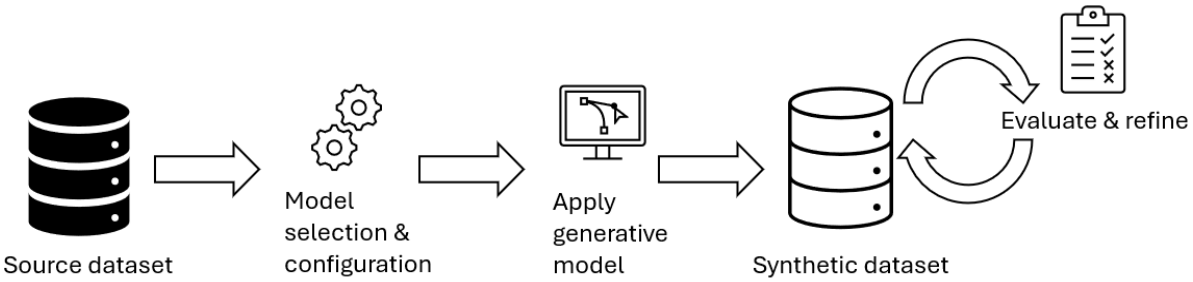
APPENDIX 1: About synthetic data

To develop synthetic data at scale, but without any one-to-one mapping, requires:

- A 'source dataset', containing 'real' records
- A generative model, and
- A new table or file, in which to house the synthetic records created.

These three elements are illustrated in the following diagram.[20]



These three elements can be further described as follows.

A source dataset

This is the original dataset containing records about a group of 'real' individuals, such as customers, patients or students. The synthetic data to be generated will be expected to emulate the statistical properties of this source data.

Some source datasets will include data fields that clearly and directly identify individuals, from direct identifiers such as names, date of birth and unique numbers. Others might include indirect or 'quasi' identifiers such as age, gender and postcode, which in combination can render some individuals unique in the dataset (i.e. distinguishable from the rest of the group) and thus 'identifiable' in law.

Others again might have already robustly controlled for both direct and indirect identifiers via de-identification techniques, but the 'attribute' data is itself rich enough that some individuals will be unique in the dataset, and thus 'identifiable' in law. For example, even without any direct or indirect identifiers about a patient, the parts of a patient record that record event dates (such as doctor visit, hospital admission, surgery) and clinical information (such as conditions or treatment) can themselves render a patient unique in the dataset.

[20] Diagram adapted from UTS, "Synthetic Lumos - Technical Transfer", presentation slides, 20 December 2021

In all such cases, the source dataset contains 'personal information' as defined in privacy law.

To demonstrate using an example, in the source system of hospital admissions dataset, you might have 50% male and 50% female patients. You might also know from the source dataset that 5% of patients have had a caesarean section. If you break down that figure by gender, it represents 10% of female patients and 0% of male patients.

A generative model

This is the statistical model used to generate the synthetic data. It is derived from the source dataset. Statistical tables are created from the source dataset to generate a probabilistic model.

> "A generative model is able to 'learn' the statistical properties of the source data without making strong assumptions about the underlying distributions of variables and correlations among them". [21]

The statistical properties of the source data might include, for example, the distribution of patients across gender, age ranges, geographic regions, as well as reason for hospital admission, but *without* necessarily correlating one data field (such as gender) to another (such as reason for hospital admission).

In the generative model for a hospital admissions dataset, the model will show that 50% of patients are male and 50% are female. Separately, it will show that 5% of all patients were admitted to hospital for a caesarean section.

The synthetic data

This is the data generated from the generative model: many thousands of individual records of 'fake' individuals.

Because the data about individual synthetic 'patients' is generated by the generative model, unless correlations were designed into the generative model, attributes will be distributed across the synthetic patient records according to the statistical properties of the source dataset.

In the synthetic version of our hypothetical hospital admissions dataset, 5% of male patients will show a caesarean section as the reason for hospital admission, as will 5% of female patients.

This hypothetical scenario is intended to illustrate the 'privacy protection vs. data utility' trade-off that can occur when creating synthetic data. Privacy protection is strengthened by

[21] Office of the Privacy Commissioner of Canada, "When what is old is new again – The reality of synthetic data", OPC blog post, 12 October 2022; available at https://priv.gc.ca/en/blog/20221012

not mapping correlations, however more simplistic representations of populations in synthetic datasets may then limit the quality or accuracy of results. In this scenario, if the correlation between “gender” and “reason for hospital admission” is not maintained, then the synthetic dataset will indicate a percentage of male patients who have had caesarean sections. Whether this will impact a project’s outcomes will depend on the use case and the purpose for the analysis. For example, if the use case is to create dataset to train nursing students to use a patient management system, the synthetic dataset may be suitable for this purpose. However, the synthetic dataset would not be suitable for a use case that involves testing or researching a clinical hypothesis. The more that complex correlations are maintained from the source dataset (i.e. in order to make the synthetic dataset ‘more real’), then the greater the likelihood of replicating unique patients from the source dataset, which leads to increased re-identification and privacy risk.