GENERATING SYNTHETIC DATA FOR RESEARCH AND ANALYSIS

IMDA PET SANDBOX – KAJIMA CASE STUDY

INFOCOMM MEDIA DEVELOPMENT AUTHORITY

Contents

Use Case Background

1. As the building owner of the "The GEAR", Kajima collects personal data of building inhabitants through sensors (e.g. CCTVs) and visitor management systems for the purpose of monitoring, surveillance, and building clearance.
2. To use the collected data for purposes other than what it was collected for, individual consent[1] would be required. However, it may not be feasible for Kajima to seek every individual's consent, every time.
3. Therefore, Kajima is seeking a privacy-centric solution that could potentially enable collected data – especially if there is personal or commercially sensitive data within - to be used freely without the need for consent, such that it can be easily shared and used.
4. Kajima's initial use of the synthetically generated data is to share with researchers to analyse and measure the "connectedness" of certain zones within the building, with the objective to improve the well-being (e.g. through better services) for building inhabitants.

Use Case Details

5. **A proof of concept (POC) was conducted seeking to evaluate and understand:**
    a. The effectiveness of the synthetic data across 3 key metrics (e.g. utility, fidelity and privacy)
    b. PDPA obligations regarding the use of synthetic data (SD)

6. **Key POC stakeholders:**
    a. **Kajima** – Building owner seeking to leverage collected data for analysis and research purposes
    b. Betterdata – Provider of SD Generation solution

7. **The Betterdata solution includes:**
    a. Metadata profiler[2] to profile Kajima's raw datasets. Profiling occurs within Kajima's environment, only the final report of the profiling process is shared with Betterdata.

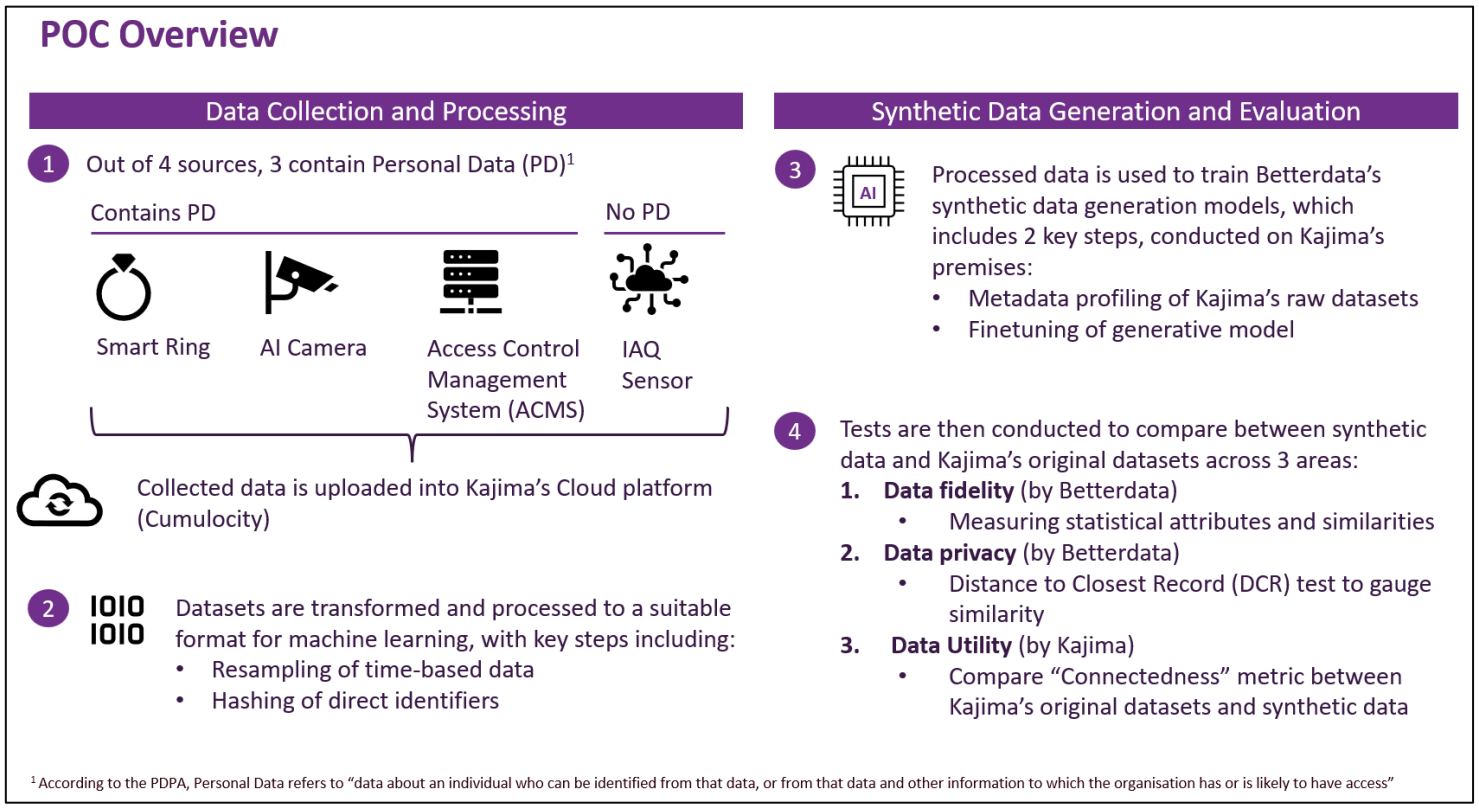
[1] Currently, Kajima seeks and obtains the requisite consent from individuals before using their data for purposes beyond what they were initially collected for

