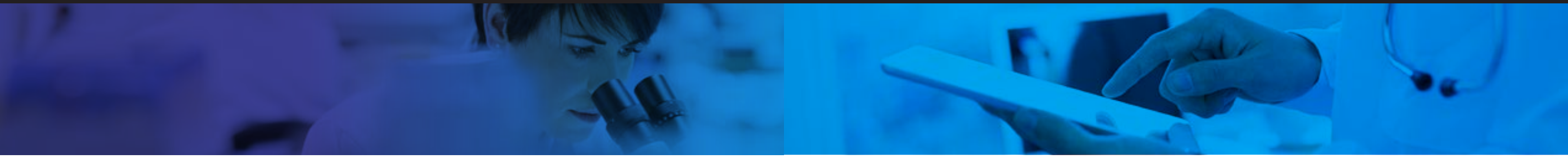
Delivering on the Promise of Synthetic Data



Replica Analytics develops unique technologies for generating privacy protective synthetic data that maintains the statistical properties of real data. We enable fast and effective access to high utility data while meeting regulatory obligations.

Methods employed by Replica Analytics combine machine learning tools to generate synthetic data, and privacy assurance on synthetic data to ensure that the probability of identifying individuals is very small.

Core Synthetic Data Software and Services

Replica Analytics provides software for the generation of synthetic data based on real datasets, and for performing privacy assurance on existing synthetic datasets. We also provides both of these capabilities as a service for clients who want to outsource data synthesis and privacy assurance.

With deep expertise in privacy and data analysis, Replica Analytics is a trusted partner for data synthesis.

Built for Handling Complex Data

Responsive, Reliable and Trustworthy Synthetic Data

Enabling AI Innovation through Rapid Data Access

WWW.REPLICA-ANALYTICS.COM

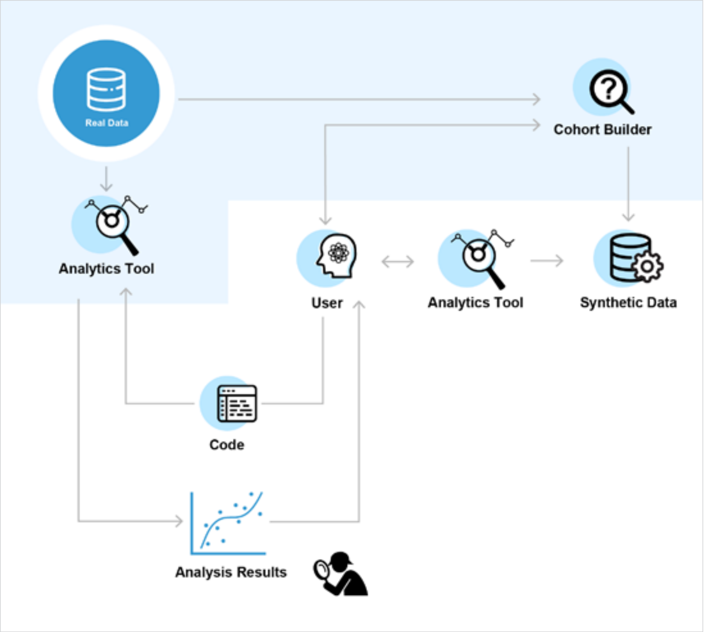
Replica Analytics

Delivering on the Promise of Synthetic Data

Replica Software

Replica Synthesis

Replica Synthesis software ingests real data, and builds data synthesis models to generate high utility synthetic datasets.



Key features of Replica Synthesis include:

