SYNTHETIC DATA GENERATION 101

FEB. 9, 2022 | 11AM EDT

Presented by

Dr. Khaled El Emam,
SVP and General Manager,
Replica Analytics

Synthetic Data Generation 101

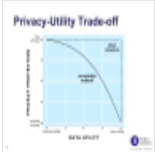
Khaled El Emam

February 9, 2022

Replica Analytics
AN AETION COMPANY

Agenda

Introduction to Synthesis 1 General description of what synthetic data is and general use cases



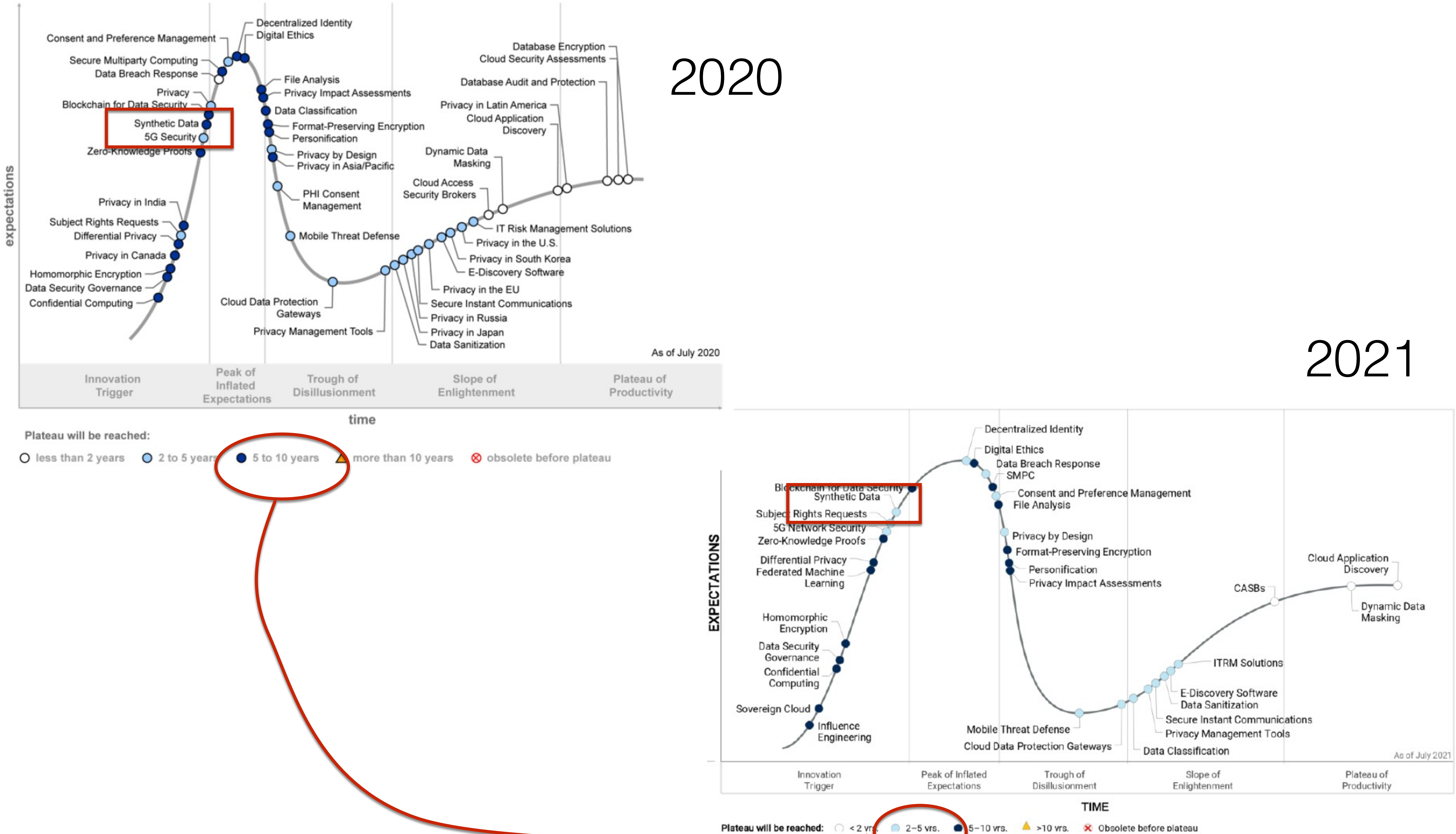
Privacy and Utility 2 An examination of privacy risks and the utility of synthetic data

FAQs 3 Some commonly asked questions about synthetic datasets

Additional Use Cases 4 Beyond data sharing, SDG has additional applications

Replica Analytics
AN AETION COMPANY

The adoption of synthetic data has been accelerating quite rapidly



Gartner
Hype Cycle for Privacy, 2021

Replica Analytics
AN AETION COMPANY

Gartner predicts synthetic data will have a non-trivial impact on privacy violations and sanctions

Replica Analytics
AN AETION COMPANY

The Erosion of Trust ?

The New York Times

Your Data Were 'Anonymized'? These Scientists Can Still Identify You

Computer scientists have developed an algorithm that can pick out almost any American in databases supposedly stripped of personal information.

Opinion | THE PRIVACY PROJECT

Twelve Million Phones, One Dataset, Zero Privacy

By Stuart A. Thompson and Charlie Warzel

DEC. 19, 2019

theguardian

'Anonymised' data can never be totally anonymous, says study

Findings say it is impossible for researchers to fully protect real identities in datasets

You're very easy to track down, even when your data has been anonymized

A new study shows you can be easily re-identified from almost any database, even when your personal details have been stripped out.

by **Charlie Jee** Jul 23, 2019

ACM TECHNEWS

'Anonymized' Data Can Never Be Totally Anonymous, says Study

By The Guardian

HUFFPOST

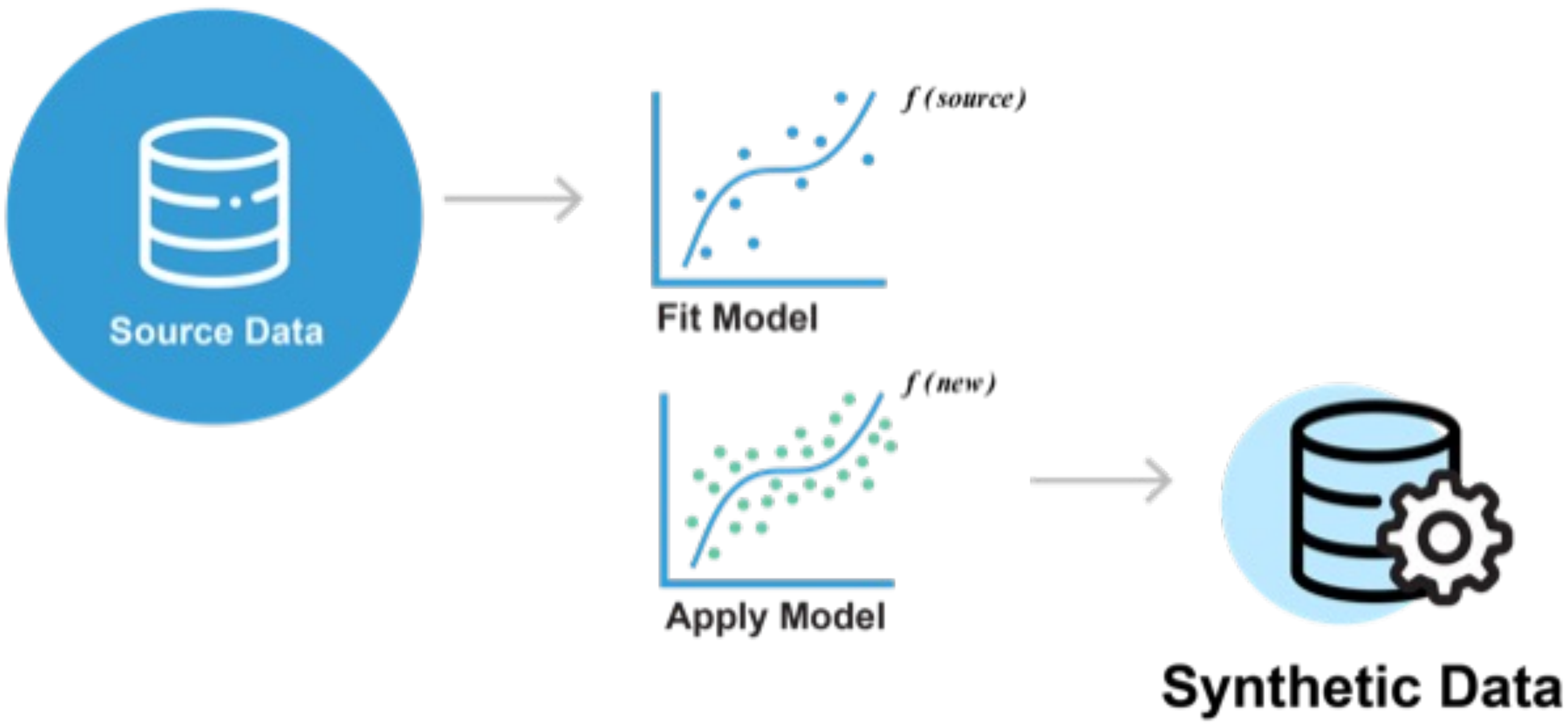
Online Profiling and Invasion of Privacy: The Myth of Anonymization