[2] The metadata was crucial in understanding the schema, data types, time resolutions, and intended usage of each dataset. It enabled Betterdata to prepare the data appropriately for training their generative models and to ensure the synthetic datasets preserve critical statistical and structural properties necessary for Kajima's research.

b. **Generative Model** trained on Kajima's real datasets and subsequently deployed at Kajima's premise and air-gapped environment to be further finetuned.

POC Overview and Steps

8. There are 2 key phases to the POC: **1) Data collection and processing and 2) Synthetic data generation and evaluation**



9. **Step 1** – Kajima currently collects data from multiple sources. *(See Annex 1 for metadata feature lists)* The POC includes the use of 4 datasets containing interactions between the building and its occupants:

a. **Access Control Management System (ACMS)** – Dataset contains static (non-time series), record-based personal information (e.g. email address, company info) regarding building occupants; leverages facial recognition camera to provide clearance to registered users.

b. **Facial Recognition Artificial Intelligence (FRAI)** – Dataset contains event-driven (irregular time series) facial recognition data capturing movement details of building occupants.
c. **Smart Ring** – Dataset contains regular time series health reports and regular heart rate data with variable frequencies depending on activity captured daily. Wearable device for building tenants that measures individual statistics, e.g. user heart rates, etc.
d. **Indoor Air Quality Sensor (IAQ)** – Dataset contains regular sequential time series environmental data, e.g. temperature and humidity; there is no personal data in IAQ datasets.

10. **Step 2** – Datasets are then transformed through an extensive process into a Wide Data Format necessary for the generative models to generate synthetic data; doing so also reduces feature dimensionality and enhances model performance.

    a. Selecting, normalising and concatenating a few datasets
    b. Filtering to specific zone names on selected specific hours
    c. Analysing the structure and characteristics of the raw data based on its metadata, i.e. data schema, time pattern and statistical distributions
    d. Resampling all time-based data to 5-minute intervals
    e. Hashing of direct identifier fields (i.e. email addresses)

