Guide on Synthetic Data Generation

JOINTLY DEVELOPED WITH

A*STAR

(Agency for Science, Technology and Research - Singapore)

Presented to CANON

22 October 2024

pdpc
PERSONAL DATA PROTECTION COMMISSION SINGAPORE

Personal Data Protection Commission (PDPC)

Responsibility

The PDPC was set up in 2013 to administer and enforce the Personal Data Protection Act (PDPA). The PDPC serves as Singapore's main authority in matters relating to personal data protection and will represent the Singapore Government internationally on data protection related issues.

Our Mission

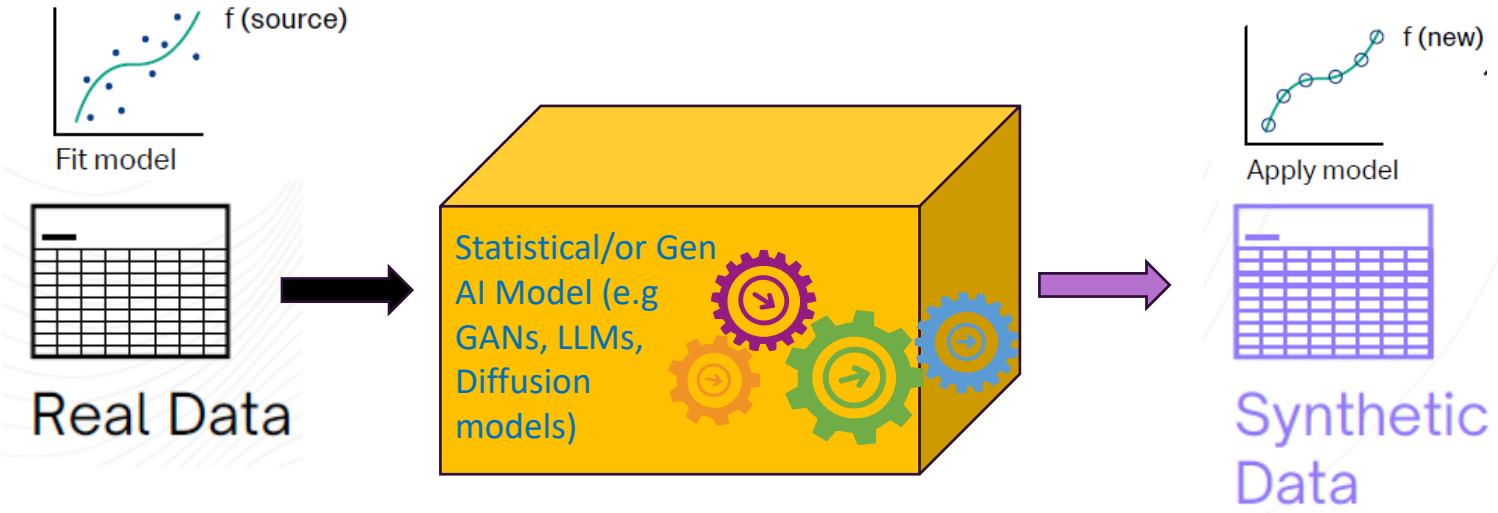
- To promote and enforce personal data protection to foster an environment of trust among businesses and consumers, contributing to a vibrant Singapore economy.
- To provide guidance to private sector organisations to address AI governance issues when using artificial intelligence solution.
- To contribute to strengthening Singapore's position as a trusted hub for businesses and cross border data flows.

