Synthetic Data as a Privacy Enhancing Technology

Lucy Mosquera

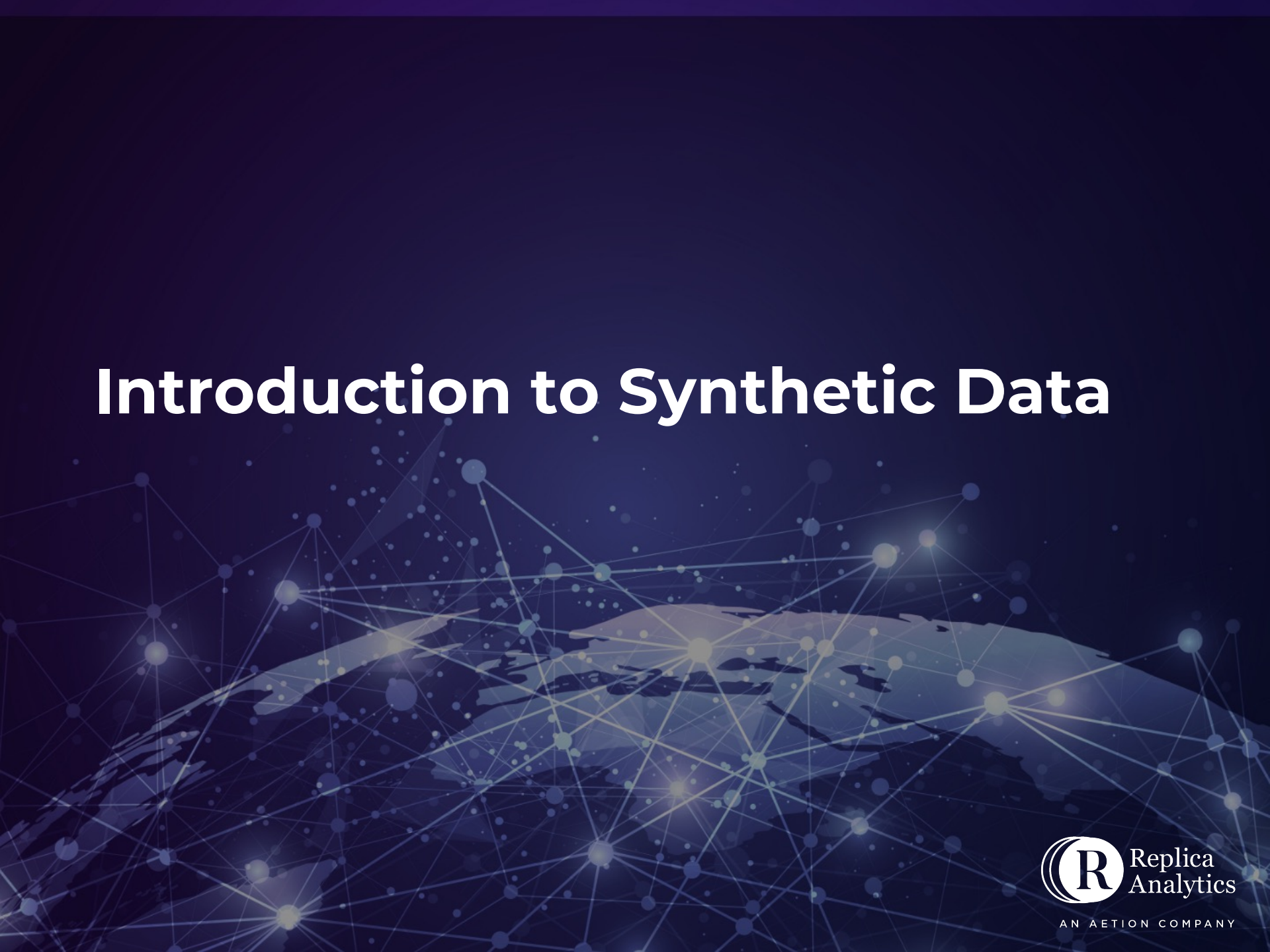
Director of Data Science, Replica Analytics