11. **Step 3 and 4** – SD was assessed based on the following: Utility and Fidelity tests compared SD with Kajima's original datasets

    a. **Data Utility** – The test is evaluated based on the Average Centrality Metric (ACM) which quantifies how "well-connected" individuals are within a defined space and is computed using overlapping time spent by individuals within the same zone. A higher ACM value indicates denser interactions or stronger linkages between individuals. The objective of the test is to ensure that the metric on datasets achieve the same score, to ensure its similarity.
    b. **Data Fidelity** – The test evaluates statistical distribution, based on a series of metrics across model, features, distance, graphical evaluation and conditions dissimilarity. *(See Annex 2 for more information on metrics).* Lower metric values mean distribution to real data is closer and indicate higher fidelity, i.e. a more effective SD model.
    c. Data Privacy – To evaluate the privacy of SD, the Distance to Closest Record (DCR) – a test that measures the similarity between a synthetic to the nearest real record was conducted (A DCR of 0 indicates an exact match, presenting a high risk of re-identification. A lower DCR value generally suggest strong similarity to real data, an increased privacy risk). The DCR captures both exact

and near-exact similarities, offering a comprehensive view of re-identification risk.

Regulatory Learnings

12. Kajima sought Practical Guidance (Guidance) from the Personal Data Protection Commission (PDPC) on the following:

    a. Whether the business improvement exception under the PDPA can be relied on to generate SD without consent; and
    b. Whether the generated SD is considered personal data such that the Data Protection Provisions under the Personal Data Protection Act 2012 (PDPA) would apply.

PDPC's assessment

Whether the business improvement exception under the PDPA can be relied on to generate SD without consent

13. PDPC is of the view that the business improvement exception may apply where Kajima's use of personal data in training the SD generation model is for any of the following purposes (including where the SD generated contributes towards these purposes):

    a. For Kajima to improve or enhance its goods or services, or develop new goods and services (e.g., building consultancy services through insights generated from the use of SD)
    b. For Kajima to improve or enhance its methods or processes, or develop new methods or processes, for business operations (e.g., better building design and facilities management processes); and/or
    c. For Kajima to learn about or understand behaviour and preferences of individuals (e.g., better understanding of how building occupants interact with the surroundings/environment).

14. To rely on the exception, Kajima will need to ensure that the purpose cannot be reasonably achieved without using the personal data in an individually identifiable form, and that a reasonable person would consider the use of personal data for such purpose appropriate in the circumstances.

Whether the generated SD is considered personal data such that the Data Protection Provisions under the Personal Data Protection Act 2012 (PDPA) would apply

15. Personal data is defined in Section 2 of the PDPA to refer to data, whether true or not, about an individual who can be identified (a) from that data; or (b) from that data and other information to which the organization has or is likely to have access. While synthetic data is generally fictitious data, it is not inherently risk-free due to possible re-identification risks such as singling out attacks, linkability attacks and inference attacks.

16. In this POC, PDPC notes that the following safeguards have been implemented:

    a. **Pseudonymising of raw data.** Hashing of email IDs in the raw data is performed prior to generating the synthetic data. This reduces the risk of identifiers being used in model training and reproduced when generating synthetic data.
    b. **Using less granular data for model training.** In this POC, data records at 5-min intervals were used instead of data records collected seconds apart. This reduces the granularity of records used and lowers the likelihood of re-identification of individuals from the training data.
    c. Removing synthetic data records that are similar/nearly identical to the training data. This involves identifying and removing records below the 5% DCR (Distance to Closest Record) threshold in the final synthetic dataset.
    d. **Implementing controls to limit access to the synthetic data.** In this POC, the SD generated will only be shared in a controlled manner with selected researchers.

17. Taking into consideration the above, PDPC is of the view that the generated synthetic data **would generally not be considered personal data** if it has had data protection best practices incorporated both during and after its generation process, and it has been assessed that there is no serious possibility of re-identification. These best practices include sufficient safeguards being put in place to manage risks of re-identification of individuals, such as having contractual agreements to outline the responsibilities of the research organisation in respect of the synthetic data (e.g., safeguarding the synthetic data from unauthorised access and disclosure, to prohibit attempts to re-identify individuals), or conducting periodic reviews to assess risks of re-identification, especially if intended for public release. In addition, while DCR-based filtering can address singling out attacks, Kajima should consider implementing additional safeguards to lower the risks of linkability and inference attacks, to strengthen the overall privacy protection afforded to the generated synthetic data.