Privacy Enhancing Technologies (PET)

Categories of PETs	PETs	Examples of applications (non-exhaustive)
Data obfuscation	Anonymisation/pseudonymisation techniques	• Secure storage • Data sharing and retention • Software testing
	Synthetic data generation	• Privacy-preserving AI machine learning • Data sharing and analysis • Software testing
	Differential privacy	• Expanding research opportunities • Data sharing
	Zero knowledge proofs	• Verifying information without requiring disclosure (e.g., age verification)
Encrypted data processing	Homomorphic encryption	• Secure data stored in cloud • Computing on private data that is not disclosed
	Multi-party computation (including private set intersection)	• Computing on private data that is not disclosed
	Trusted execution environments	• Computing using models that need to remain private • Computing on private data that is not disclosed
Federated analytics	Federated learning	• Privacy-preserving AI machine learning
	Distributed analysis	

Definition of Synthetic Data

Synthetic data is data that has been generated using a purpose-built mathematical model or algorithm, which replicates patterns and statistical properties of real data. As a result, performing analysis on synthetic data can produce similar results as using the real-world data.



Generating Synthetic Data

Fit a sample of real data into a Statistical/or Gen AI model to generate a set of new data with the same inherent statistical distribution

Barriers faced in adoption of Synthetic Data

Synthetic data is not inherently risk-free!

Lack of knowledge about Use Case-to-Synthetic Data fit

Education and awareness is needed if widespread adoption is to be achieved

There is a general lack of understanding of what synthetic data offer as a solution. Thus, companies cannot envision how it will apply for their use case

Unclear about regulatory boundaries of use

Measure of data utility and/or privacy risk threshold

There are many examples of approaches to measure data utility and/or privacy risk threshold but currently there is NO unified standard risk threshold and this topic is still in academic debates

Tension between data utility and privacy

Finding the right balance between data utility and privacy

There is often a trade-off between privacy and utility whenever synthetic data is applied ie the more synthetic data method protects privacy, there is a decrease in utility of data

Use case archetype 1 : Generate training datasets for AI Models

Types of Use Cases	Key Benefits	Good Practices to generate synthetic data
Augment data for AI/ML models	Synthetic data addresses the challenge of the user having to obtain large volumes of labelled data needed for training and testing AI/ML models due to costs, legal regulations, and proprietary rights.	• ***Add noise* to or reduce granularity*** of the synthetic data points. • ***Such fictitious new data points will generally not be considered personal data***. *If the statistical properties/ characteristics of the synthetic data is representative of the population in question and not significantly skewed towards a specific individual/ group of individuals used as source training data, adding of noise might not be necessary as re-identification risks are generally low.
	Augmenting training datasets with synthetically generated labelled data can be more ***cost-effective***, especially when the source datasets are sparse.	
Increase data diversity for AI/ML models	Synthetic data can be used to simulate rare events or augment under-represented groups in training AI models.	
	Diverse datasets can be useful in improving performance of AI/ML models.	