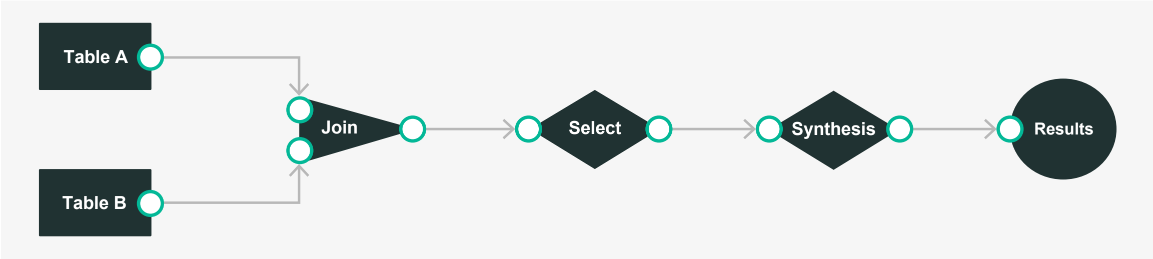
- Synthetic Cohort Builder to query and integrate data from multiple data sources and generate synthetic variants of these cohorts.
- Validation Server to re-run analytics code on the original data.
- Produces a synthesis report which describes the data, methodology, the synthesis results, the utility results, and any limitations. The report template can be easily customized.
- Customizable synthesis parameters.
- Can be deployed on the cloud or on-premises. Clients do not have to share their data to synthesize it.
- Comprehensive REST API for integration with multiple and varied front ends.
- Generates detailed data utility results.
- The software is built for small and large datasets.
- SDKs supporting multiple data science and software engineering end-users.
- Flexible synthesis plan specifications to handle complex datasets.

Replica Software

Replica Synthesis

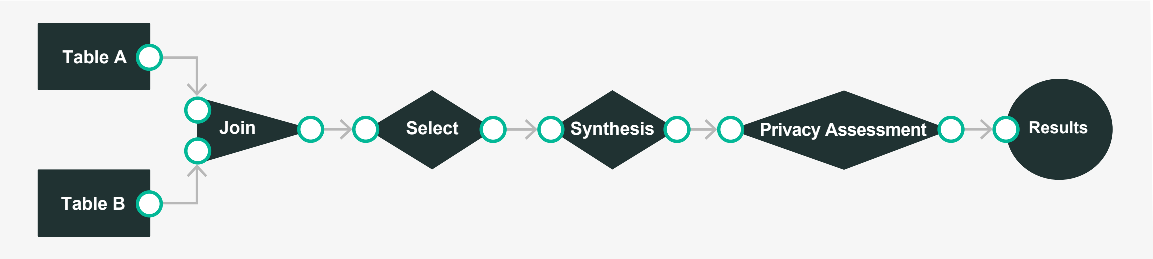
Data Synthesis

The basic functionality of Replica Synthesis allows the user to define a cohort from the source data and return a synthetic version of that cohort. A report summarizing the synthesis and the utility assessment is sent directly to the user.



Privacy Assurance

This workflow combines data synthesis and privacy assurance capabilities. The identity disclosure (privacy) risk is computed directly on the synthesized data when it is generated and the privacy assurance results are sent directly to the user.



Replica Analytics

Delivering on the Promise of Synthetic Data

Replica Software

Replica Synthesis

Unique solution to evaluate the identity disclosure risk of synthetic data. If you are generating or using synthetic data, your legal team will want assurances that this synthetic data is not personal information. Replica Synthesis implements a complete assessment methodology for this type of risk specifically designed for synthetic data.

"One of the key capabilities of Replica Analytics is the ability to quantitatively assess the privacy risks for a synthetic data set. Based on many years of experience in this area, we have developed a model and technology to convincingly measure the likelihood of a synthetic record being matched to a real person."

Khaled El Emam, Co-Founder, Replica Analytics

Key features of privacy assurance include:

- Empirically measures the likelihood that synthetic records can be matched to real individuals.
- Produces a detailed report describing the risk assurance methodology as well as the results of the risk assessment.
- Employs robust risk measurement models to evaluate the probability of matching synthetic records to real people under different types of attack.
- The privacy assurance models also account for multiple types of privacy risks simultaneously.

Given the risks of not processing personal information properly under various legal regimes, privacy assurance gives you the compliance evidence needed.

