Synthetic Data Generation and Evaluation

SERENE-RISC Workshop
October 22, 2020

Duc Phong Le

Roadmap

- **Problem statement**
- **Methods for synthesizing data**
  - Challenges in Synthetic Data Generation
  - Process of Synthetic Data Generation
  - Approaches
- **Evaluation of Synthetic Data**
  - Utility Assessment
  - Privacy Assessment

Problem Statement

CIC

- **Data** is getting more and more important
  - Analyzing data for making business decisions
  - Using data to train and validate machine learning models
  - Sharing data between clients to organizations, organization to organization, etc.
- **Sensitive data & Data privacy**
  - <u>Sensitive data</u>: data wherever it's loss could cause damage or distress to people/devices
  - <u>Data Privacy</u> is the necessity to preserve and protect any sensitive data, collected by any organization, from being accessed by a third party
- <u>One solution</u> that can provide data privacy is **synthetic data**

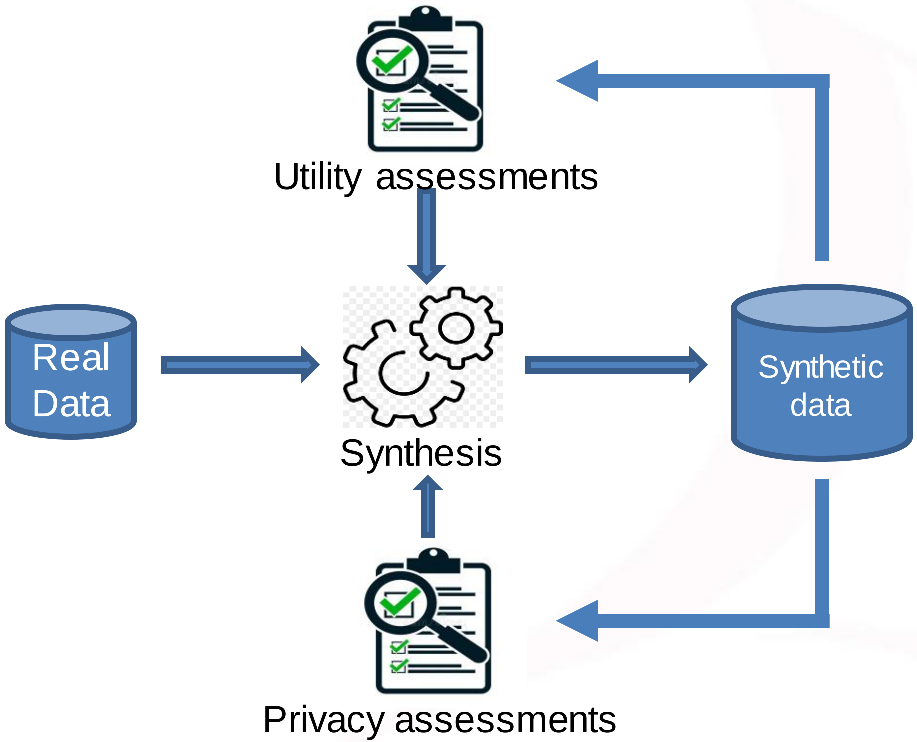
Generation Methods

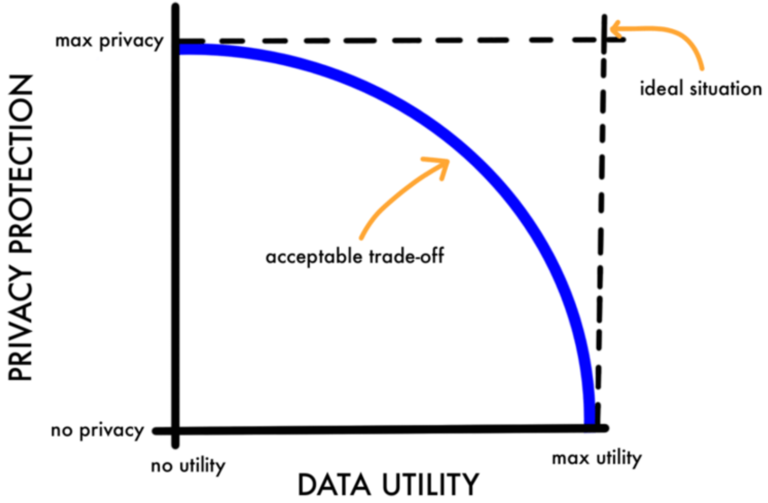
Challenges in Synthetic Data Generation

- **Providing data utility**
  - Have the same statistical properties, e.g., distribution, correlation, structure
  - All queries to the synthetic data would lead to the same result as to the original data

- **Providing the data privacy**
  - **Anonymity** based on regulatory standards, e.g., GDPR requirements
    - Singling out: possibility to identify an individual
    - Linkability: ability to link two record concerning the same data subject