Use case archetype 2 : Data analysis and collaboration

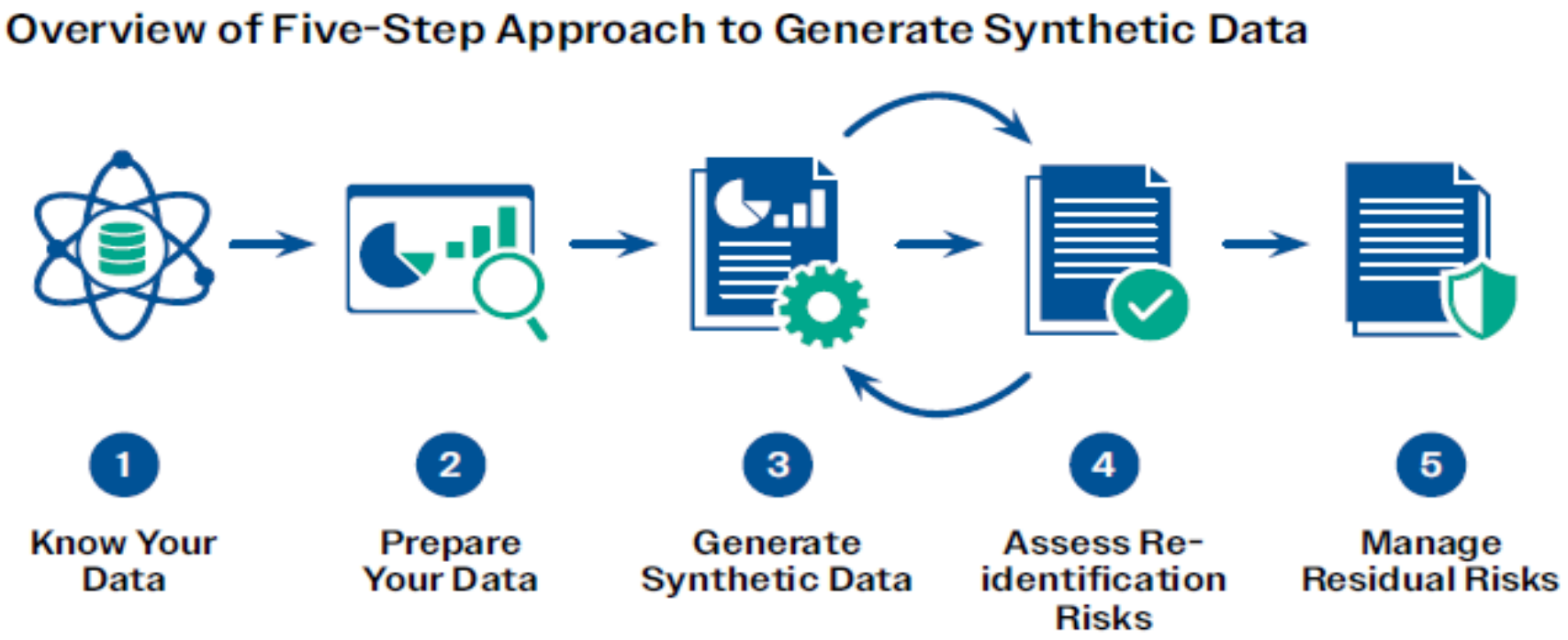
Types of Use Cases	Key Benefits	Good Practices to generate synthetic data
Data sharing and analysis	Underlying trends or patterns,and biases of the data are useful for data analytics regardless of whether the data source is real or synthetic.	Balance the trade-offs between data utility and data protection by ***incorporating data protection measures throughout the synthetic data generation process***, for example: <u>Data Preparation</u> • Remove outliers from source data • Pseudonymise source data • Employ data minimisation and generalise granular data <u>Synthetic data generation</u> • Add noise before or after as well as during synthetic data generation <u>Post-synthetic data generation</u> • Incorporate technical, contractual, and governance measures to mitigate any residual re-identification risks
	Synthetic data can enable data sharing for analysis especially in industries and sectors, e.g., healthcare, ***where the source data can be sensitive.***	
Previewing data for collaboration	Synthetic data can be used in data exploration, analysis, and collaboration to provide stakeholders with a representative preview of the source data without exposing sensitive information.	
	This enables stakeholders to explore and understand the structure, relationships, and potential insights within the data **to *gain assurance of the data quality*** before finalising any agreement or collaboration.	

Use case archetype 3 : Software Testing

Types of Use Cases	Key Benefits	Good Practices to generate synthetic data
System development/ Software testing	Organisations can use synthetic data instead of production data to ***facilitate software development.***	***Focus on generating synthetic data that follows semantics*** e.g., format, min/max values and categories, of source data instead of the statistical characteristics and properties.
	Use of synthetic data can help organisations ***avoid data breaches*** in the event of the development environment being compromised.	