18. If there is a serious possibility that individuals can be re-identified from the SD, such SD would be considered personal data for the purposes of the PDPA.

19. Please note that PDPC's views above are confined to the context of the proposed POC. Organisations should seek further guidance from PDPC if they intend to use SD in other situations, e.g., for commercial purposes.

Results and Next Steps

20. Whilst the POC successfully validated the potential of SD to unlock secure, privacy-preserving access to Kajima's smart building datasets, a comparison of Kajima's raw datasets to the synthetically generated data showed that further refinement was required to preserve inter-variable relationships across datasets to improve the effectiveness and privacy of SD. Key areas include:

    a. **Training of SD Generation model should support multiple types of data inputs (e.g. relational time-series, event-driven data generation) in their original formats, without need for significant data transformation[3]:** Preserving the original schema of Kajima's datasets would ensure that inter-table dependencies can be captured without losing granularity.
    b. **Hyperparameters (i.e. "settings" defined before actual model training) need to be dynamically tuned to ensure potential changes across the use case are accurately accounted for:** A key limitation of the POC was untuned hyperparameters, which resulted in inter-variable relationships not being preserved. Enabling automated intelligent tuning would be a key measure to address that.
    c. **Number of unique IDs between original and synthetic datasets should be the same:** To ensure that interaction patterns in resultant synthetic data are not diluted, checks should be undertaken to ensure no. of unique IDs in SDG and original datasets are similar. A difference in number (e.g. 10 vs. 100) would misrepresent interactions, resulting in differing ACM and fidelity values, lowering the effectiveness of SD as values should be similar.
    d. **Enhance data privacy evaluations of generated SD with additional innovative techniques:** In addition to the DCR evaluation, Betterdata would subsequently conduct a differential privacy (DP) audit of the SD – a statistical test that empirically estimates the level of DP parameter ε. The evaluation involves the measurement of the average distance between a tiny set of randomly generated audit examples, "canaries", added to the training set of

[3] In the POC, Betterdata introduced a Wide Data Format, which entailed significant revision to the metadata and the original schema of each dataset, so that all 4 datasets could be captured in a single table

Whilst the format was designed to reduce feature dimensionality to enhance model performance while capturing implicit relationships, the **key trade-off was the loss of the original multi-table structure, as the data was consolidated into an interval time-series format containing inherently event-driven data.**

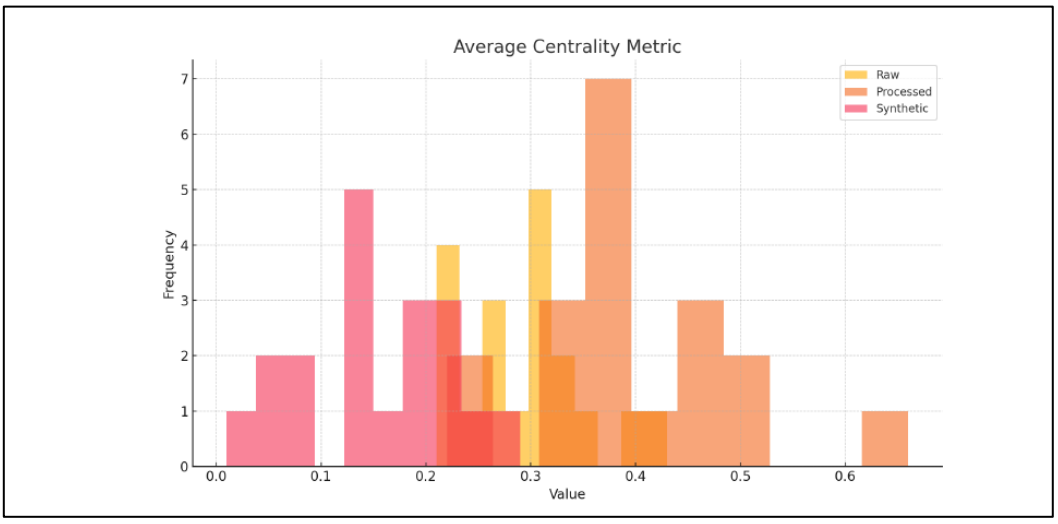
real records to their nearest synthetic records as a proxy measure of training-data memorisation. A smaller canary-to-synthetic data distance, i.e. a larger ε value, would indicate a higher risk of re-identification for the real records.