Agenda

- Introduction to synthetic data generation and use cases
- How to assess synthetic data in terms of utility and privacy
- Data synthesis as a privacy enhancing technology

Replica Analytics
AN AETION COMPANY

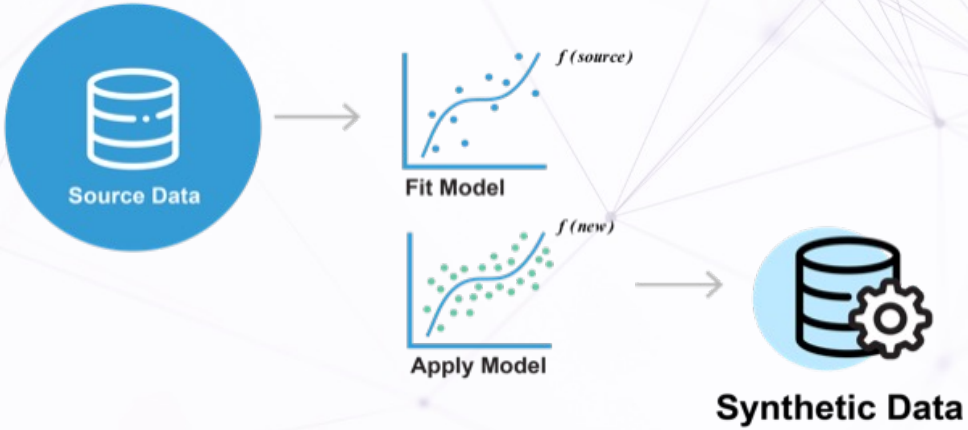
Introduction to Synthetic Data



Replica Analytics

AN AETION COMPANY

The Synthesis Process



COU1A	AGECAT	AGELE70	WHITE	MALE	BMI
United States	2	1	1	1	33.75155
United States	2	1	1	0	39.24707
United States	1	1	1	0	26.5625
United States	4	1	1	1	40.58273
United States	5	0	0	1	24.42046
United States	5	0	1	0	19.07124
United States	3	1	1	1	26.04938
United States	4	1	1	1	25.46939