Handbook on Key Considerations and Best Practices in Synthetic Data Generation

PDPC recommends a set of good practices and considerations for generating synthetic data through the five-steps approach to reduce re-identification risks



Step 1 – Know Your Data

Before embarking on any synthetic data project, it is necessary to have a clear understanding of the purpose and use cases of the synthetic data and the source data that the synthetic data is to mimic. This will help to determine whether use of synthetic data might be relevant and identify the possible risks of using the synthetic data. Some of the considerations may include:

- Where general trends/insights of source data are sensitive, organisations should take note that the use of synthetic data will not offer any protection to the trends/insights since they will be replicated in the synthetic data.
- Where the synthetic data is intended to be released publicly, organisations may have to prioritise data protection over data utility in such circumstances.
- Where relevant, organisations should also put in place proper contractual obligations on recipients of synthetic data where necessary to prevent re-identification attacks on the data.

❖ The main goal is to establish objectives prior to synthetic data generation to determine an acceptable risk threshold of the generated synthetic data and the expected utility of the data. This will help provide organisations with the appropriate benchmarks to assess any trade-offs between data protection risks and data utility.

Step 2 – Prepare your data

When preparing the source data for generating synthetic data, it is important to consider the following:

- What are the key insights that needed to be preserved in the synthetic data?
- Which are the necessary data attributes for the synthetic data to meet the business objectives?

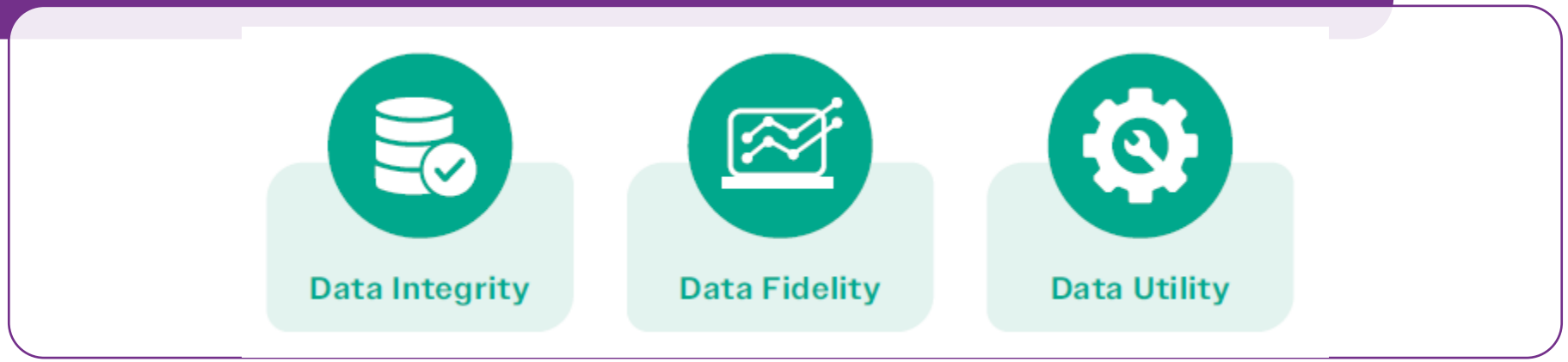
❖ Understanding the key insights to be preserved will ensure that the generated synthetic data will be able to meet the business needs. At this point, organisations should consider whether outlier trends and insights are necessary to be preserved for the business objectives.

❖ Selection of the relevant attributes from source data for generating synthetic data is a good practice for applying data minimization principles.

Step 3 – Generate Synthetic Data

There are many different methods to generate synthetic data, for example, sequential tree-based synthesisers, copulas, and deep generative models (DGMs). Organisations need to consider which methods are most appropriate, based on their use cases, data objectives, and types of data.

After generating synthetic data, it is a good practice for organisations to perform the following checks on the quality of the generated synthetic data:



- As a good practice, organisations should validate the integrity of the generated synthetic data against the data dictionary of source data

- Organisations should use performance metric(s) for data fidelity and data utility

Step 4 – Assess Re-identification Risks

Re-identification (or privacy) risk assessment for synthetic data is an attack-based evaluation. It evaluates how successful an adversary is able to determine an individual belongs to the source dataset or derive undisclosed details of an individual.