21. Tests conducted across utility, fidelity and privacy yielded the following results

Evaluation Dimension	Key Metric / Observation	Implication
Statistical fidelity	Contextual-FID (C-FID) = 0.0498; good alignment in t-SNE plots	High similarity between synthetic and real data; structurally realistic generation achieved
Utility	Lower AVM in synthetic data due to email ID expansion	Interaction patterns and co-location signals diluted; requires refinement for downstream use cases
Privacy	~50 records with DCR ≈ 0; remainder > 5% distance	Most records meet privacy expectations; filtering recommended before external sharing

Summary of key findings from the POC

a. **Data Utility**
   i. Comparison of the ACM metric between Kajima's raw, processed and synthetically generated data were observed to have varying differences, with the difference greatest between processed and SD. There were two issues observed:
      A. Transformation of the original datasets during initial processing had impacted the relationship between variables, changing inter-variable dependencies, causing original and processed datasets to have different values.
      B. ACM for processed and resultant SDG were also different due to untuned hyperparameters in generative models.
   ii. Given the results, more refinement on how data is to be transformed and enabling the models to support more data formats may be required, to ensure a higher likelihood that synthetic data can match ACM patterns in real data, and that intervariable dependencies may be preserved

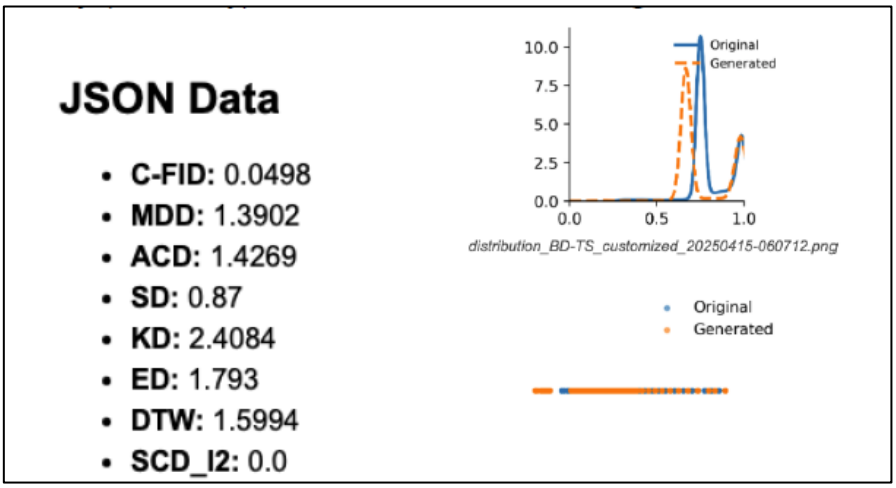


Histogram of ACM values for raw, processed and synthetic datasets

22. **Data fidelity**

a. This was evaluated based on 8 metrics (C-FID, MDD, ACD, SD, KD, ED, DTW, SCD_l2) where values closer to 0 would indicate high similarity between the real and synthetic datasets.

b. The metric values listed under JSON Data displayed strong alignment with values close to 0. The t-SNE plot (on the right) visually demonstrates that the real and synthetic data points form overlapping clusters, indicating strong structural similarity and high data fidelity, affirming the closeness of the two datasets.