    - Inference: capability of deducing one attribute value from other attribute values
  - **Differential privacy**: a theoretical privacy requirement

Process of Synthetic Data Generation





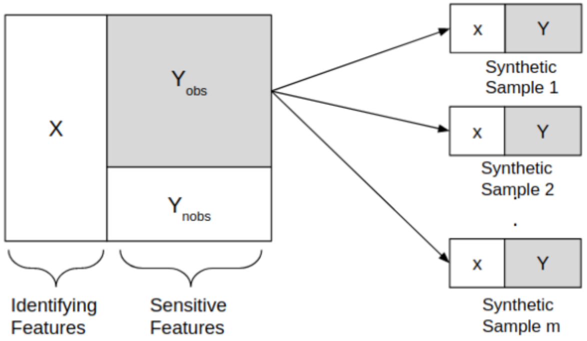
Synthetic Data Generation Approaches

- Synthetic data generation has been researched for nearly three decades
  - The first data synthesizer was introduced in 1993 by Rubin in the context of Statistical Disclosure Limitation (SDL) [1]
  - Three categories: Fully / Partially / Hybrid synthetic data
- Three main approaches:
  - Imputation-based methods
  - Full joint probability distribution methods
  - Generative Adversarial Networks (GAN)-based methods

[1] Rubin D. B. *Statistical disclosure limitation*. Journal of Official Statistic, 1993.

Imputation-based Methods

- Imputation was initially introduced in statistics as a technique to fill in missing data with substituted values in 1986
  - Given X and Y_obs, synthetically generate Y_nobs
- Proposed as a fully synthetic data generator by Rubin in 1993
  - Treating sensitive data as missing data
  - Releasing randomly sampled imputed values



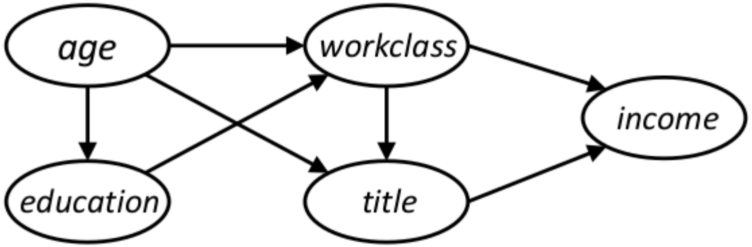
Sample *m* datasets from imputed population and release them publicly

- Randomness in the dataset due to sampling from population and imputed values

$$\begin{array}{lll} \text{DOB, Marital Status, Gender}, \textit{Income} & \rightarrow & \text{Census dataset} \\ \underbrace{\text{DOB, Marital Status, Gender}}_{\text{Identifying features}}, \textit{HIVStatus} & \rightarrow & \text{Health dataset} \end{array}$$

