{TV}(\mathbb{P}, \mathbb{Q}).$$

Example of Fidelity Evaluation (2)

- $f : \mathbb{R}^p \to \{0,1\}$ is a classifier used to discriminate real and synthetic samples, where 1 is the label for real and 0 is the label for synthetic.
- The expected classification error can be written as

$$R(f) = \mathbb{E}_X \left[I(f(X)=1)\frac{\mathbb{Q}(x)}{\mathbb{P}(x)+\mathbb{Q}(x)} + I(f(X)=0)\frac{\mathbb{P}(x)}{\mathbb{P}(x)+\mathbb{Q}(x)} \right],$$

where $X \sim \mathbb{D}$.

- The minimal risk $R(f^*)$ is given as

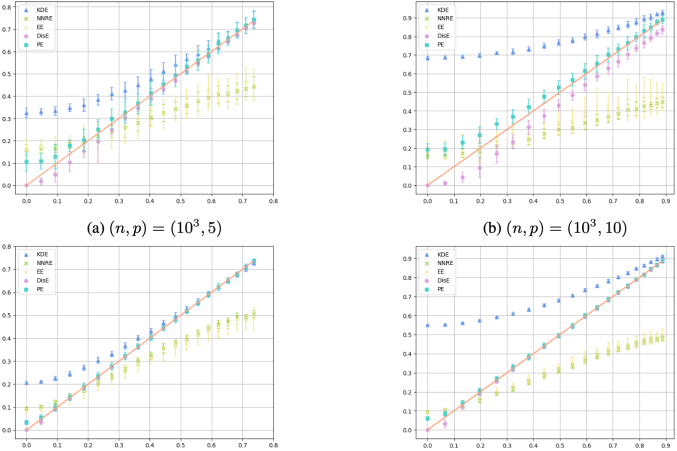
$$R(f^*) = \frac{1}{2}\int_{\mathbb{R}^p} \min\{\mathbb{P}(x), \mathbb{Q}(x)\} dx = \frac{1}{2} - \frac{1}{2}\mathrm{TV}(\mathbb{P},\mathbb{Q}).$$

- Since $R(\widehat{f}) \geq R(f^*)$, we obtain

$$\mathrm{TV}(\mathbb{P},\mathbb{Q}) \geq 1 - 2R(\widehat{f}),$$

where $\widehat{f}$ is any feasible classifier. Therefore, $\widehat{f}$ provides a means to establish a lower bound for the total variation distance between the distributions of real and synthetic data.distributions.

Example of Fidelity Evaluation (2)



(a) $(n,p) = (10^3, 5)$

(b) $(n,p) = (10^3, 10)$

(c) $(n,p) = (10^4, 5)$

(d) $(n,p) = (10^4, 10)$

Figure 2: True total variation (x-axis) versus estimated total variation (y-axis) in cases $(n,p) \in \{10^3, 10^4\} \times \{5, 10\}$ under varying disparity between two Gaussian distributions.

1. Synthetic Data Generator (SDG) Evaluation
2. Fidelity, Utility, and Privacy
3. Auditing Synthetic Data - Fidelity Evaluation
4. Auditing Synthetic Data - Utility Evaluation
5. Auditing Synthetic Data - Privacy Evaluation
6. Potential Research Direction

Utility-driven Evaluation

- Goal: the **accuracy** of a model estimated with synthetic data should reflect its accuracy on real data
- A particularly important application is the development of machine learning (ML) models to perform inference and classification tasks from data
- There are broadly two directions of research aiming to evaluate the suitability of synthetic data for ML training:
  (1) evaluating the performance of models trained on synthetic data
  (2) evaluating whether the relative performances of different models are similar on synthetic and real data.

Utility-driven Evaluation

- Goal: the **accuracy** of a model estimated with synthetic data should reflect its accuracy on real data
- A particularly important application is the development of machine learning (ML) models to perform inference and classification tasks from data
- There are broadly two directions of research aiming to evaluate the suitability of synthetic data for ML training:
  (1) evaluating the performance of models trained on synthetic data
  (2) evaluating whether the relative performances of different models are similar on synthetic and real data.
- The first approach assumes that analysts will train a machine learning model on the synthetic data, and use this model directly on real, future data. It is important that the accuracy of a model estimated with synthetic data reflects its accuracy on real data.

Utility-driven Evaluation

- Goal: the **accuracy** of a model estimated with synthetic data should reflect its accuracy on real data
- A particularly important application is the development of machine learning (ML) models to perform inference and classification tasks from data
- There are broadly two directions of research aiming to evaluate the suitability of synthetic data for ML training:
  (1) evaluating the performance of models trained on synthetic data
  (2) evaluating whether the relative performances of different models are similar on synthetic and real data.
- The first approach assumes that analysts will train a machine learning model on the synthetic data, and use this model directly on real, future data. It is important that the accuracy of a model estimated with synthetic data reflects its accuracy on real data.
- The second approach assumes that analysts use synthetic data to select a model, which is then trained on a real dataset (for better real-world performances). Crucially, the goal is that model development on synthetic data reflects model development on real data.

Utility-driven Evaluation

- Common metrics used to evaluate model performance
  1. **Accuracy:** Measures the overall correctness of predictions made by a model
  2. **Precision:** Measures the proportion of true positive predictions among all positive predictions made by the model. Useful when minimizing false positives is important.
  3. **Recall:** Measures the proportion of actual positive cases that are correctly identified by the model. Useful when capturing all positive cases is important.
  4. **F1-score:** Harmonic mean of precision and recall, providing a balance between them.