02/20/2013 12:23 pm ET | **Updated** Apr 22, 2013

Replica Analytics

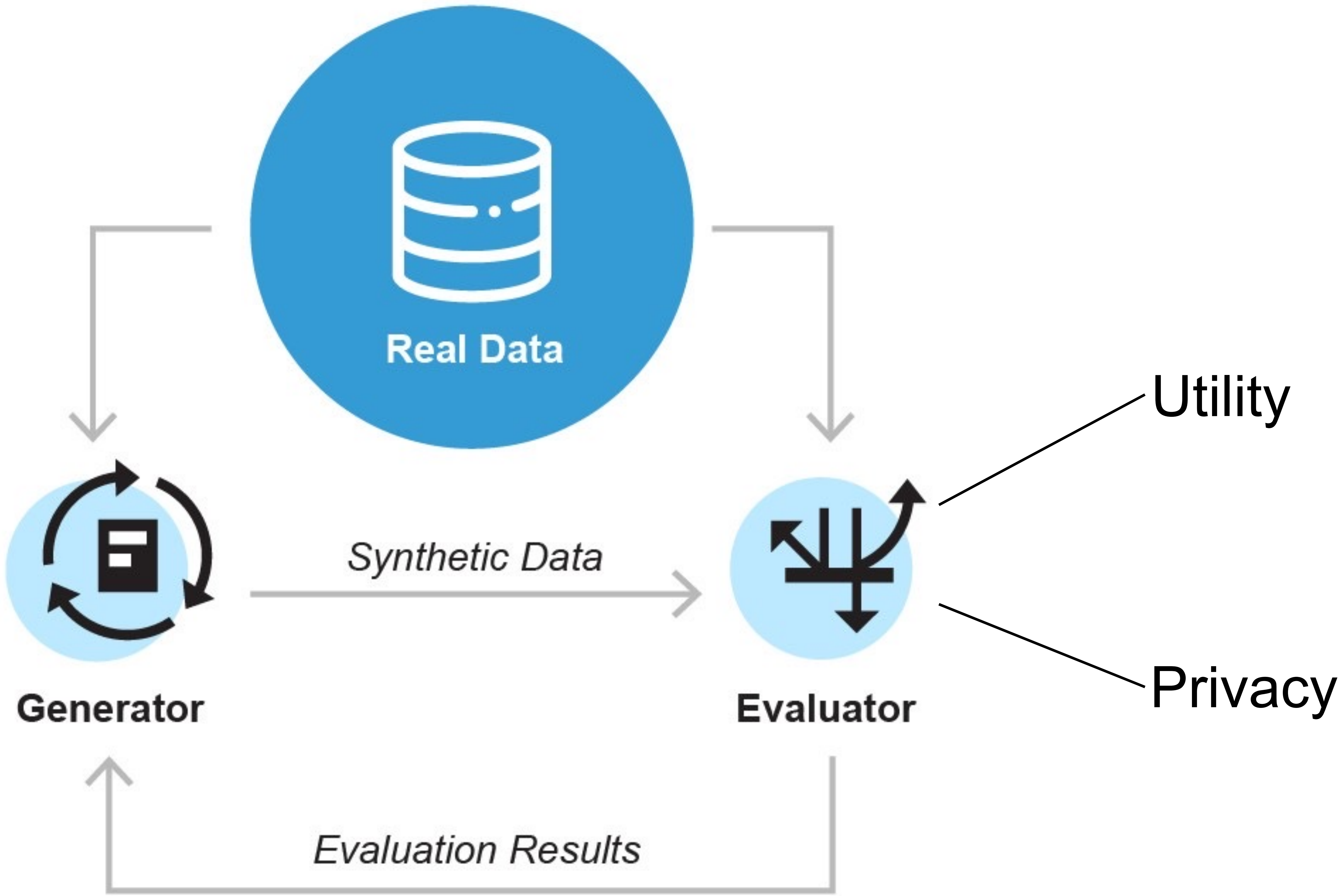
AN AETION COMPANY

The Synthesis Process



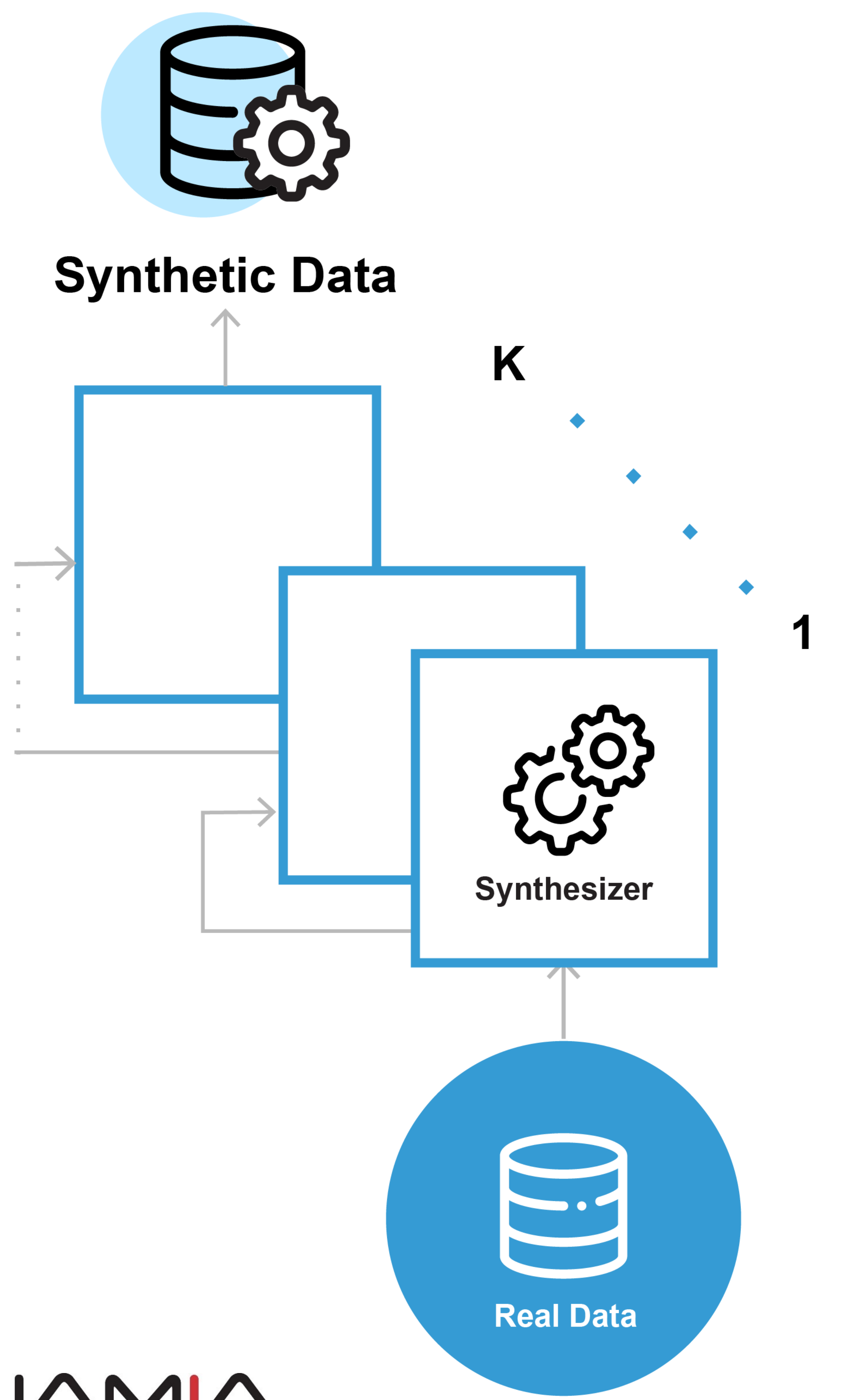
COU1A	AGECAT	AGELE70	WHITE	MALE	BMI
United States	2	1	1	1	33.75155
United States	2	1	1	0	39.24707
United States	1	1	1	0	26.5625
United States	4	1	1	1	40.58273
United States	5	0	0	1	24.42046
United States	5	0	1	0	19.07124
United States	3	1	1	1	26.04938
United States	4	1	1	1	25.46939

Training a generative model uses a utility – privacy loss function



Replica Analytics
AN AETION COMPANY

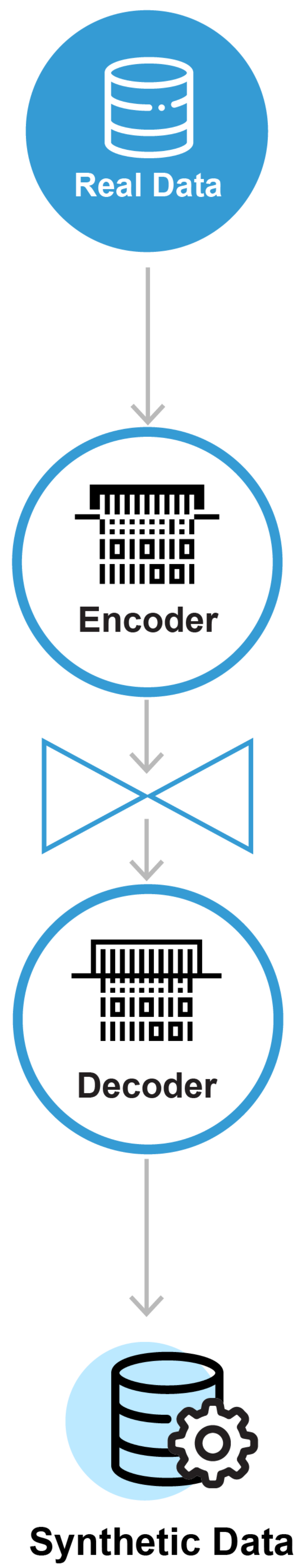
Sequential synthesis utilizes multiple machine learning methods in a sequence



JAMIA

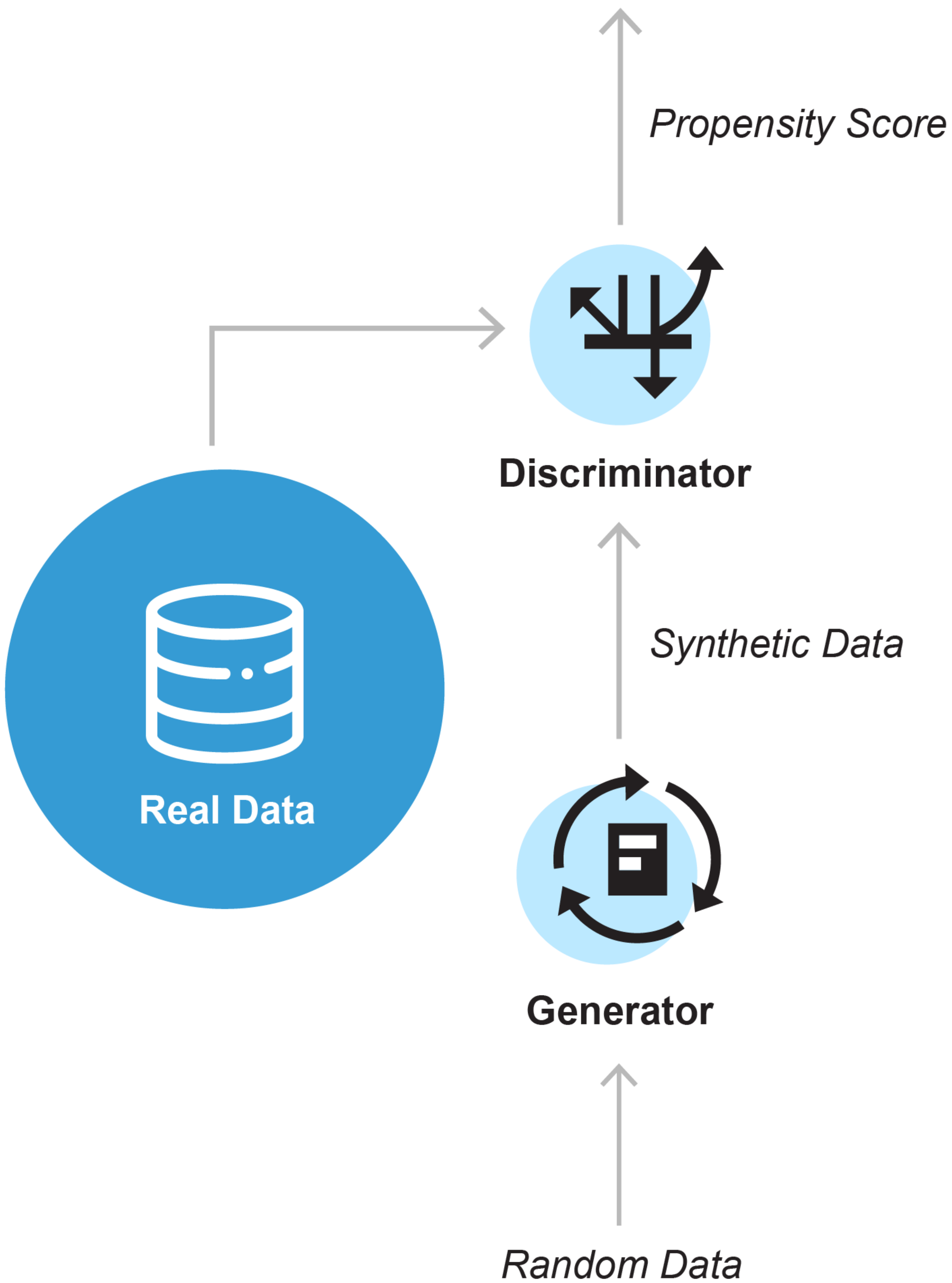
Replica Analytics
AN AETION COMPANY

Variational Auto Encoder (VAE) compresses the input into a latent space



Replica Analytics
AN AETION COMPANY

Generative Adversarial Network (GAN)