Replica Analytics

AN AETION COMPANY

Use Cases for Synthetic Data

Discover Artificial Intelligence

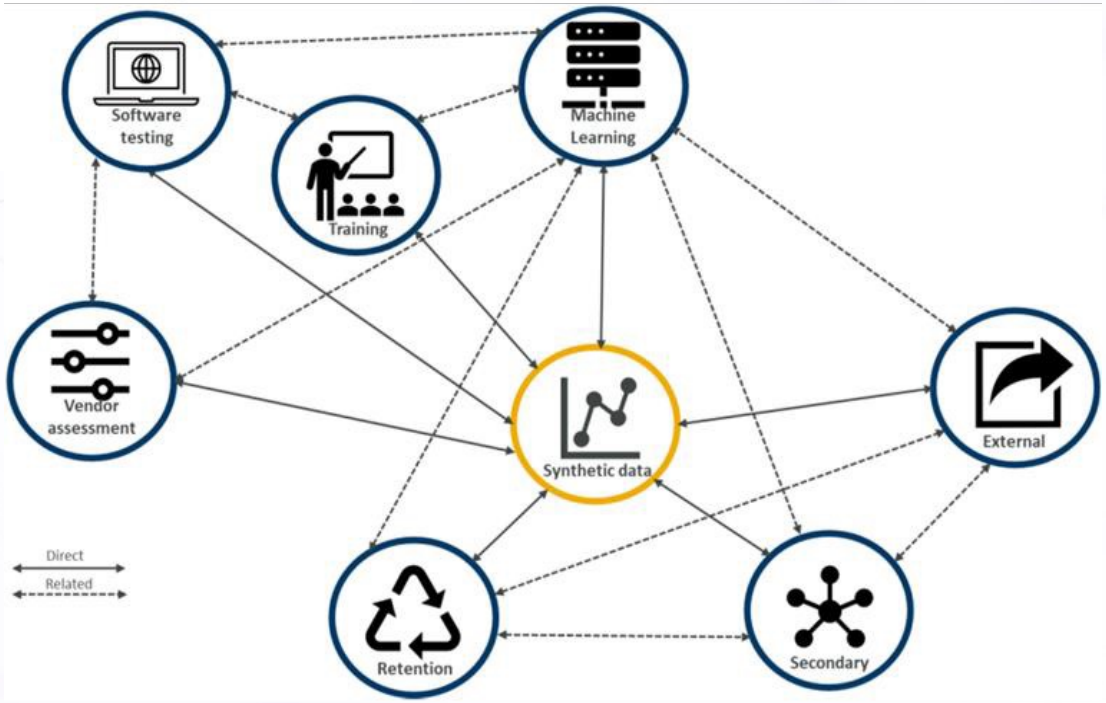
Discover

Review

Synthetic data use: exploring use cases to optimise data utility

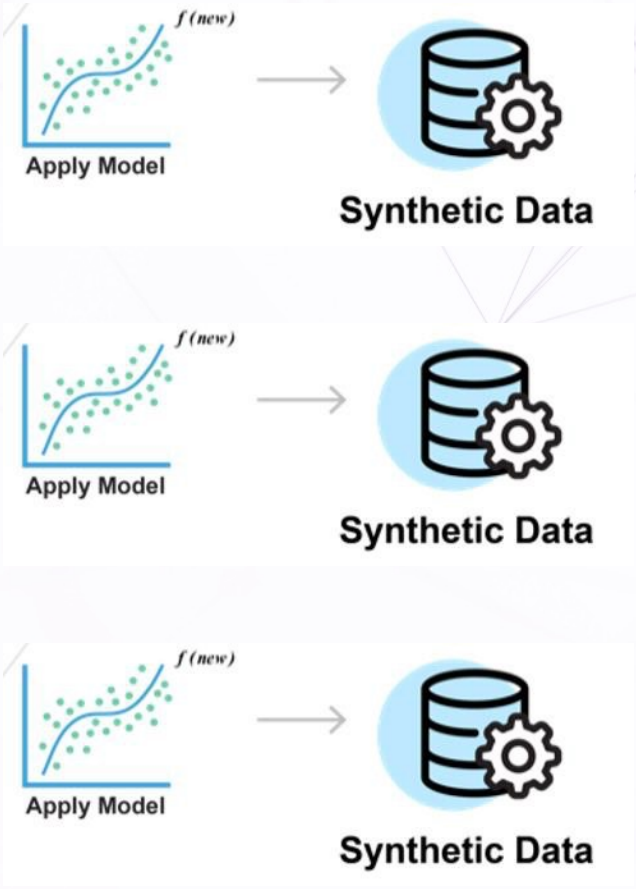
Stefanie James[1] · Chris Harbron[2] · Janice Branson[3] · Mimmi Sundler[4]

Received: 12 November 2021 / Accepted: 7 December 2021
Published online: 13 December 2021

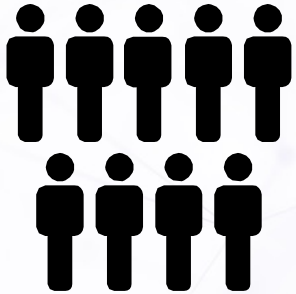


Replica Analytics
AN AETION COMPANY

A simulator exchange allows synthetic data to be made available without sharing actual data