Utility-driven Evaluation

- Common metrics used to evaluate model performance
  1. **Accuracy:** Measures the overall correctness of predictions made by a model
  2. **Precision:** Measures the proportion of true positive predictions among all positive predictions made by the model. Useful when minimizing false positives is important.
  3. **Recall:** Measures the proportion of actual positive cases that are correctly identified by the model. Useful when capturing all positive cases is important.
  4. **F1-score:** Harmonic mean of precision and recall, providing a balance between them.
- Beaulieu et al. (2019) evaluate their synthetic data generation method by measuring the accuracy of classifiers trained on synthetic datasets on the real sensitive medical data used to generate the synthetic data
- Patki et al. (2016) distribute synthetic datasets and real datasets randomly to teams of data scientists, and evaluating whether teams working on real and synthetic datasets would arrive at approximately the same conclusions
- Tao et al. (2021) trained a XGBoost classifier on synthetic data and evaluated its performance on real data for a range of different tabular datasets.

1. Synthetic Data Generator (SDG) Evaluation
2. Fidelity, Utility, and Privacy
3. Auditing Synthetic Data - Fidelity Evaluation
4. Auditing Synthetic Data - Utility Evaluation
5. Auditing Synthetic Data - Privacy Evaluation
6. Potential Research Direction

Privacy-driven Evaluation

- **Goal:** protect against the threat models (membership inference; attribute inference; reconstruction attack)
- **DP verification:** checking that an algorithm meets DP requirements, only possible to show that an algorithm is likely to be differentially private
  - to understand what values of ϵ (privacy loss) make the most sense in a specific context
  - querying the algorithm in search of violations of the privacy definition, for both given δ and ϵ
  - running known attacks against it

Privacy-driven Evaluation

- **Goal:** protect against the threat models (membership inference; attribute inference; reconstruction attack)
- **DP verification:** checking that an algorithm meets DP requirements, only possible to show that an algorithm is likely to be differentially private
  - to understand what values of ϵ (privacy loss) make the most sense in a specific context
  - querying the algorithm in search of violations of the privacy definition, for both given δ and ϵ
  - running known attacks against it

Threat	Attacker's knowledge of Targeted Individual	Attacker's goal
Membership inference	Partial/Entire record	Determine if Targeted Individual was in the original data
Attribute inference	Partial record	Recover missing attributes of Targeted Individual's data
Reconstruction attack	N/A	Recover entire records from the original data

Privacy-driven Evaluation

- **Empirical privacy evaluation of datasets themselves:**
  - Although differential privacy cannot be established for a dataset in isolation, practitioners in the field of synthetic data have made use of a hold-out test set to evaluate the privacy of generated synthetic data.
  - Nearest-Neighbour distance ratio (NNDR):
    1. Split the data into a training set, and a test set, then use the training set to train the model.
    2. Once trained, draw samples from the trained model, and calculate the distance from these samples to the training data with the distance from the test set to the training data.

Privacy-driven Evaluation

- **Empirical privacy evaluation of datasets themselves:**
  - Although differential privacy cannot be established for a dataset in isolation, practitioners in the field of synthetic data have made use of a hold-out test set to evaluate the privacy of generated synthetic data.
  - Nearest-Neighbour distance ratio (NNDR):
    1. Split the data into a training set, and a test set, then use the training set to train the model.
    2. Once trained, draw samples from the trained model, and calculate the distance from these samples to the training data with the distance from the test set to the training data.
- **Attacks against private synthetic data:**
  - From the perspective of attackers: what can a motivated attacker learn about users in the dataset?
  - This approach helps to evaluate whether a system protects user privacy in a given context, and compare methods built with different privacy definitions in mind.

Example of Privacy Evaluation

- The main method to evaluate privacy risks in synthetic data was proposed by Stadler et al.(2020). They propose a general methodology to apply membership and attribute attacks on any synthetic data generation model.
- They assume black-box access to the SDG method, and specifically, being able to retrain the SDG model on new data.

Example of Privacy Evaluation

- The main method to evaluate privacy risks in synthetic data was proposed by Stadler et al.(2020). They propose a general methodology to apply membership and attribute attacks on any synthetic data generation model.
- They assume black-box access to the SDG method, and specifically, being able to retrain the SDG model on new data.
- Findings:
  1. Synthetic data either does not prevent inference attacks or does not retain data utility
  2. Synthetic data does not provide a better trade-off between privacy and utility than traditional anonymisation techniques
  3. Synthetic data leads to a highly variable privacy gain and unpredictable utility loss.
- Analysing synthetic data alone is not sufficient to properly understand information leakages. In order to audit synthetic data, it is necessary to be able to understand how it was generated.

1. Synthetic Data Generator (SDG) Evaluation
2. Fidelity, Utility, and Privacy
3. Auditing Synthetic Data - Fidelity Evaluation
4. Auditing Synthetic Data - Utility Evaluation
5. Auditing Synthetic Data - Privacy Evaluation
6. **Potential Research Direction**

Potential Research Direction

- **DP-verification:** For differential privacy auditing, how to choose appropriate two adjacent datasets to observe the information leakage?
- **Privacy evaluation:** With the observation that GANs can be vulnerable to membership inference attacks, how can these attacks be ported to the setup where the attacker has access to data generated by the model, rather than the model itself?
- **Multi-Faceted Evaluation Framework:** There are trade-off between utility and privacy of synthetic data, what kind of evaluation framework can be developed to comprehensively evaluate the three aspects of synthetic data?

Thank you!