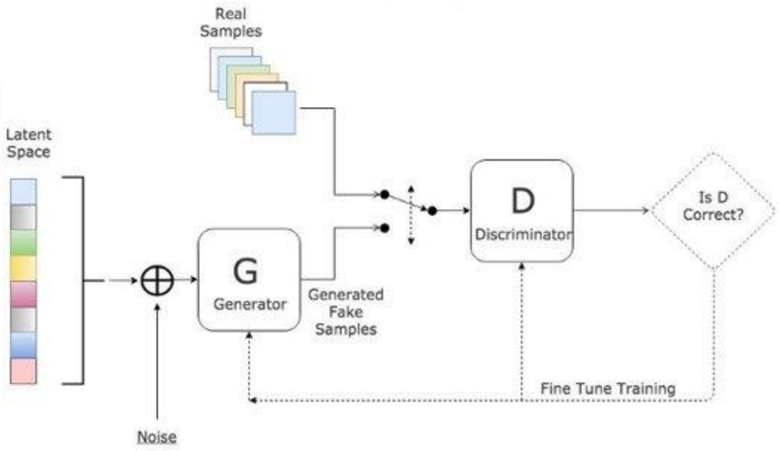
Full probability joint distribution methods

- Imputation-based methods may not preserve the correlation between attributes
- Full probability joint distribution methods learn and build a model about:
  - Marginal distributions
  - Joint distributions
  - Correlations between variables.
- For example, using Bayesian networks:
  - A probabilistic graphical model representing a set of variables and their conditional dependencies via a directed acyclic graph (DAG)
  - For example, the relationship between education and income, age and health, etc



Generative Adversarial Networks (GAN)

- GAN (Generative Adversarial Networks):
  - a popular class of Deep Neural Networks (DNN)
  - produces two joint-trained networks **Generator** and **Discriminator**
  - Generator: generates synthetic data intended to be similar to the training data
  - Discriminator: tries to discriminate the synthetic data from the true training data
  - These two networks contest in a game often in the form of **zero-sum** game




Time to
evaluate

Testing Criteria for Synthetic Data

CIC

Goal

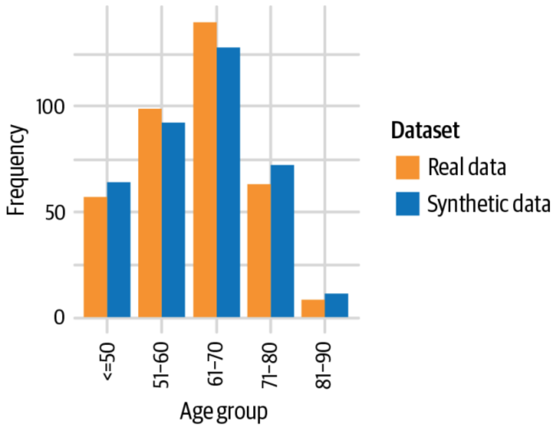
- Evaluate the Utility and Privacy of Synthetic Data

Two approaches

- Statistical Measurements
  - Evaluate the similarity between two datasets through statistical information
- AI-based Measurements
  - Using ML algorithm to train/test real and synthetic data then measure/compare similarity metrics metrics

Utility Assessments

- Univariate Distributions
  - Basic Stats such as mean, median, histograms, etc
- Joint-distributions
  - Compare joint-distributions of variables in real data and synthetic data
- Correlation between variables
  - Compare correlation matrices of the real data and synthetic data
- Machine learning score similarity
  - For ex: accuracy, F-1 score for classification and MRE (mean relative error) for regression tests



Privacy Assessments

- **Differential Privacy** (DP): theoretical privacy requirements
  - Common methods to realize DP include Laplace, Gaussian, Exponential and Global sensitivity mechanisms

- **Requirements from Privacy Acts:**
  - For example: Singling out, Linkability, Inference
  - <u>Possible measurements</u>: Hitting rate, Record linkage, and Distance to the Closest Records (DCR)

CIC

Q & A

Thank you for your attention...