There are seven common use cases for synthetic data

1. Machine learning

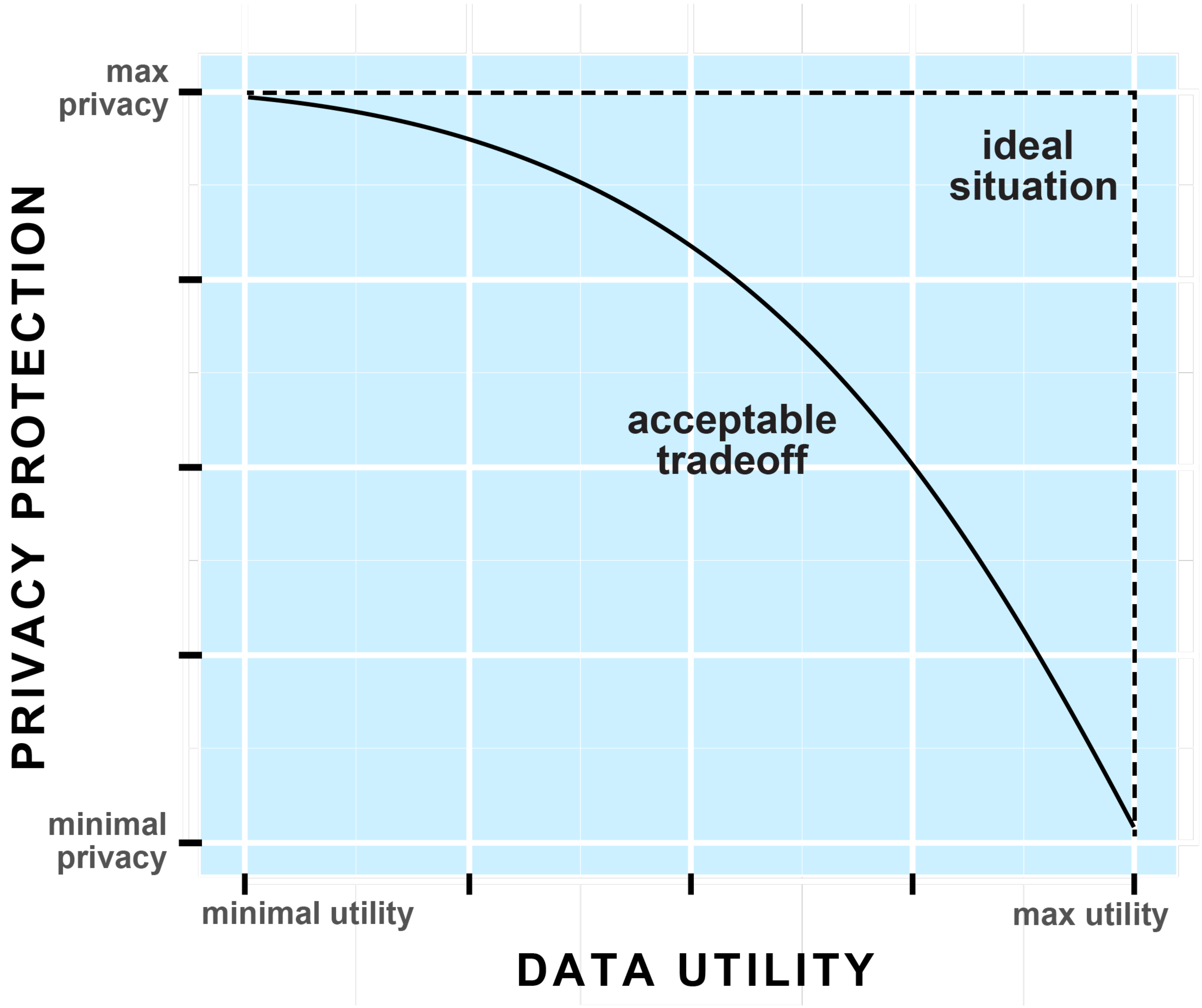
   (model evaluation, data augmentation, sharing ML models)
2. Software testing
3. Education, training, and hackathons
4. Data retention
5. Vendor assessment
6. Internal secondary use

   (exploratory and detailed analytics)
7. External data sharing

Replica Analytics

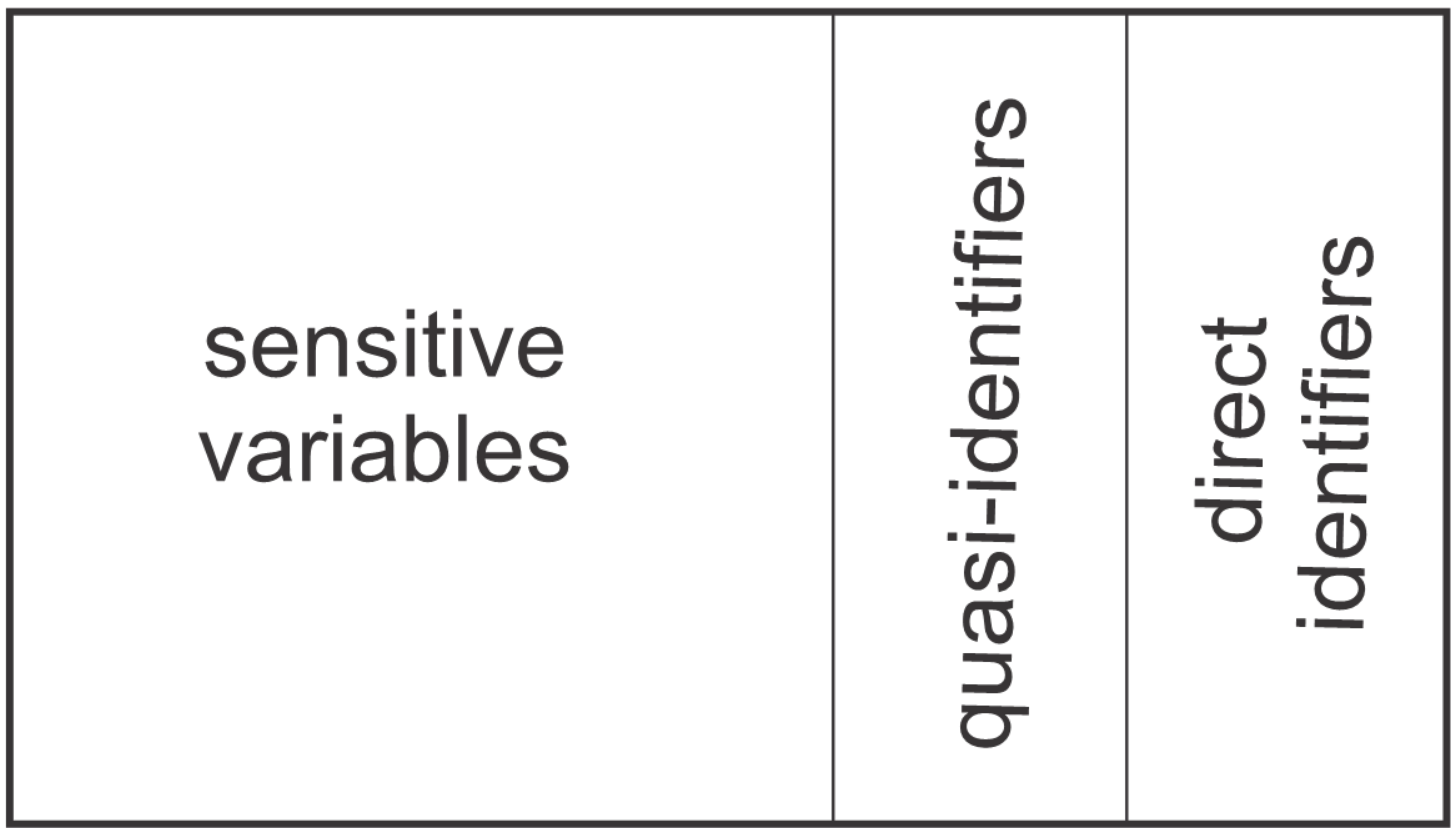
AN AETION COMPANY

Privacy-Utility Trade-off



Replica Analytics
AN AETION COMPANY

Variables in a dataset can be classified into one of three types

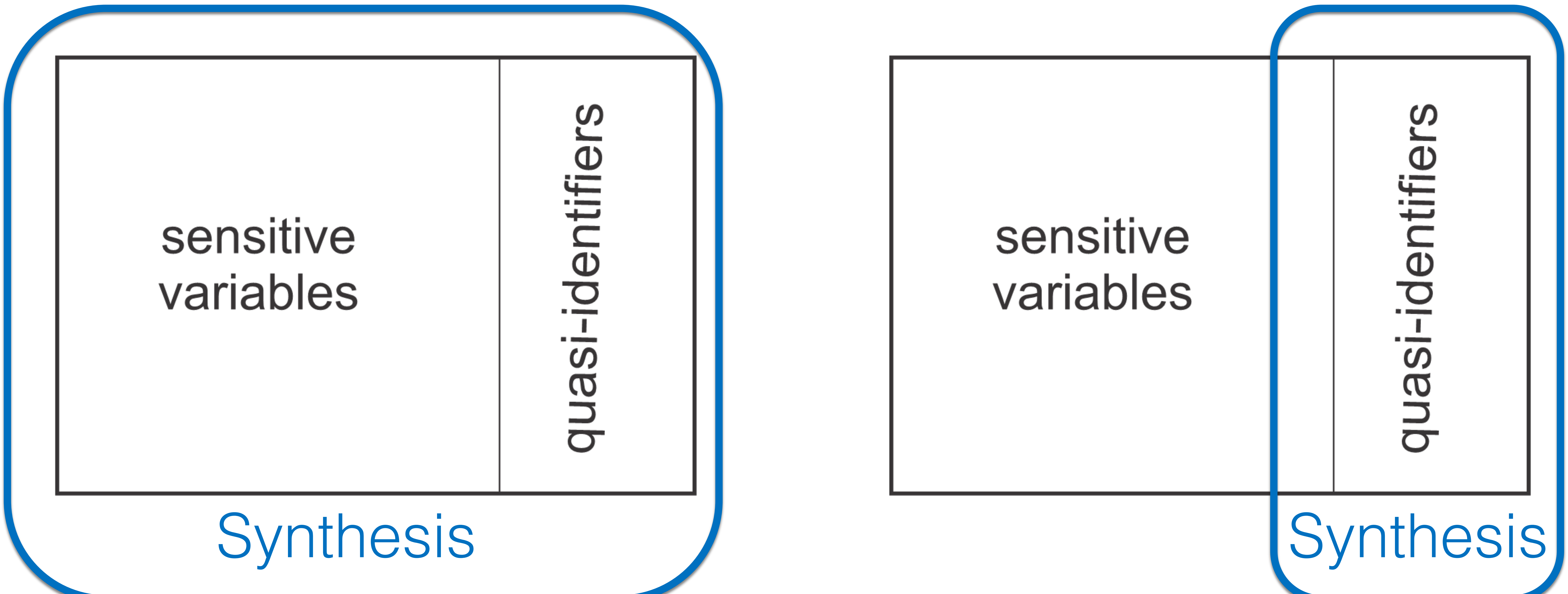


Replica Analytics
AN AETION COMPANY

Two Synthesis Strategies

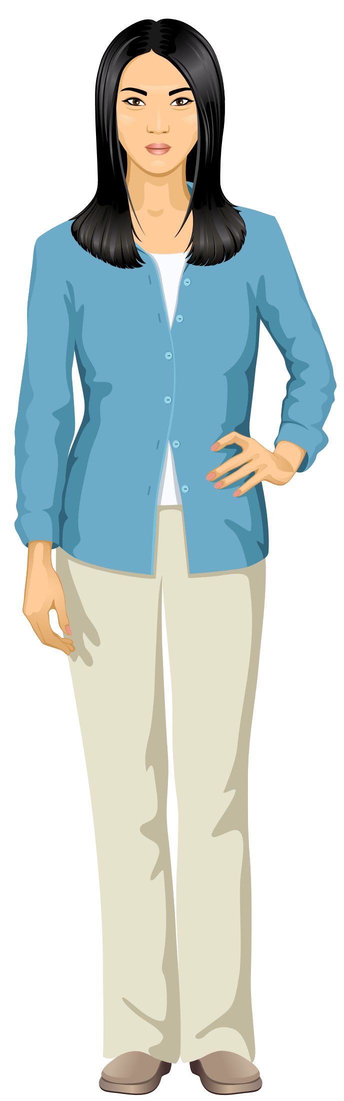
Full Synthesis
Synthesize all variables

Partial Synthesis
Synthesize quasi-identifiers



Replica Analytics
AN AETION COMPANY

Attribution disclosure: find a similar record in the synthetic data and learn something new



Quasi-identifiers

New Information

Sex	Year of Birth	NDC
Male	1975	009-0031
Male	1988	0023-3670
Male	1972	0074-5182
Female	1993	0078-0379
Female	1989	65862-403
Male	1991	55714-4446
Male	1992	55714-4402
Female	1987	55566-2110
Male	1971	55289-324
Female	1996	54868-6348
Male	1980	53808-0540