Data Consumers

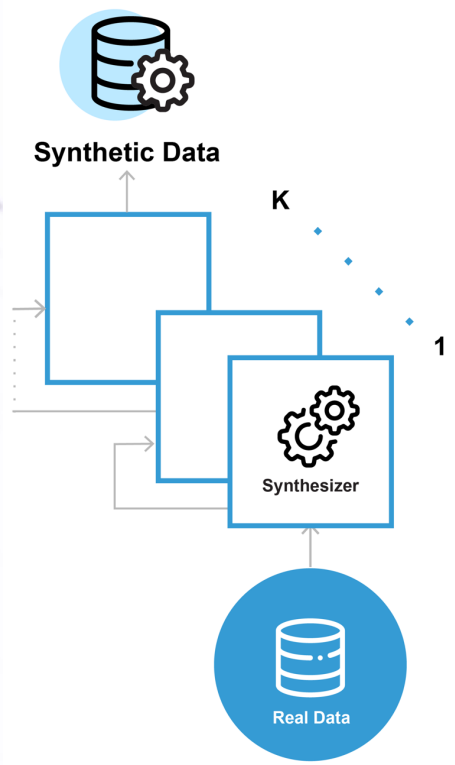


Replica Analytics

AN AETION COMPANY

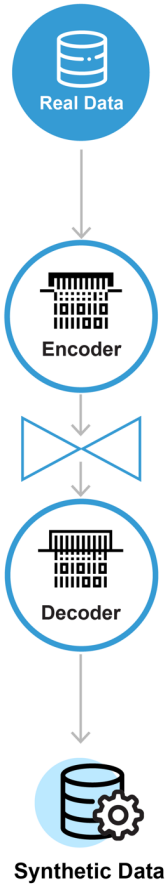
Sequential Synthesis

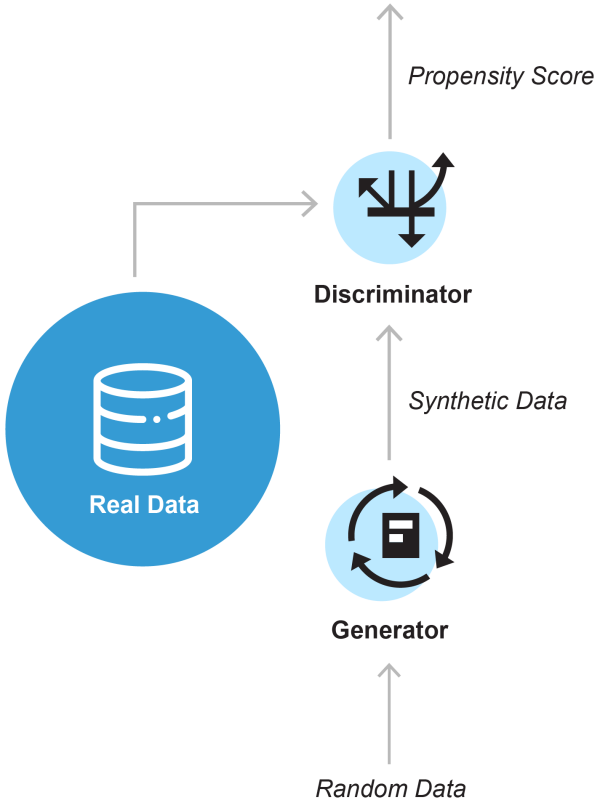
- Iterative synthesis model
- Can use different types of models for each variable in the dataset
- Model fitting can be parallelized