The different types of re-identification attacks (commonly referred to as privacy attacks) on synthetic datasets are:

- **A** Singling Out Attack
- **B** Linkability Attack
- **C** Inference Attack

- As a good practice, organisations should assess and perform the re-identification risk assessment after synthetic data has been generated and data quality checks has been performed.
- Ensure that the re-identification risk levels for the three key re-identification attacks are within acceptable risk level based on the organisation's internal acceptance criteria.
- Organisations may also need to engage the synthetic data solution provider to perform the re-identification risk assessments.

Step 5 – Manage Residual Risks

- Identify any potential residual risks such as (1) new insights derived from synthetic data (2) potential impact on groups of individuals (3) parties receiving synthetic data (4) changing environment (5) model leakage etc

- Identify any data breach risks for example (1) loss of fully synthetic data that was not intended for public release (2) loss of synthetic data generator model, parameters and/or synthetic data.

Best Practices and security controls

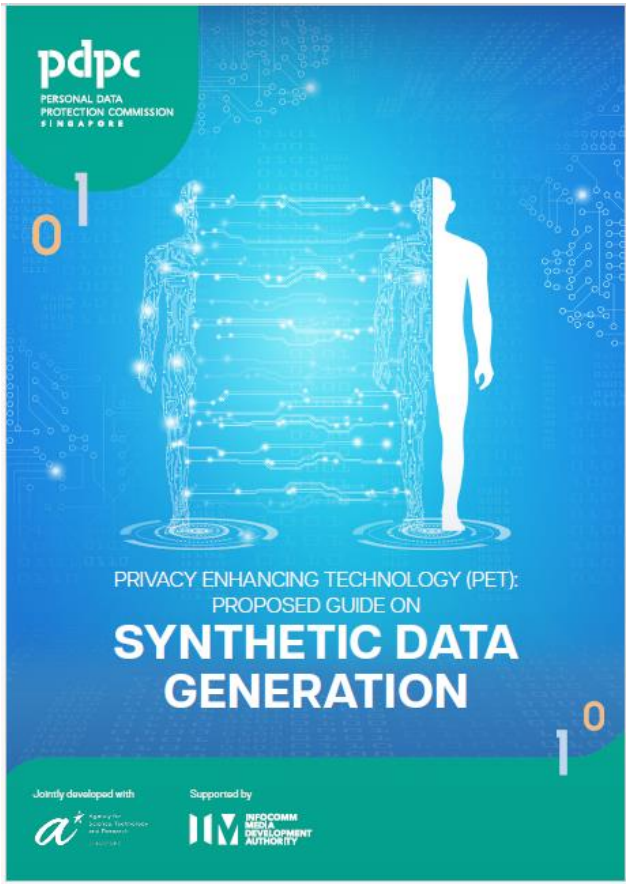
- Implement appropriate mitigation controls (governance, technical, process, contractual) to minimize the identified risks.

- Incorporate incident management plans to investigate root cause of any data breach incidents involving synthetic data, implement internal safeguards against such occurrences in the future and assess if such breach would be notifiable to PDPC.

❖ As a good practice, organisations should document all the potential residual and data breach risks as well as the safeguard measures and seek the approval from management and key stakeholders as part of organisation's enterprise risk framework.

Summary

- First publication this year on 15 July 2024. It was launched as part of PET Series at PDP Seminar 2024 & PET Summit 2024
- Agency for Science, Technology and Research (A*STAR) has been a key collaborator to co-develop the guidance. With PDPC
- Simplified synthetic data generation into 5 steps and provided best practices and key considerations for each step to reduce re-identification risks of synthetic data
- https://go.gov.sg/pdpc-sdg



Thank you

www.facebook.com/pdpcsg

www.linkedin.com/company/pdpcsg

www.instagram.com/pdpcsg

t.me/pdpcsg

www.pdpc.gov.sg

Scan the QR code
for more information.

pdpc
PERSONAL DATA
PROTECTION COMMISSION
SINGAPORE