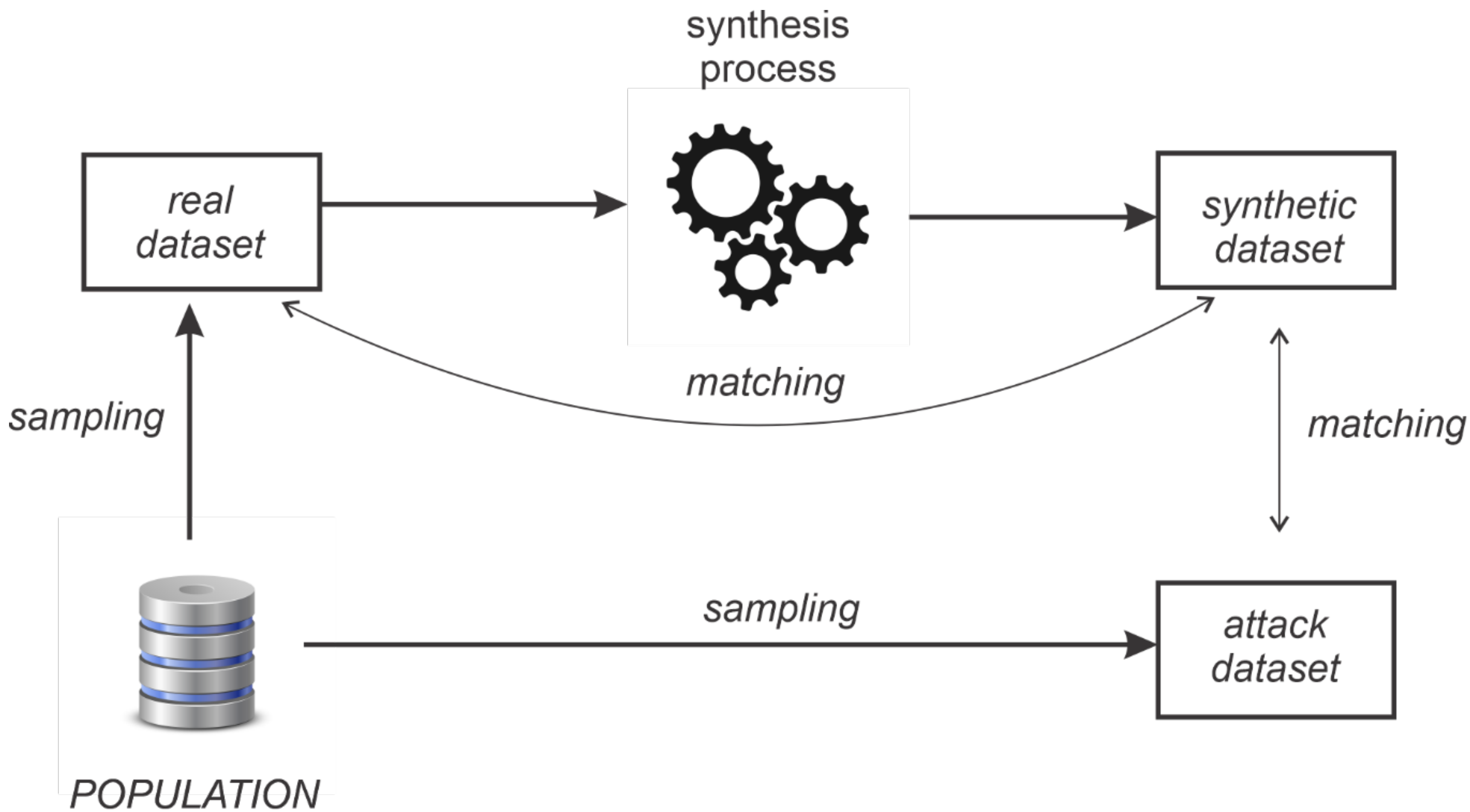
Evaluations of attribution risks show that it is low in multiple studies across multiple datasets

Dataset	Fully Synthetic Data	Original Data
Washington Hospital Data (Discharge)	0.0197	0.098
Canadian COVID-19 Data (Public Health)	0.0086	0.034

A commonly used risk threshold = 0.09

Replica Analytics
AN AETION COMPANY

Membership disclosure



One way to classify utility metrics is as broad and narrow

broad metrics → narrow metrics

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.

There are multiple use cases for these metrics:

- Dataset vs model
- Use case: rank / hyperparameter tuning / communicate

narrow metrics

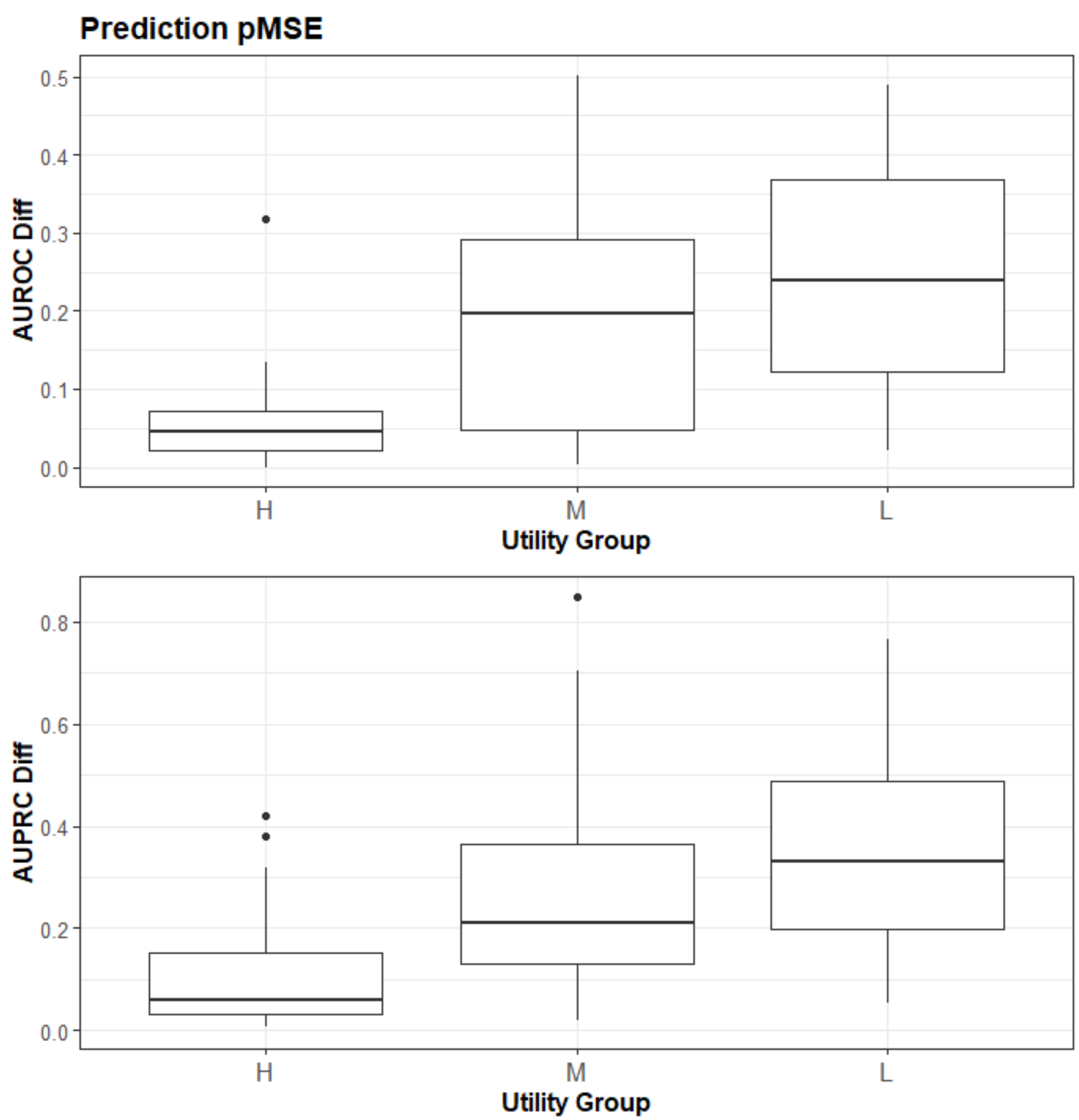
These are workload-specific and are what is of most interest to the data users. However, all the possible workloads will not be known in advance and therefore we have to consider representative workloads when developing and evaluating utility metrics.

There are different types:

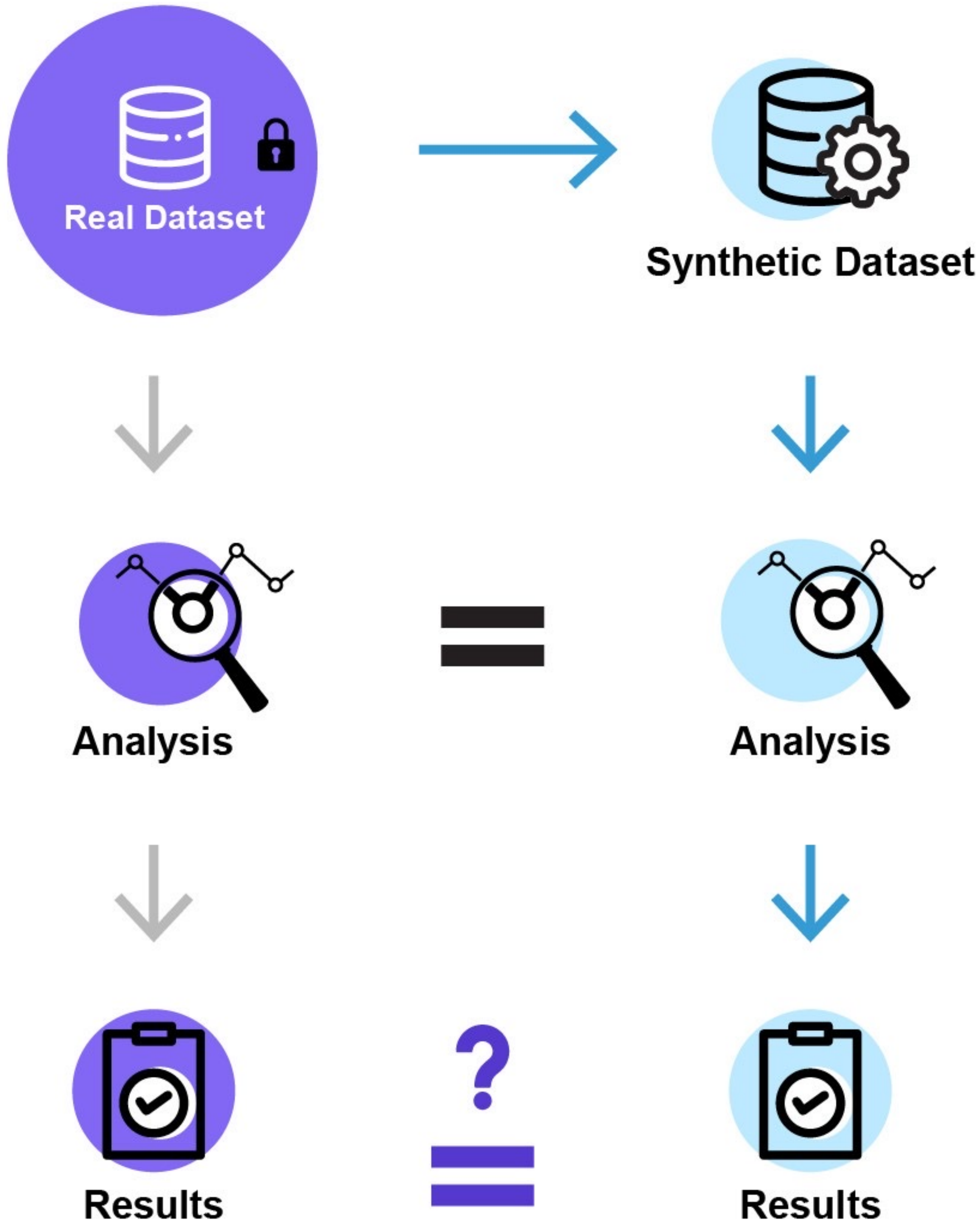
- Information loss metrics
- Inferential validity metrics

