pdpc
PERSONAL DATA
PROTECTION COMMISSION
SINGAPORE

PRACTICAL GUIDANCE TO KAJIMA

Background

1. Kajima Development Pte Ltd is the regional property development arm of Kajima Corporation, involved in construction, engineering, and real estate business. In collaboration with its research and development arm at Kajima Corporation, Kajima Development Pte Ltd (referred to collectively as “Kajima”) as the building owner of the Kajima Lab for Global Engineering, Architecture & Real Estate (also known as The GEAR building), Kajima collects data from various sources, including personal data of building occupants through sensors and visitor management systems for the purpose of monitoring, surveillance and building clearance.

2. Kajima is implementing a Proof of Concept (“POC”) to generate synthetic data using datasets that capture interactions between the building/environment and the occupants. In this POC, Kajima has engaged Betterdata as the provider of the synthetic data generation model. The generated synthetic data will be used for conducting analysis and research to optimise services that can improve well-being of building occupants. Kajima also intends to share the generated synthetic data with researchers for this purpose.

3. The data sources involved in this POC are as follows:

a. Access Control Management System (ACMS) – Dataset contains static (non-time series), record-based personal information (e.g. email address, company information) regarding building occupants. It uses 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. This is a wearable device for building tenants that measures individual statistics (e.g. user heart rates).

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.

4. A brief description of the key steps involved in the POC is as follows:

   a. **Data collection and pre-processing.** Data collected from the four sources will have their email addresses hashed, and matched based on the hashed email addresses. Without gaining access to the raw data, Betterdata analyses the metadata (i.e., data schema, time pattern and statistical distributions) to identify data pre-processing requirements. Kajima transforms the hashed data into a format that is suitable for machine learning. This processing takes place in Kajima's premises with access controls.

   b. **Model training.** The processed data is used to train Betterdata's synthetic data generation model. All training activities take place on-premise within Kajima's environment[1].

   c. **Generation of synthetic data.** Synthetic data is generated using the trained model, also performed within Kajima's environment.

   d. **Evaluation of synthetic data.** Kajima assesses the generated synthetic data on three fronts, namely (i) data privacy; (ii) data fidelity; and (iii) data utility. Of relevance to this document is on data privacy, where the Distance to Closest Record (DCR)[2] is used to measure the similarity between a synthetic data record to the nearest real record (also known as training/source data).

5. 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 synthetic data without consent; and

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

PDPC's assessment

[1] The computer used during the POC is located in a secured area within Kajima's premises, and access is strictly limited to Kajima personnel. Betterdata staff were only allowed to enter the area under the supervision and escort of Kajima members, and no remote access was permitted at any time. In addition, all work performed by Betterdata during the POC was conducted under the direct supervision of Kajima staff. The computer itself is protected by a login password known only to authorized Kajima members.

[2] A DCR value of 0 indicates an exact match, presenting a high risk of re-identification. A lower DCR value generally suggests strong similarity to real data, which indicates an increased privacy risk.

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

6. PDPC is of the view that the business improvement exception may apply where Kajima's use of personal data in training the synthetic data generation model is for any of the following purposes (including where the synthetic data 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 synthetic data);

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).

7. 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 synthetic data is considered personal data such that the Data Protection Provisions under the Personal Data Protection Act 2012 (PDPA) would apply

8. 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.

9. 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 synthetic data generated will only be shared in a controlled manner with selected researchers.

10. 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[3] 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 is a method to handle 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.

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

12. 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 synthetic data in other situations, e.g., for commercial purposes.

END OF DOCUMENT

[3] Refer to safeguards and best practices in Table 8 of PDPC's Proposed Guide on Synthetic Data Generation.