Data Fidelity evaluation values

23. **Data Privacy**

a. The DCR metric calculates the minimum distance between each synthetic and real record, where a **DCR of 0** indicates an exact match, which presents a high risk of re-identification. Lower DCR values in general would suggest stronger similarity to real data, and therefore increased privacy risk.
b. For the POC, **approximately 1.4% of synthetic records had DCR values close to 0**, indicating they were either identical or nearly identical to real records. An option to filter out these synthetic records with DCR values close to 0 before any data sharing was provided.
c. However, the average DCR of synthetic records from their real counterparts was about 5% of the maximum range, which indicates a reasonable level of privacy protection for practical use.

While the synthetic data is generally fictitious and may not be considered personal data on its own, it is not inherently risk-free. It presents potential re-identification risks when combined with other datasets, due to its design goal of closely representing the original data patterns.

The risk increases when multiple data attributes are released together, as the intersection of these features increases the likelihood of unique pattern identification. As such, data minimisation and outlier treatment are crucial steps in reducing such risks, particularly for granular data involving detailed demographics or high-frequency measurements.

Annex 1 – Data Features used in POC

ACMS | Occupant info	**FRAI** | Movement events	**IAQ** | Indoor environment	**Smart Ring** | Health metrics
4 features	14 features	22 features	19 features
emailAddress	time	time	day
company	type	source	score
department	email	ktg_iaq.H.unit	temperature_deviation
userType	accessPointType	ktg_iaq.H.value	temperature_trend_deviation
	source	ktg_iaq.P.unit	participant_id
	age	ktg_iaq.P.value	vascular_age
	gender	ktg_iaq.T.unit	level
	mood	ktg_iaq.T.value	day_summary
	energy	ktg_iaq.Temp.unit	recovery_high
	stress	ktg_iaq.Temp.value	stress_high
	wellbeing	ktg_iaq.co2.unit	activity_balance
	engagement	ktg_iaq.co2.value	body_temperature
	watchlistName	ktg_iaq.lux.unit	hrv_balance
	score	ktg_iaq.lux.value	previous_day_activity
		ktg_iaq.pm25.unit	previous_night
		ktg_iaq.pm25.value	recovery_index
		ktg_iaq.score.unit	resting_heart_rate
		ktg_iaq.score.value	sleep_balance
		ktg_iaq.spl_a.unit	average
		ktg_iaq.spl_a.value	
		ktg_iaq.voc.unit	
		ktg_iaq.voc.value	

Annex 2 - Data Fidelity Metrics

1. **Model-based (the smaller the value, the better):**
   a. **Contextual-FID (C-FID):** Quantifies how well the synthetic time series conforms to the local context of the real time series.

2. **Feature-based (the smaller the value, the better):**
   a. **Marginal Distribution Difference (MDD):** Assesses how closely the distributions of the original and generated series align.
   b. **Auto-correlation Difference (ACD):** Assesses how well dependencies are maintained in the generated time series.
   c. **Skewness Difference (SD):** Evaluates distribution asymmetry, which is vital for the marginal distribution of time series datasets.
   d. **Kurtosis Difference (KD):** Assesses the tail behaviour of a distribution, revealing extreme deviations from the mean.

3. **Distance-based (the smaller the value, the better):**
   a. **Euclidean Distance (ED):** Deterministically assesses the similarity between generated and real data.
   b. **Dynamic Time Warping (DTW):** Captures the optimal alignment between series regardless of their pace or timing.

4. **Graphical evaluation metrics:**
   a. **Distribution plot:** Visualizes the distribution of generated time series compared to the original one within a two-dimensional space.
   b. **t-distributed Stochastic Neighbor Embedding (t-SNE):** Illuminates the difference between the input and generated time series in terms of density, spread, and central tendency to show how the generated time series closely mirrors the original's statistics.

5. **Conditions dissimilarity (the smaller the value, the better):**

L2 dissimilarity: Measures the difference between the original and synthetic conditioning variables using L2 norm. Value ranges from 0 to 1.