Replica Analytics
AN AETION COMPANY

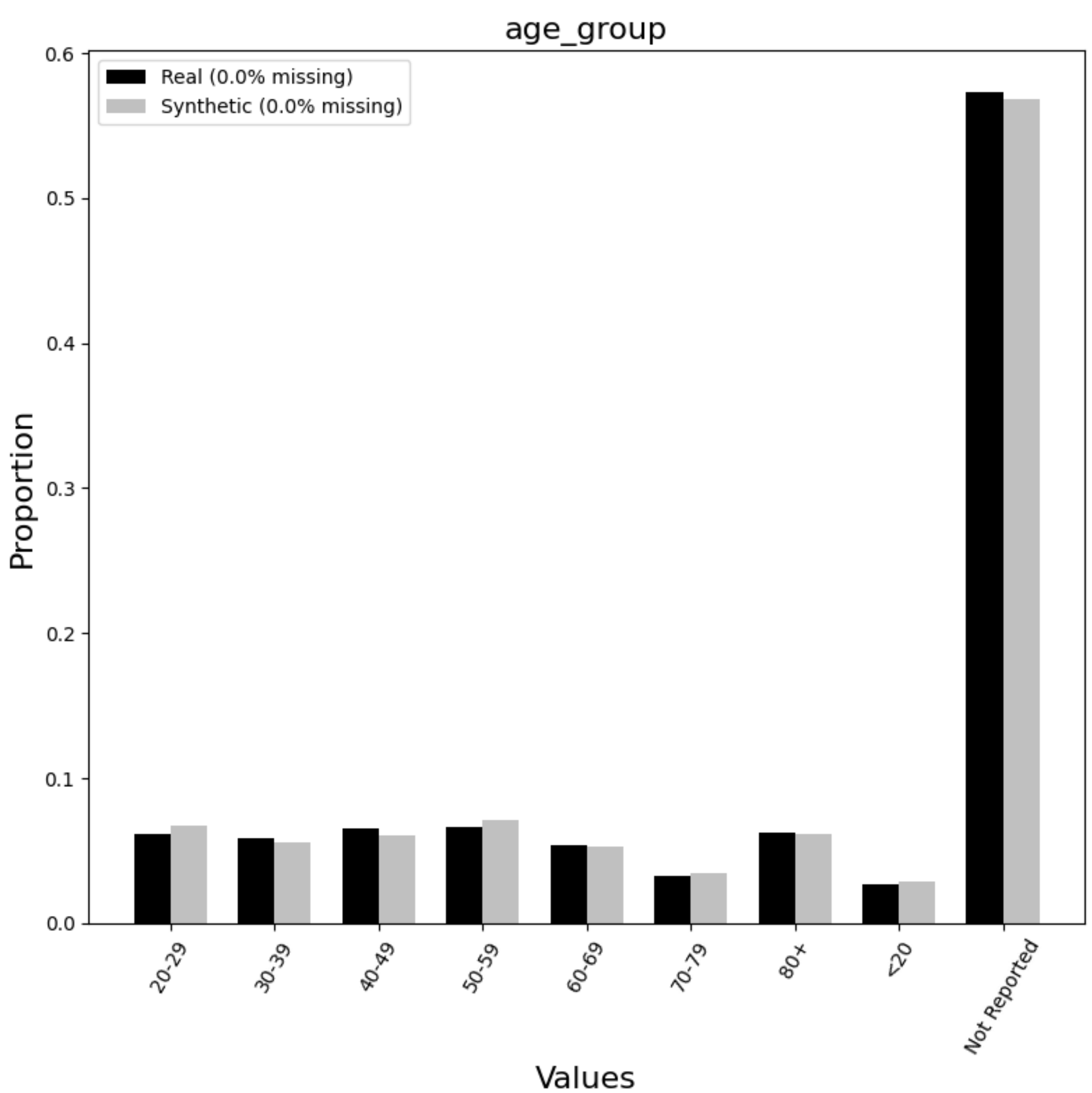
Broad utility metrics can rank SDG methods by their workload performance

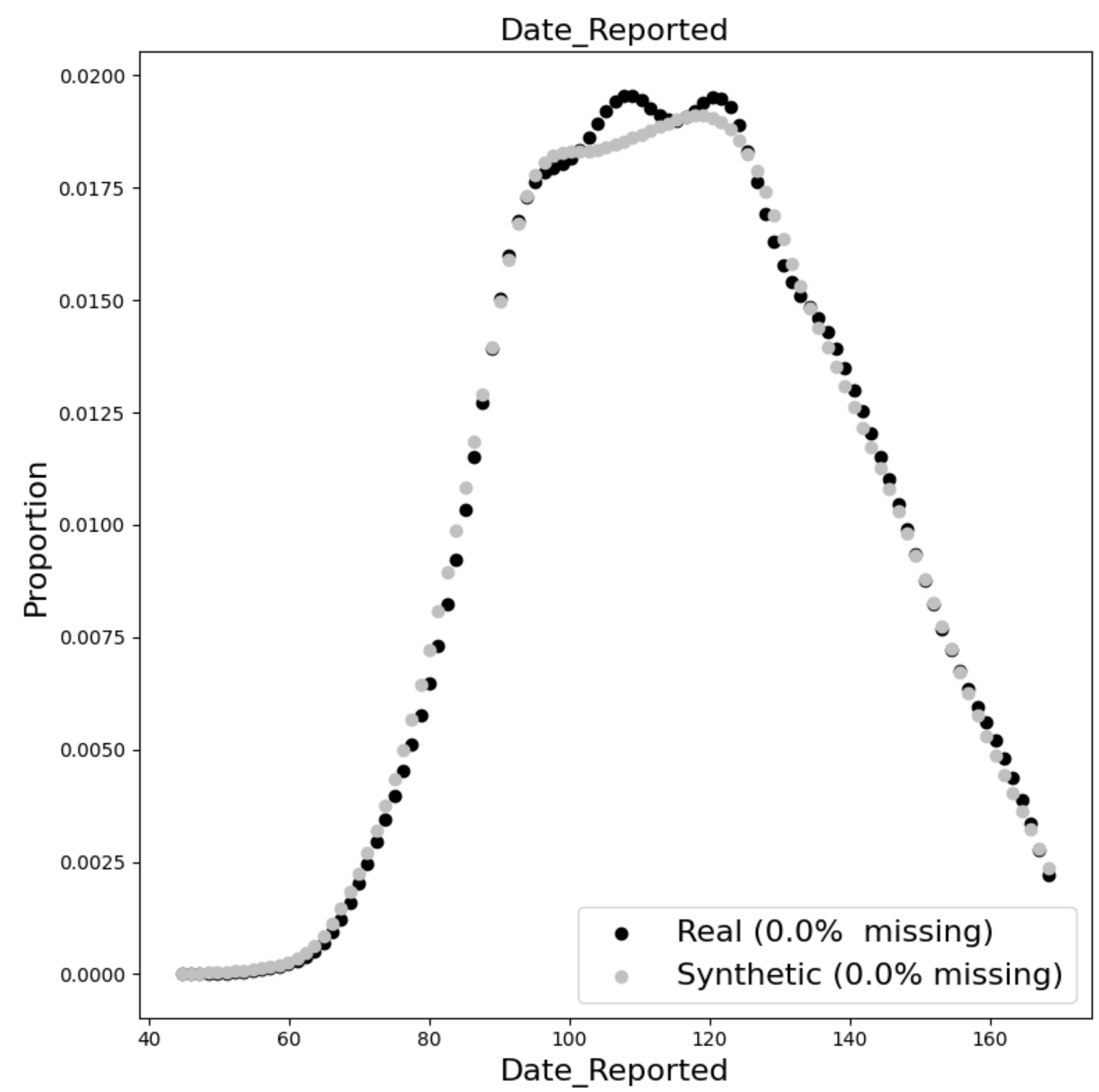


To evaluate utility one can compare the analysis results from real and synthetic data



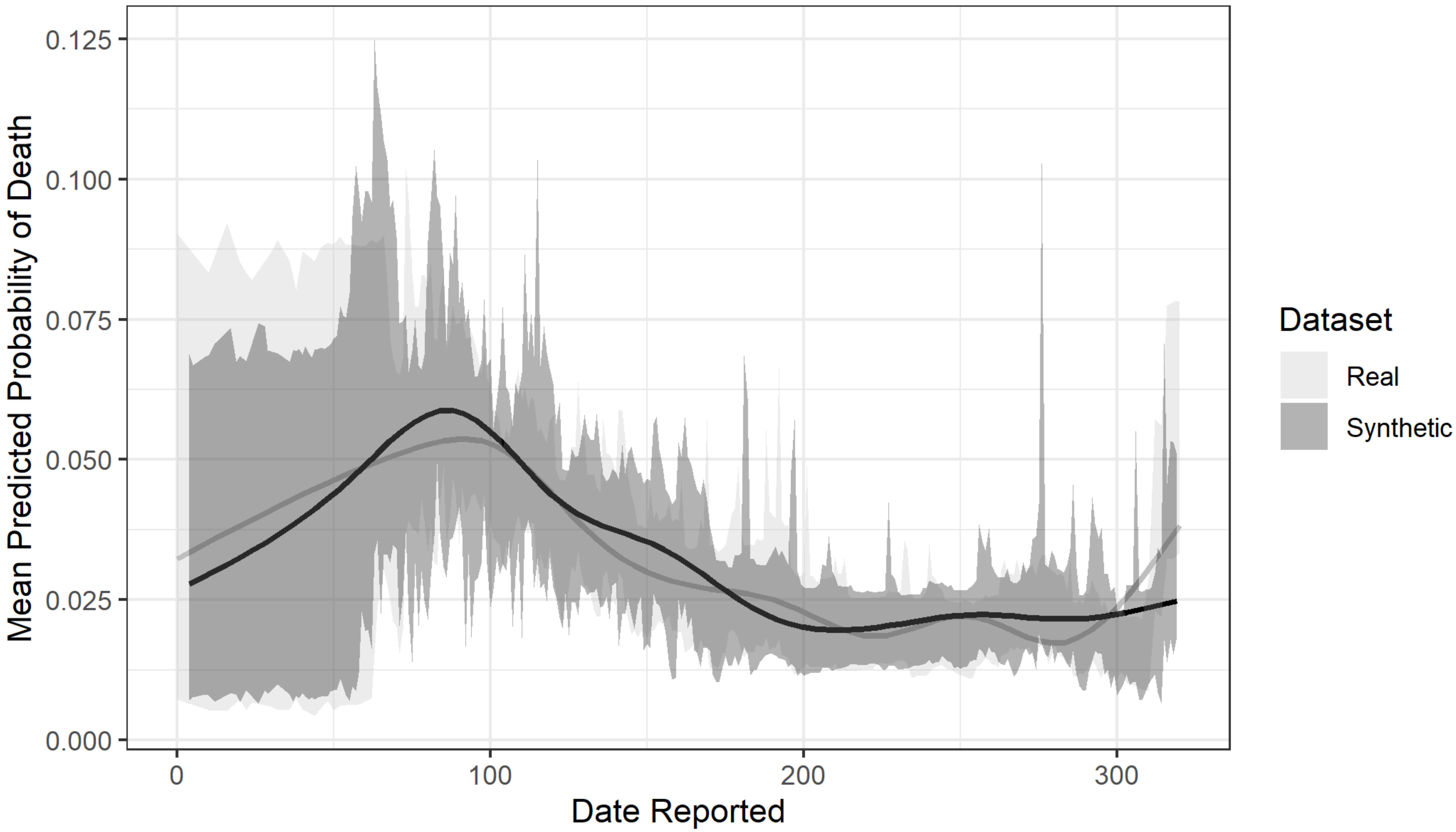
The univariate distributions of real and synthetic datasets look similar