Replica Analytics
AN AETION COMPANY

Deep Learning Synthesis Models

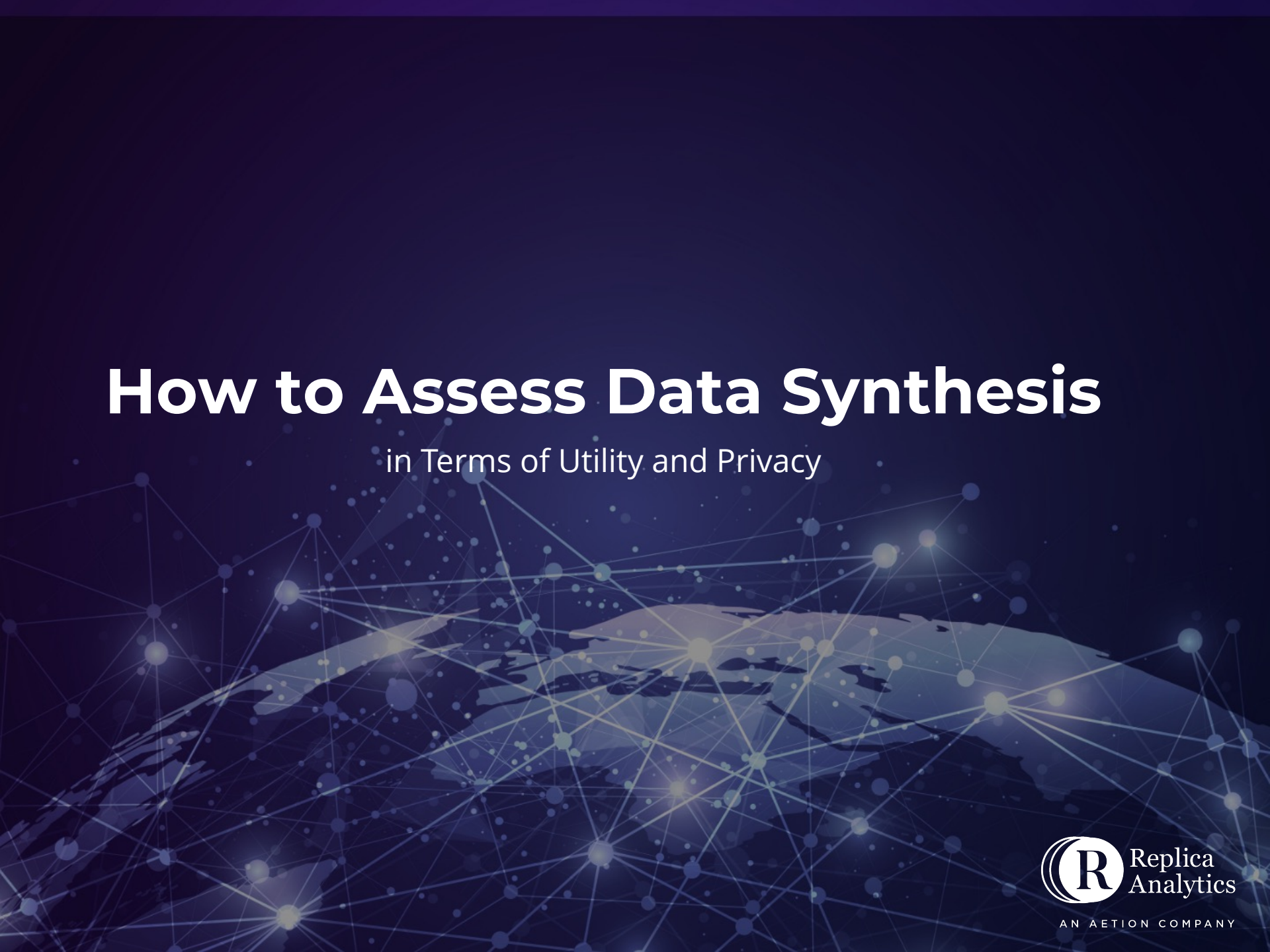




Replica Analytics
AN AETION COMPANY

How to Assess Data Synthesis

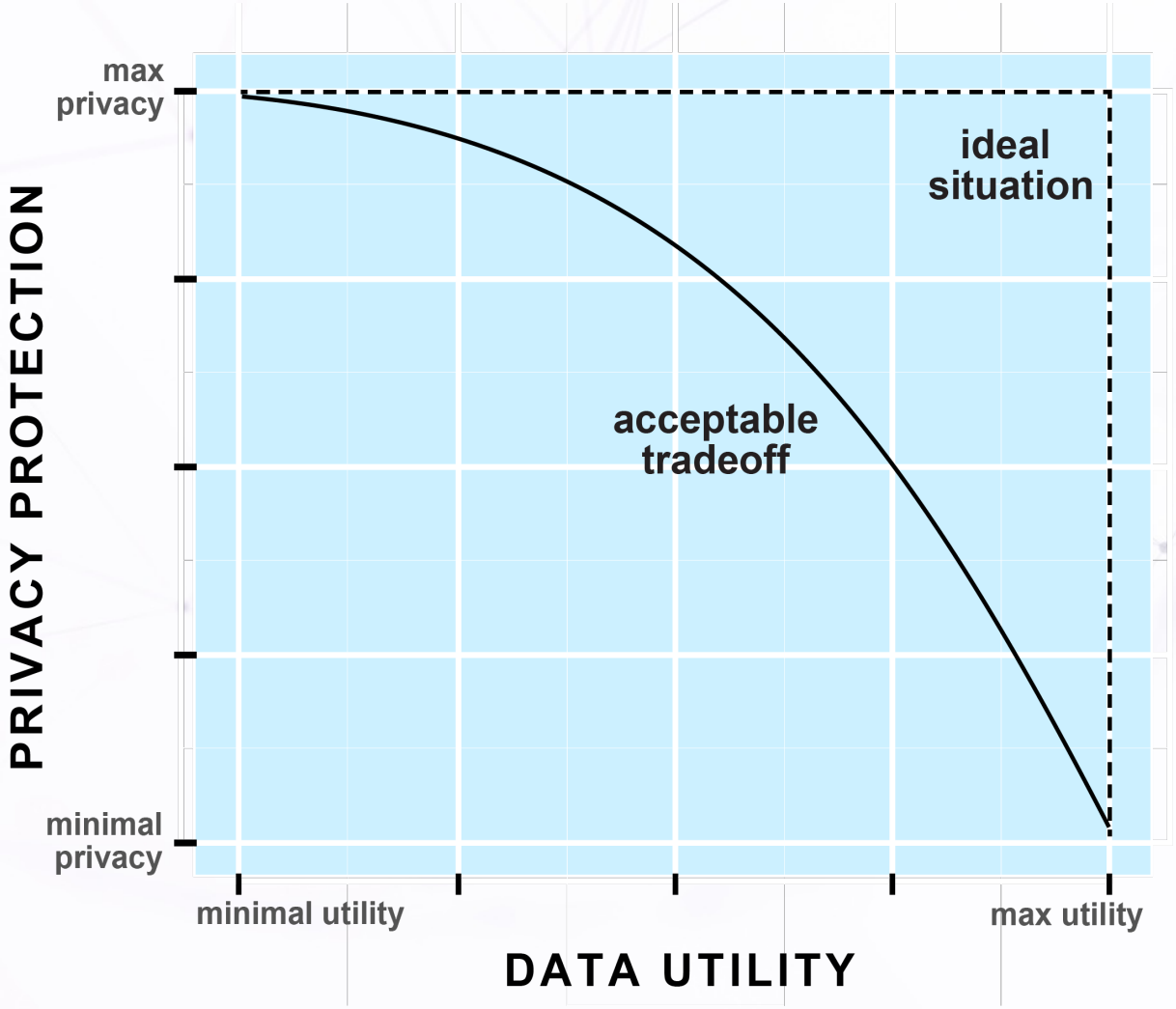
in Terms of Utility and Privacy



Replica Analytics

AN AETION COMPANY

Privacy- Utility Trade-off



Replica Analytics
AN AETION COMPANY

Utility Assessment Strategies

Broad Metrics

These are generic metrics that are easy to calculate when the generative model is built and synthetic data are synthesized.

They are only useful if they are predictive of workload-specific metrics.

