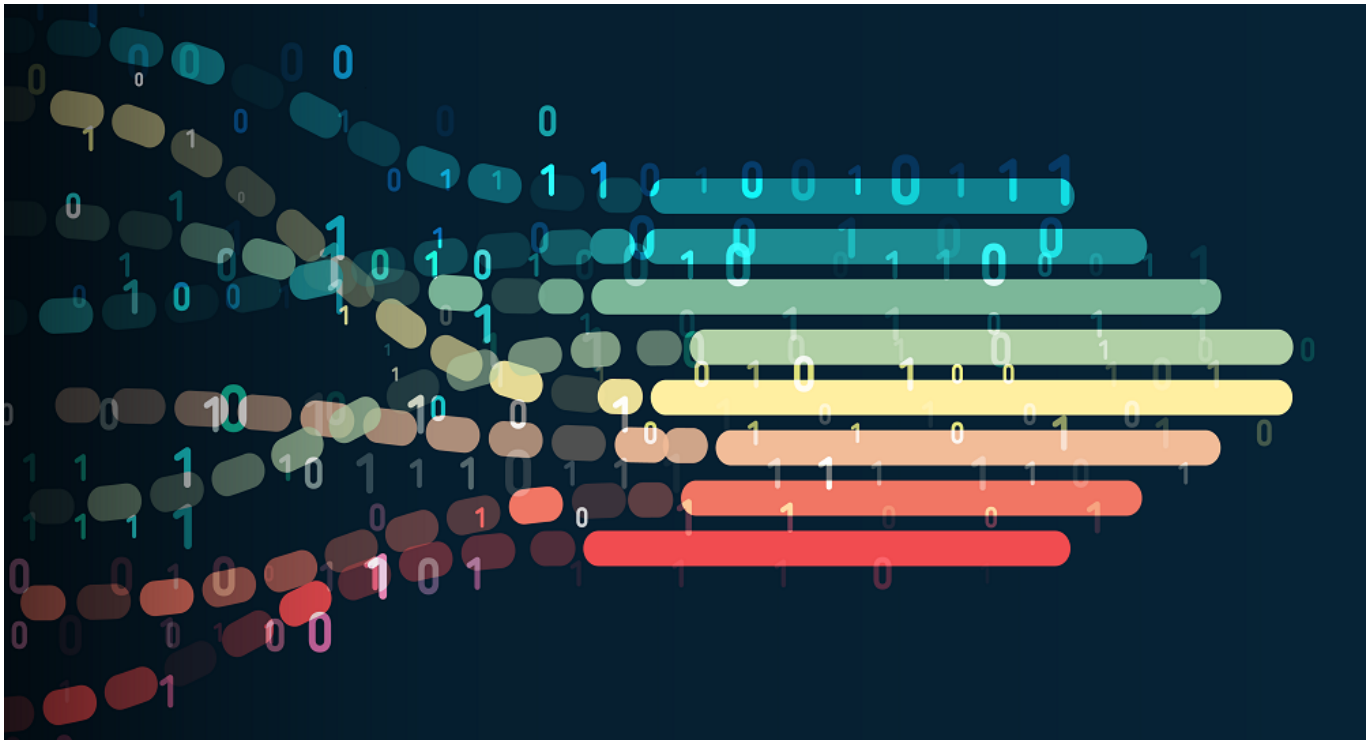
TECH OFFER

Synthetically-generated Privacy-preserving Data for Machine Learning



KEY INFORMATION

TECHNOLOGY CATEGORY:
Infocomm - Data Processing
Infocomm - Security & Privacy

TECHNOLOGY READINESS LEVEL (TRL): TRL7
COUNTRY: SINGAPORE
ID NUMBER: TO174501

OVERVIEW

Artificial Intelligence/Machine Learning (AI/ML) performance is predicated on training with good quality data.
However, such data is often difficult to acquire due to ethical concerns, logistic problems, high cost, data bias, and inherent poor data quality.

Privacy restrictions and data regulations further compound the problem of data acquisition, restricting many organisations long-term access to valuable historical data.
Ultimately, this creates the problem of incomplete or biased data which degrade the overall performance of trained AI/ML models.

This technology offer is a controlled synthetic data generation with differential privacy capability for structured (tabular) data.
Its synthetic data engine utilizes conditional GANs (cGANs) coupled with optional differential privacy to synthesize data with

similar properties as real data without the associated privacy risks.

TECHNOLOGY FEATURES & SPECIFICATIONS

The core technology is a synthetic data engine that learns the distribution of the input data and selects the column to generate based on this distribution. Gaussian noise is further added to the gradients to protect the privacy of the data.

The technology can generate data quickly: 10,000 rows, 8 columns in 8 minutes (evaluated on Nvidia GTX1080) and is mainly intended to generate synthetic datasets to address data scarcity, data privacy, and data augmentation. This generative process involves the following features:

- Conditional Generative Adversarial Networks (cGANS) generate synthetic data that mimic real data
- Sensitive data is obfuscated with statistical noise and randomization
- Definable privacy levels allowing adjustability between utility and data privacy
  (Differential privacy allows Machine Learning models to be trained on synthetic tabular data and achieve similar results as models trained on real data)
- Quality Assurance (QA) component generates reports to aid the assessment of data quality and risk metrics
- APIs for rapid integration, with full customisability

POTENTIAL APPLICATIONS

This technology can be used for the following types of structured data:

- non-time series
- time series
- multi-tables
- free-text fields

It can be applied in the following use-cases:

Data Augmentation

Increase the size of your datasets without wasting time to procure new data

Data Extrapolation

Extrapolate known data to generate unavailable or unknown data points

Bias Correction

De-bias or equalize the distribution of datasets

Targeted Generation

Generate rich data, including infrequent scenarios

UNIQUE VALUE PROPOSITION

This synthetic data generation with differential privacy technology provides accessible privacy by design - adding privacy-preserving techniques before, during or after AI training, together with the following benefits:

- Synthetic data does not require further data sanitization, providing a safe data sandbox environment
- Reduces the need to pay for additional datasets by generating missing data or de-biasing existing datasets
- Overcomes the challenges of data acquisition by enriching real data with synthetic data through controlled generation
- Synthetically generated data become your data assets, with potential for monetization as new revenue streams
- Protect real data by combining *made up* data points to make it harder to distinguish what is real even if data is compromised
- Indefinite retention time without associated compliance risks and full accessibility to rich statistical data to provide a boost to AI/ML model resilience and performance

The technology owner is looking to collaborate with technology partners in the field of AI/ML to co-develop new products/services, and for collaborators to test-bed in pilot projects.