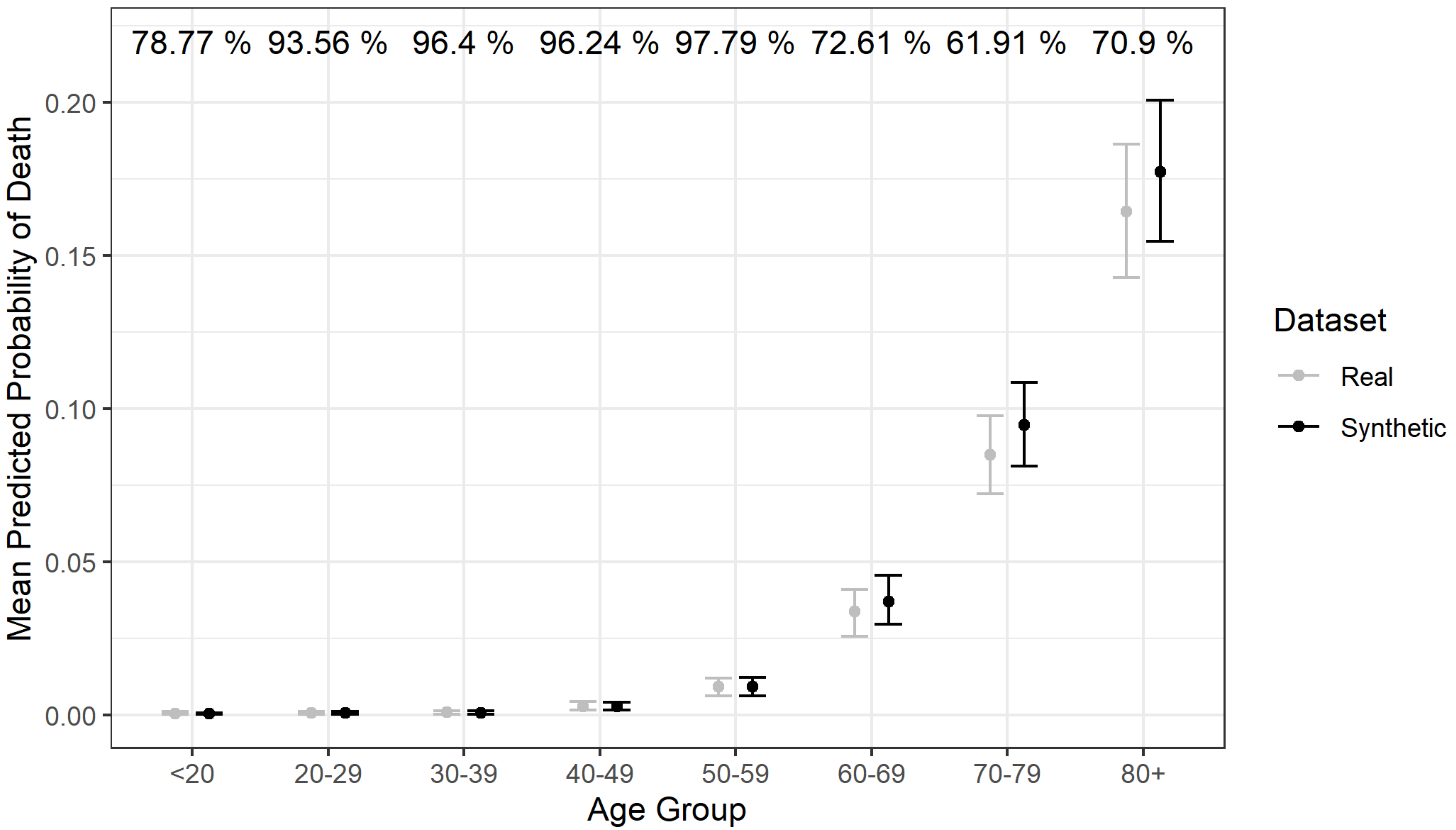
Replica Analytics
AN AETION COMPANY

Mortality over time for the Ontario COVID-19 case dataset



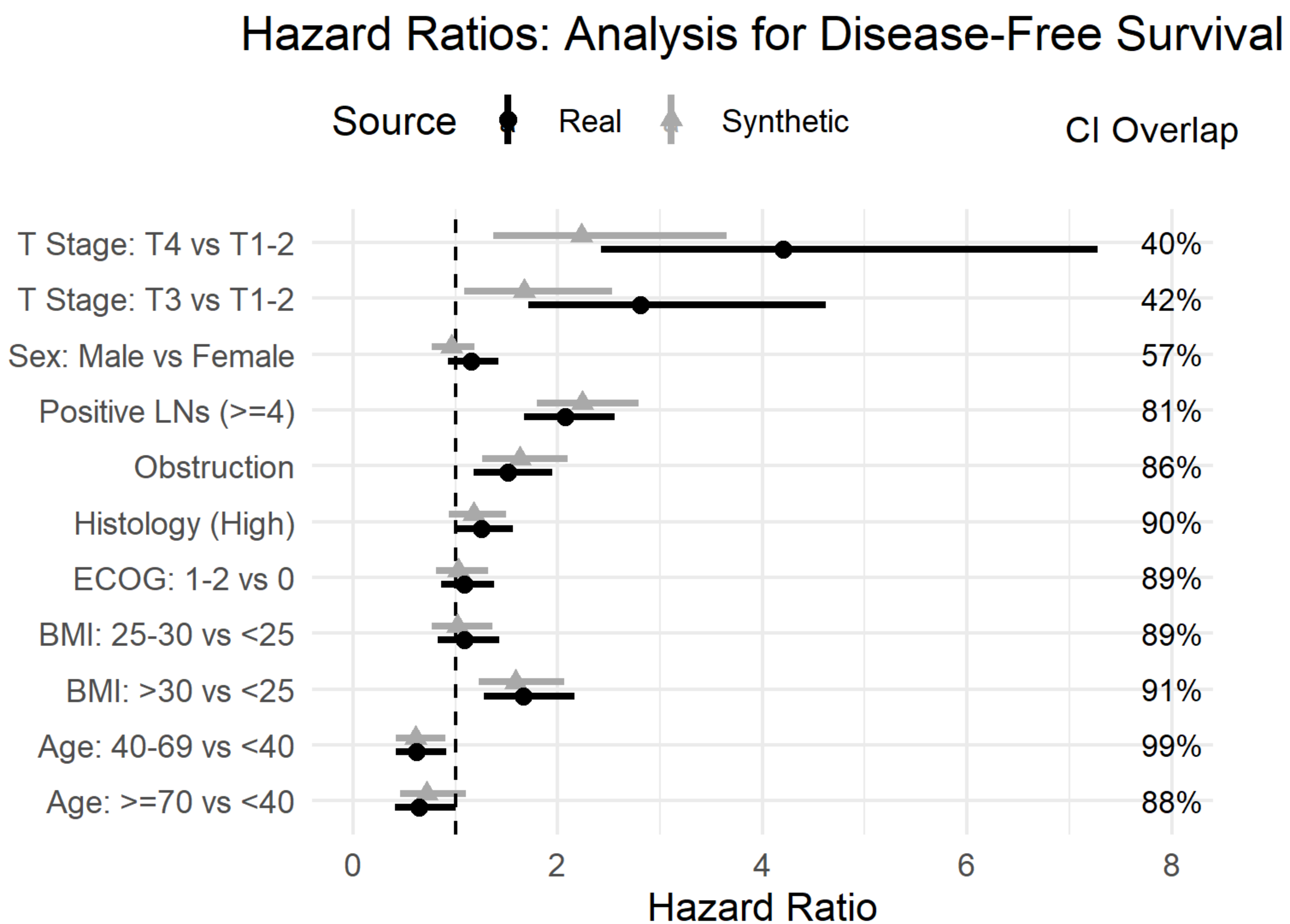
Replica Analytics
AN AETION COMPANY

Mortality by age for the Ontario COVID-19 case dataset



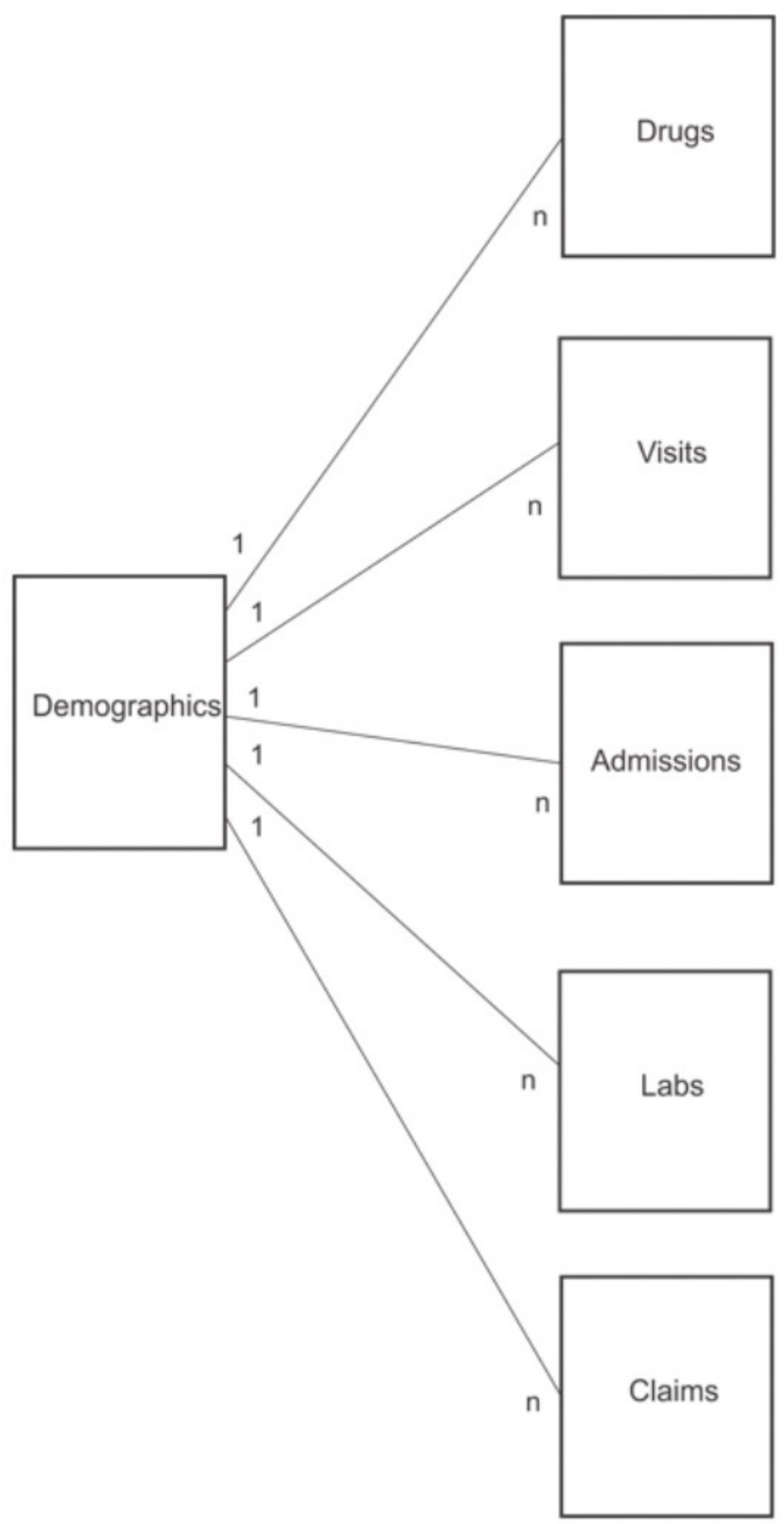
Replica Analytics
AN AETION COMPANY

Comparing real and synthetic data: Adjusted model of impact of bowel obstruction on DFS



Replica Analytics
AN AETION COMPANY

Longitudinal Data Model



Demographics
Age
Sex
Time to last day of follow-up available
Comorbidity score (elixhauser)

Drugs
Dispensed amount quantity
Relative dispensed time in days
Dispensed day supply quantity
Morphine use (binary)
Oxycodone use (binary)
Antidepressant use (binary)

Visits (ED)
Relative admission time in days
Problem code 1
Problem code 2
Resource intensity weights

Admissions (Hospital)
Relative time admitted in days
LOS
Diagnosis code 1
Diagnosis code 2
Resource intensity weight

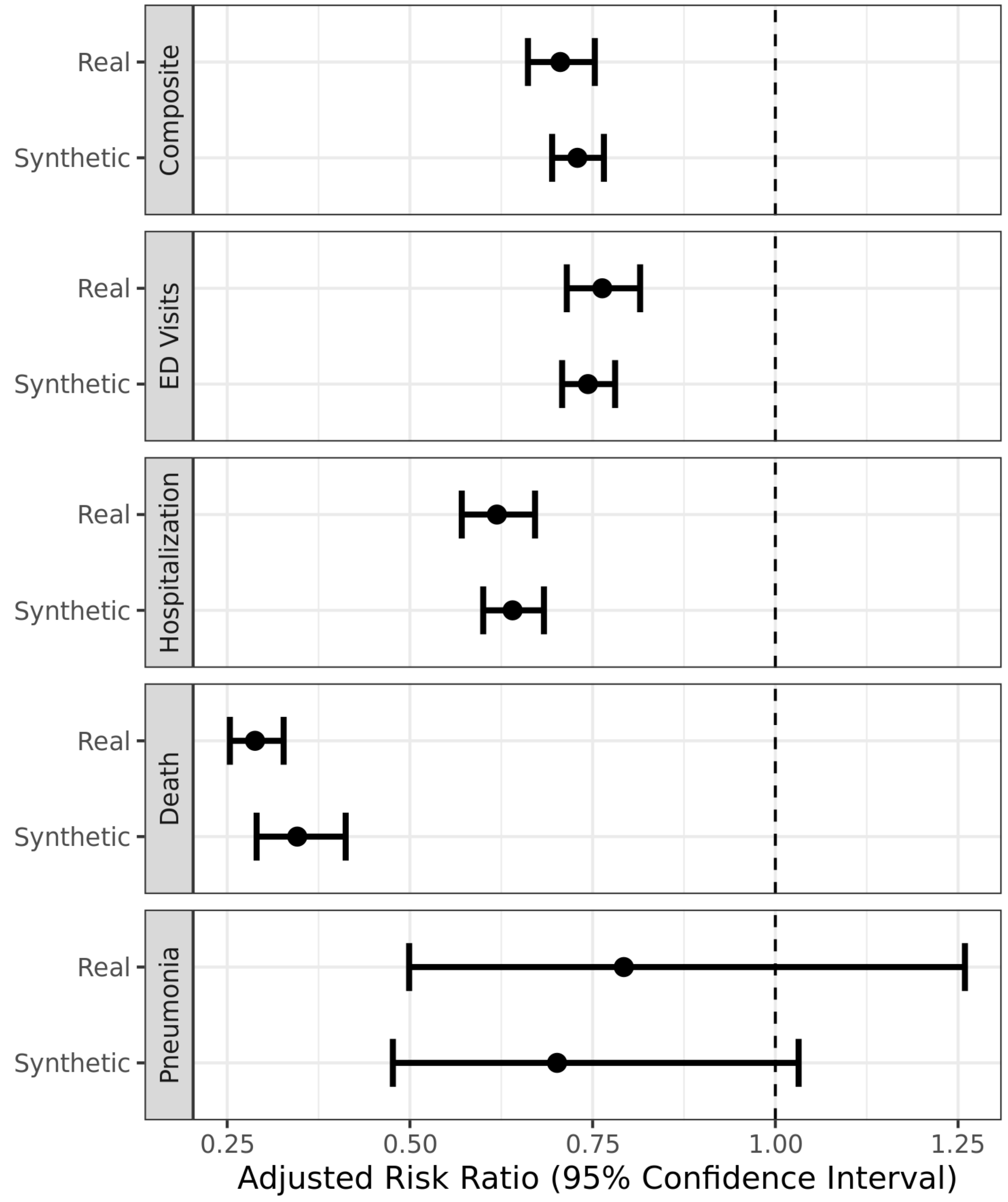
Lab
Test name
Test result (integer)
Relative time in days lab taken