Narrow Metrics

These are workload-specific and are what is of most interest to the data users. However, all the possible workloads will not be known in advance and therefore we have to consider representative workloads when developing and evaluating utility metrics.

Replica Analytics

AN AETION COMPANY

Privacy Risk Assessment Strategies

In Canadian law, **identity disclosure** is the main risk associated with de-identified data

Reidentification risk is the probability of being able to correctly match a record in a microdata sample to a real person

Since the individuals in synthetic data are not real, the privacy implications are different than with real data, they **require different strategies to assess risk**

Replica Analytics
AN AETION COMPANY

What's Different About Synthetic Data?

- In synthetic data we can classify each record as:
  1. Duplicating real individuals in their entirety due to overfitting in the synthesis model or a simple dataset
  2. Corresponding with real individuals when considering quasi-identifiers (QIs) only
  3. Do not correspond with real individual when considering QIs only

Replica Analytics
AN AETION COMPANY

What's Different About Synthetic Data?

- In synthetic data we can classify each record as:
  1. Duplicating real individuals in their entirety due to overfitting in the synthesis model or a simple dataset → Corresponds to traditional reidentification risk
  2. Corresponding with real individuals when considering quasi-identifiers (QIs) only → Corresponds to modified reidentification risk
  3. Do not correspond with real individual when considering QIs only → Does not pose a privacy risk

Replica Analytics
AN AETION COMPANY

Attribution Disclosure: Find a similar record in the synthetic data and learn something new

Quasi-identifiers

Sensitive variables

Sex	Yearof Birth	NDC
Male	1985	009-0031
Male	1988	0023-3670
Male	1982	0074-5182
Female	1983	0078-0379
Female	**1989**	**65862-403**
Male	1981	55714-4446
Male	1982	55714-4402
Female	1987	55566-2110
Male	1981	55289-324
Female	1986	54868-6348
Male	1980	53808-0540

Replica Analytics

AN AETION COMPANY

Learning Something New

		Similarity in Real Sample	
		Individual is Similar to Others	Individual is an Outlier
Similarity Between Real & Synthetic Samples	Individual's Synthetic Information Similar to Real Information	Low Attribution Risk	High Attribution Risk
	Individual's Synthetic Information Different from Real Information	Low Attribution Risk	Low Attribution Risk
Note: This table only applies to records that match between the synthetic and real data, and hence have passed the first test for what is defined as meaningful identity disclosure.			

A synthetic record matching a real individual is harmful if and only if it allows an attacker to learn something new about a real individual; that could not be learned through inference on a complete dataset