Replica Analytics

Delivering on the Promise of Synthetic Data

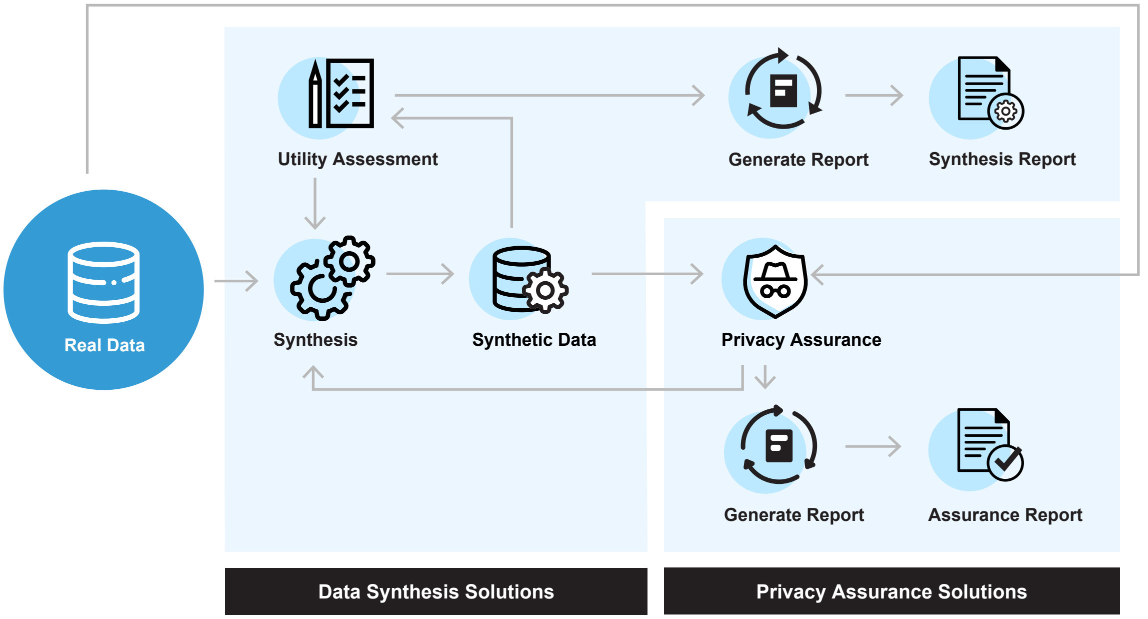
Replica Services

DATA SYNTHESIS SERVICES

- Getting a real dataset from a client and using our machine learning tools and algorithms to generate a synthetic dataset for them
- The process is largely automated – these synthetic datasets can be generated quickly and at scale
- The outputs are the synthetic dataset with a report documenting the utility characteristics of the data
- Utility assessement takes into account all possible models that can be built from the data – provides a comprehensive assessement of utility under multiple realistic conditions

PRIVACY ASSURANCE SERVICES

- Replica Analytics has the only complete privacy assurance model and process on the market today
- Either Replica Analytics or another entity (e.g., the client themselves) has generated synthetic datasets
- Replica Analytics receives the real and synthetic data and using their proprietary privacy assurance model, evaluates the privacy risks in the synthetic data
- The output is a privacy assurance report



Replica
Analytics

Delivering on the Promise of Synthetic Data

“ What Users Are Saying ”

“Synthetic clinical trial data from Replica Analytics enabled our teams to solve innovation challenges successfully. Realistic data and privacy assurance allowed us to make information available with minimal friction to access and sharing.”

Rebecca Li, Executive Director, Vivli

“Realistic synthetic data from Replica Analytics allowed us to demonstrate our data integration and analytics software to a prospective client quickly. What was going to be a months-long process to deal with the data privacy concerns and regulations happened in only days. Synthetic data helped us prove the business case and accelerate customer acquisition significantly.”

Steve Dischinger, VP of Business Development and Alliances, Cambridge Semantics

“Synthetic clinical trial data generated by Replica Analytics is being used across multiple business lines, such as data science and statistical programming. This has allowed easy and broader access to data for our internal teams and partners, and enables us to accelerate our innovation agenda.”

Head of Transparency, Global Pharmaceutical Company

Delivering on the Promise of Synthetic Data

What is Synthetic Data?

Synthetic data is generated from real data. Replica Analytics takes a real dataset and models its statistical distributions (e.g., is a variable in the data set bell-shaped?) and structure (such as the relationships among the variables). Using that model, records are generated that make up the synthetic dataset. Therefore, the values in the synthetic data are generated from the model and closely replicate many of the statistical properties of the original dataset.

If we are modelling the structure of the original data very accurately then the end result would effectively be replicating the original data, which would have negative privacy implications. Therefore, the model captures only the partial structure. There is a balancing act between how accurate the model needs to be and how close the synthetic data is to the original data. It is the classic trade-off between data utility and data privacy.