Claims
Primary diagnosis code
Provide specialty
Relative service event start date

Replica Analytics
AN AETION COMPANY

Adjusted Cox Regression

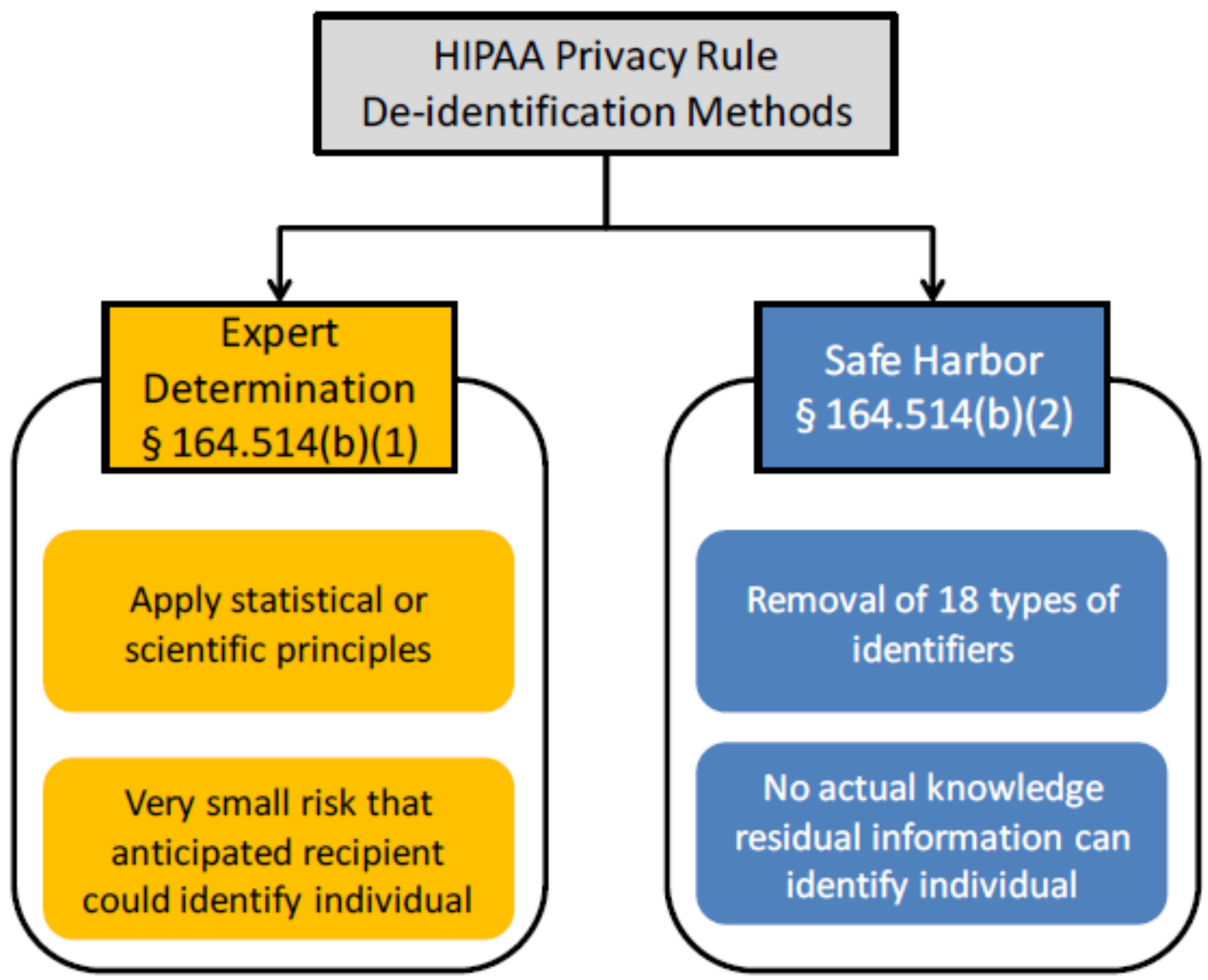
Note: Adjusted estimates include the following co-variates: age, sex, antidepressant use, Elixhauser score, ALT, eGFR, HCT; Opioid 1 served as the reference group



AN AETION COMPANY

Are controls needed to manage the privacy risks in synthetic data ?

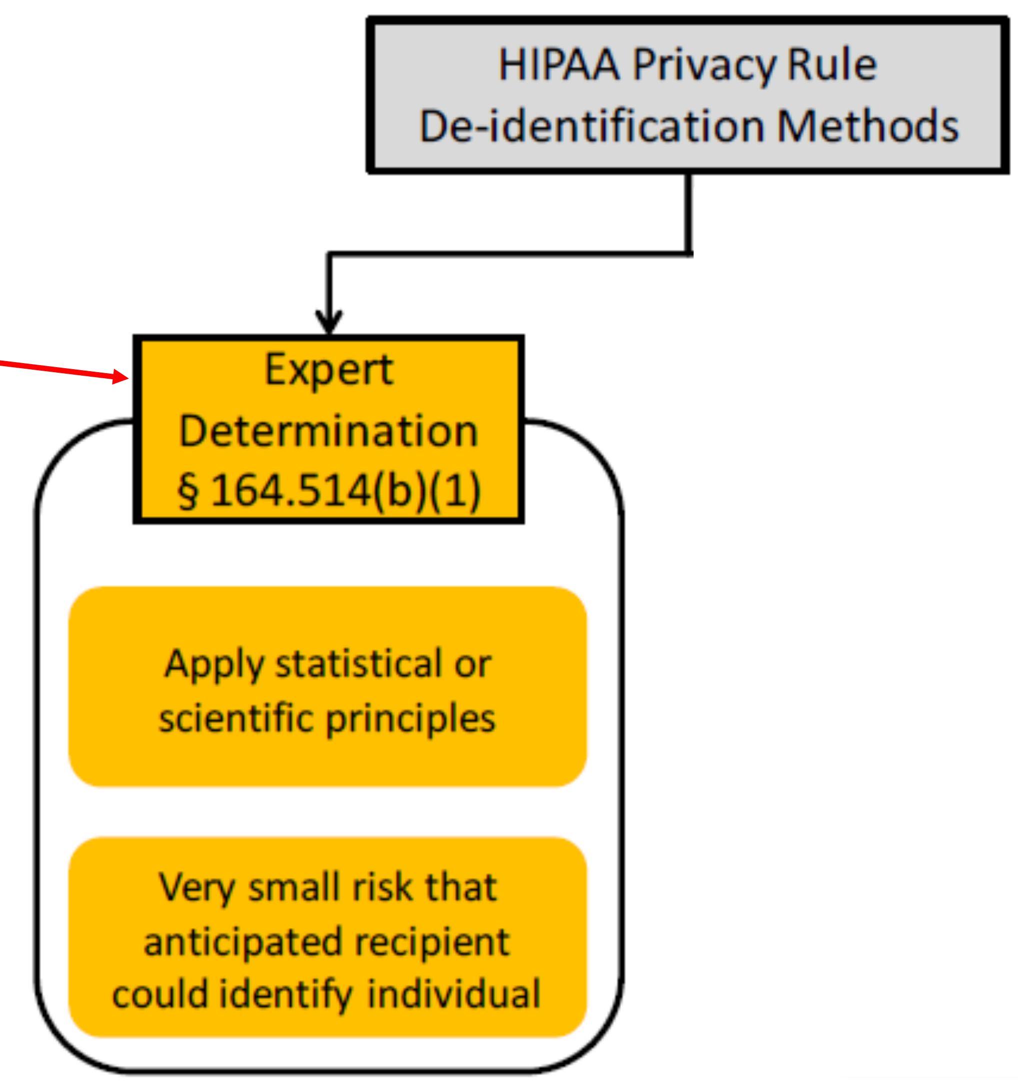
HIPAA de-identification methods are still being used – they are generally considered the gold standard



Replica Analytics
AN AETION COMPANY

The Expert Determination Method

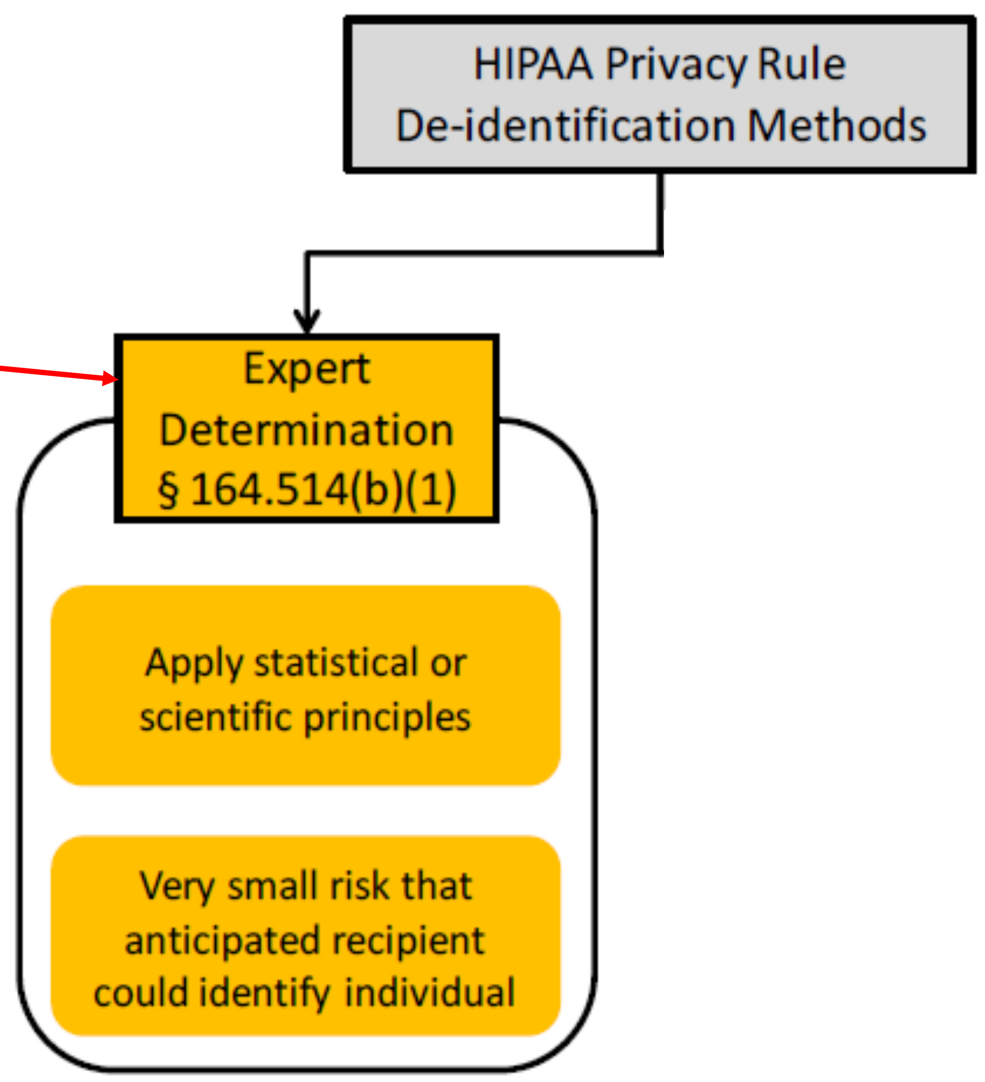
A person with appropriate knowledge of and experience with **generally accepted statistical and scientific principles and methods** for rendering information not individually identifiable