Replica Analytics
AN AETION COMPANY

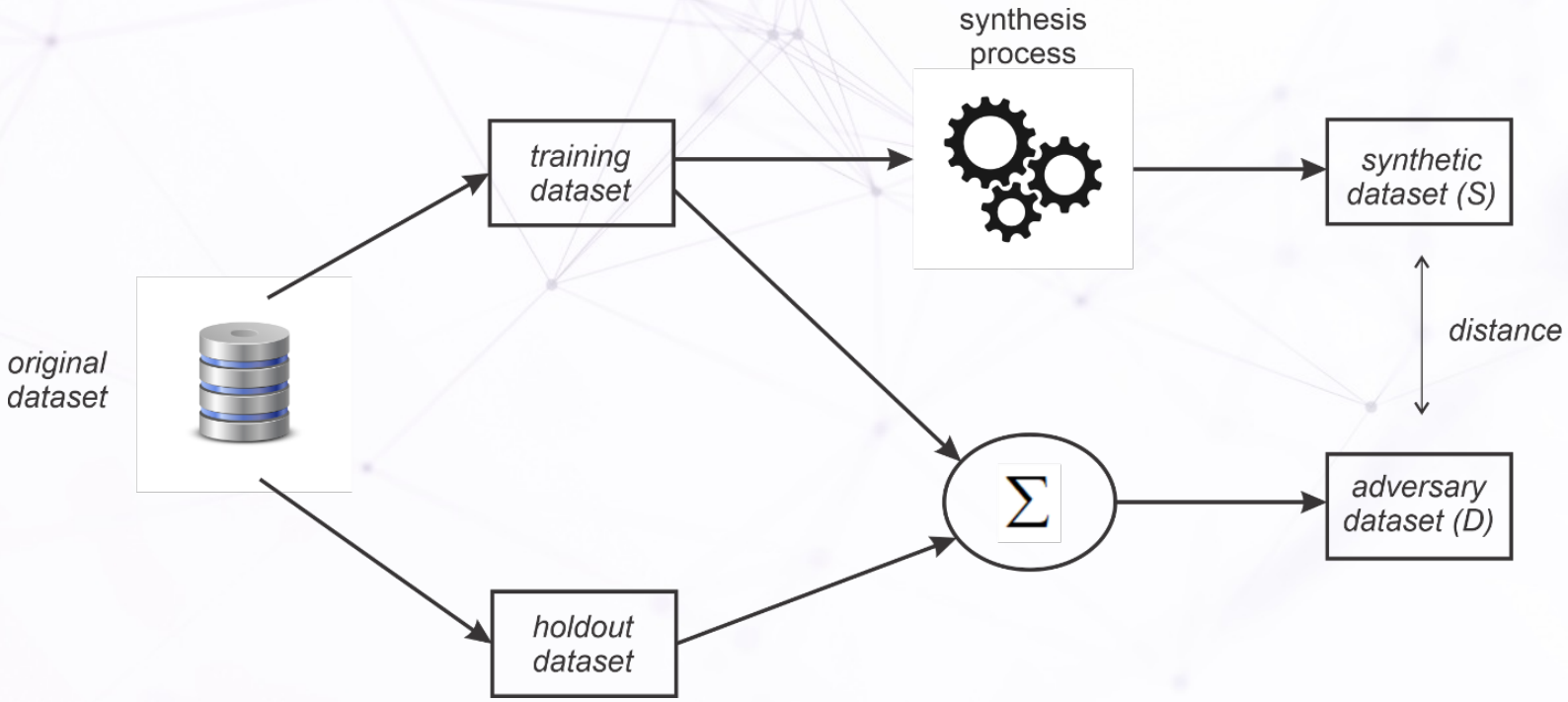
Published risk assessment results for synthetic data generated using sequential tree synthesis method:

El Emam K, Mosquera L, Bass J. Evaluating Identity Disclosure Risk in Fully Synthetic Health Data: Model Development and Validation. J Med Internet Res 2020;22(11):e23139, doi: 10.2196/23139.

	Synthetic Data Risk	
	Population-to-sample risk	Sample-to-population risk
Washington State Inpatient Database	0.00056	0.0197
Canadian COVID-19 cases	0.0043	0.0086

Replica Analytics
AN AETION COMPANY

Membership disclosure: is the distance between S and D predictive of which records are in the training dataset



Replica Analytics
AN AETION COMPANY

Data Synthesis as a Privacy Enhancing Technology

Replica Analytics

AN AETION COMPANY

Promise of Synthetic Data

- Can be applied to a wide range of dataset sizes and complexities
- Does not require expert determination of which components of a dataset have a high privacy risk
- Can produce higher utility data for small datasets that are difficult to anonymize using traditional methods
- Can be combined with other PETs or controls to produce more robust solutions
- Automate-able, scale-able

Replica Analytics
AN AETION COMPANY

Limitations of Synthetic Data

- Where synthetic data falls within current regulatory stance is uncertain
- Some use cases will only use synthetic data for analysis development but will still want to ‘validate’ results on the real data
- May be computationally intensive to generate with deep learning models
- Industry wide question for how to report privacy risks: average risk across a dataset or maximum risk observed for a given individual

Thank you!

Replica Analytics

AN AETION COMPANY