When done well the synthetic data retains sufficient statistical properties of the original data and has a very low risk of identifying individuals. This allows synthetic data to be shared more freely with minimal administrative and technical controls, and to be treated as non-identifiable data.

Why Consider Synthetic Data?

There is more than one legal way for sharing data for secondary purposes. For example, by obtaining the consent of the individuals, or by de-identifying the data. Consent can be impractical and there is strong evidence of consent bias. De-identification is a good solution. However, contemporary risk-based de-identification methods require additional administrative and technical controls to be in place, including the data recipient signing a data-sharing agreement. Sometimes implementing controls is not practical (the obvious examples being technology evaluation, competitions, and open data), or they are expensive to put in place.

Data synthesis enables the sharing of that data without having to sign a data-sharing agreement, which can be time-consuming in practice and does not scale well when there are a large number of data users. Implementing controls for processing synthetic data is either not necessary or can be limited. And the Replica Analytics synthetic data will work better in certain situations, such as when dealing with small data sets. Also, the generation of synthetic data can be quite efficient in practice.

Delivering on the Promise of Synthetic Data

What Use is Synthetic Data?

There are a number of cases where synthetic data provides an ideal solution. These include:

- **AI and data science projects.**
  One of the biggest challenges when developing AI and machine learning algorithms is getting sufficient realistic datasets to train the models and test on. Our Synthetic Cohort Builder can make it easier to get rapid data access and to augment data.

- **Proof of concept and technology evaluations.**
  Oftentimes technology developers or acquirers need to quickly evaluate whether a new technology works well in practice and they need realistic data to work with and have minimal constraints on access to that data.

- **Data exploration.**
  Organizations that want to maximize the use of their data can make synthetic versions available for exploration and assessment by potential users. If the exploration yields positive results, the users can validate the results on the real data through the Validation Server in the Replica Synthesis software.

- **Algorithm development.**
  Data analysis programs can be developed on synthetic data and then submitted to the data custodian for execution on the real data – this brings the verified code to the data rather than sharing the data itself.

- **Software testing.**
  Testing data-driven applications require realistic data for functional and performance testing. Random data cannot replicate what will happen when a system goes into production.

- **Open data.**
  Sharing complex datasets publicly is challenging because of privacy concerns. This can now be achieved by sharing synthetic data instead.

- **Hackathons and data competitions/challenges.**
  These require datasets that can be distributed widely with minimal burden on the entrants. The datasets have to be realistic to enable meaningful innovation.

With the innovative data synthesis methods being developed by Replica Analytics, the number of use cases where synthetic data is a good proxy for real data are increasing over time.

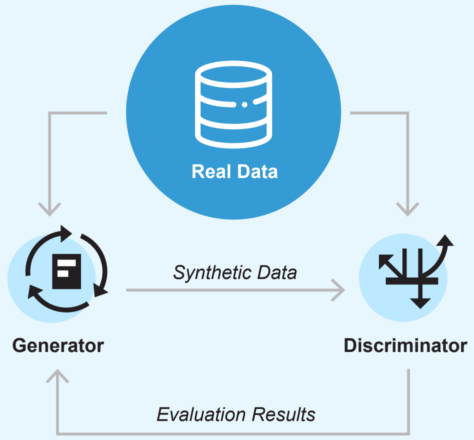
Delivering on the Promise of Synthetic Data

How is Synthetic Data Generated?

The main approach for generating synthetic data is to have a generator model and a discriminator function.

The generator takes real datasets, and using various statistical machine learning and deep learning models, generates synthetic data.

The discriminator evaluates how good that synthetic data is compared to the real data. If the differences between them are large then this information is fed back and a new generator model is trained to try to narrow these differences. Therefore, it is an iterative process of training generator models until they produce acceptable synthetic data.



The highly automated Replica Analytics pipeline will produce the datasets as well as comprehensive utility assessments for the data. This helps with creating buy-in from data analysts and other data users. Privacy Assurance Services from Replica Analytics complement the main synthesis technology to provide the necessary evidence to the legal team.

For further information:

Replica Analytics

replica-analytics.com
info@replica-analytics.com

Canadian Office

251 Laurier Avenue West, Suite 900,
Ottawa, Ontario K1P 5J6, Canada