The Expert Determination method is more consistent with modern disclosure control practices

Applying such principles and methods, determines that the **risk is very small** that the information could be used, alone or in combination with other **reasonably available information**, by an **anticipated recipient** to identify an individual who is a subject of the information

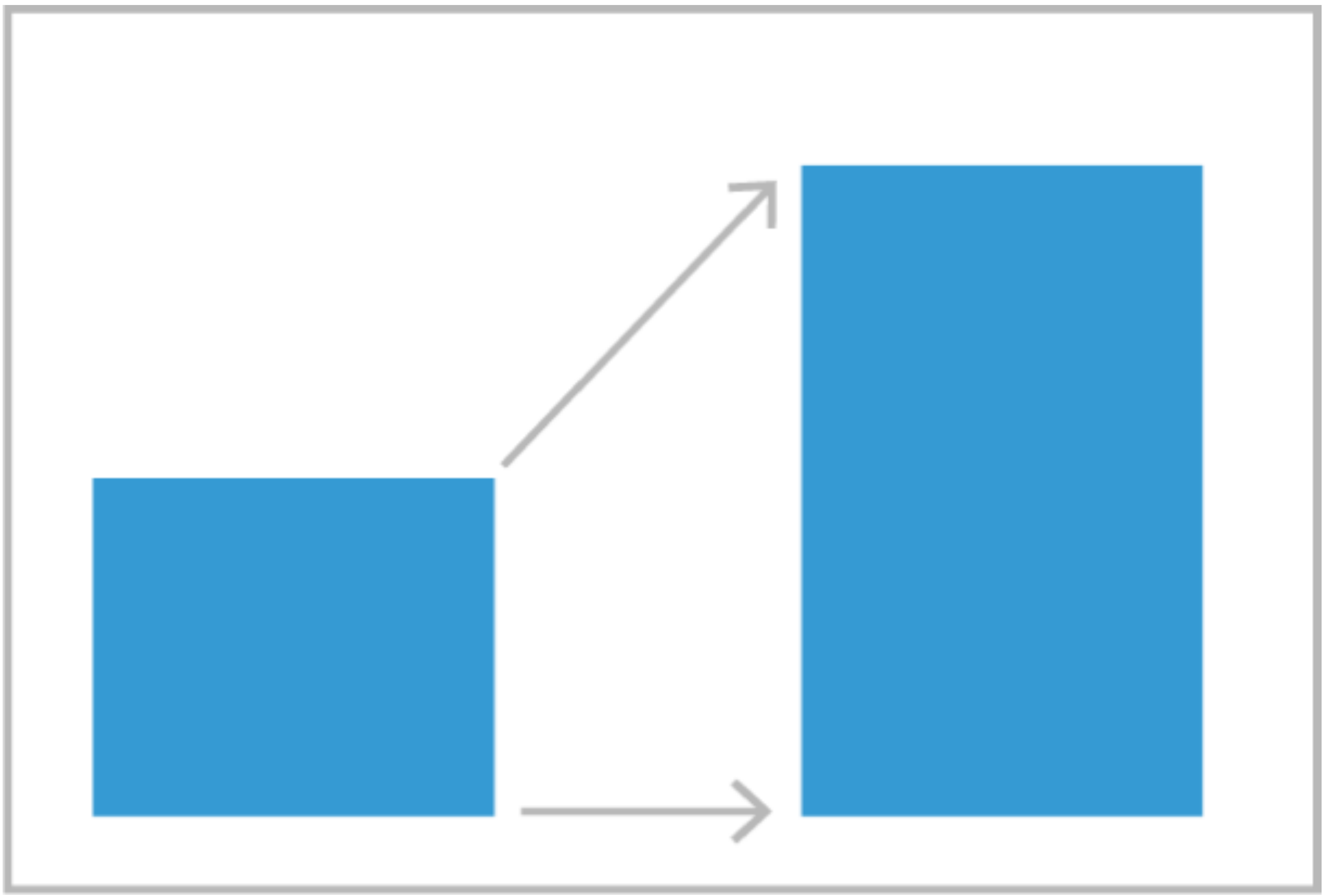
***Documents** the methods and results of the analysis that justify 'very small risk' determination*



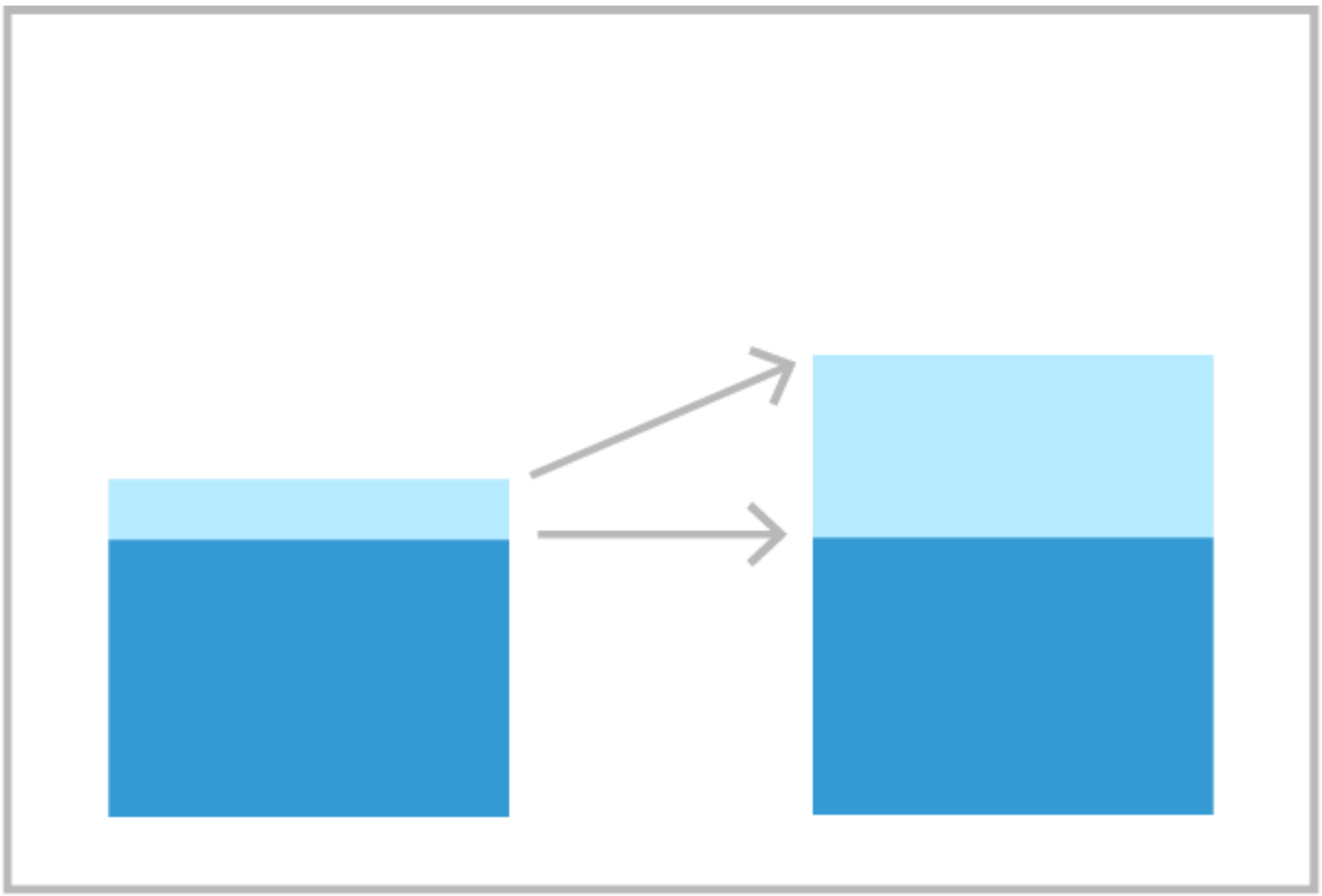
Replica Analytics

AN AETION COMPANY

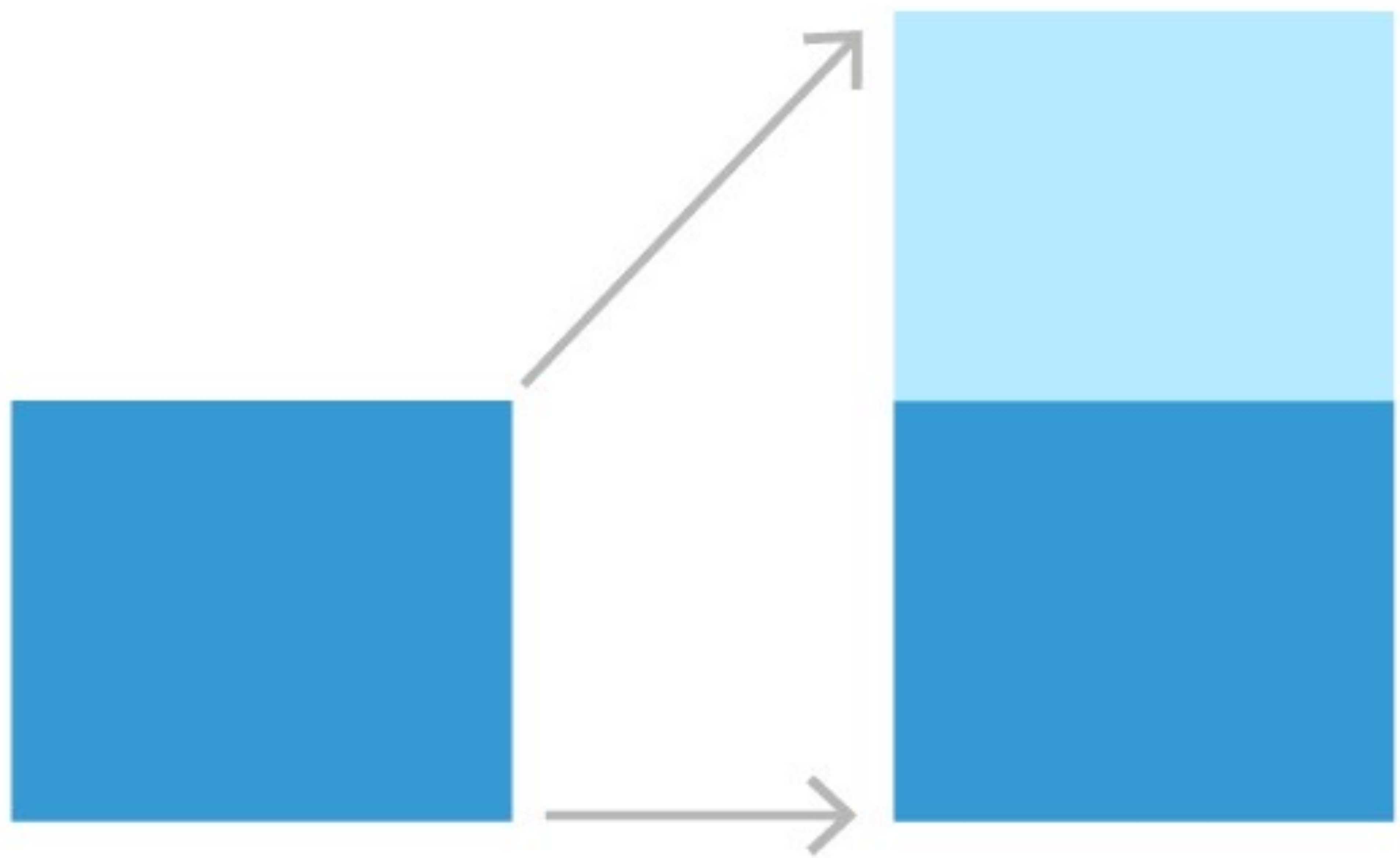
An important use case for SDG is data amplification



Amplification can also focus on a specific class in the dataset

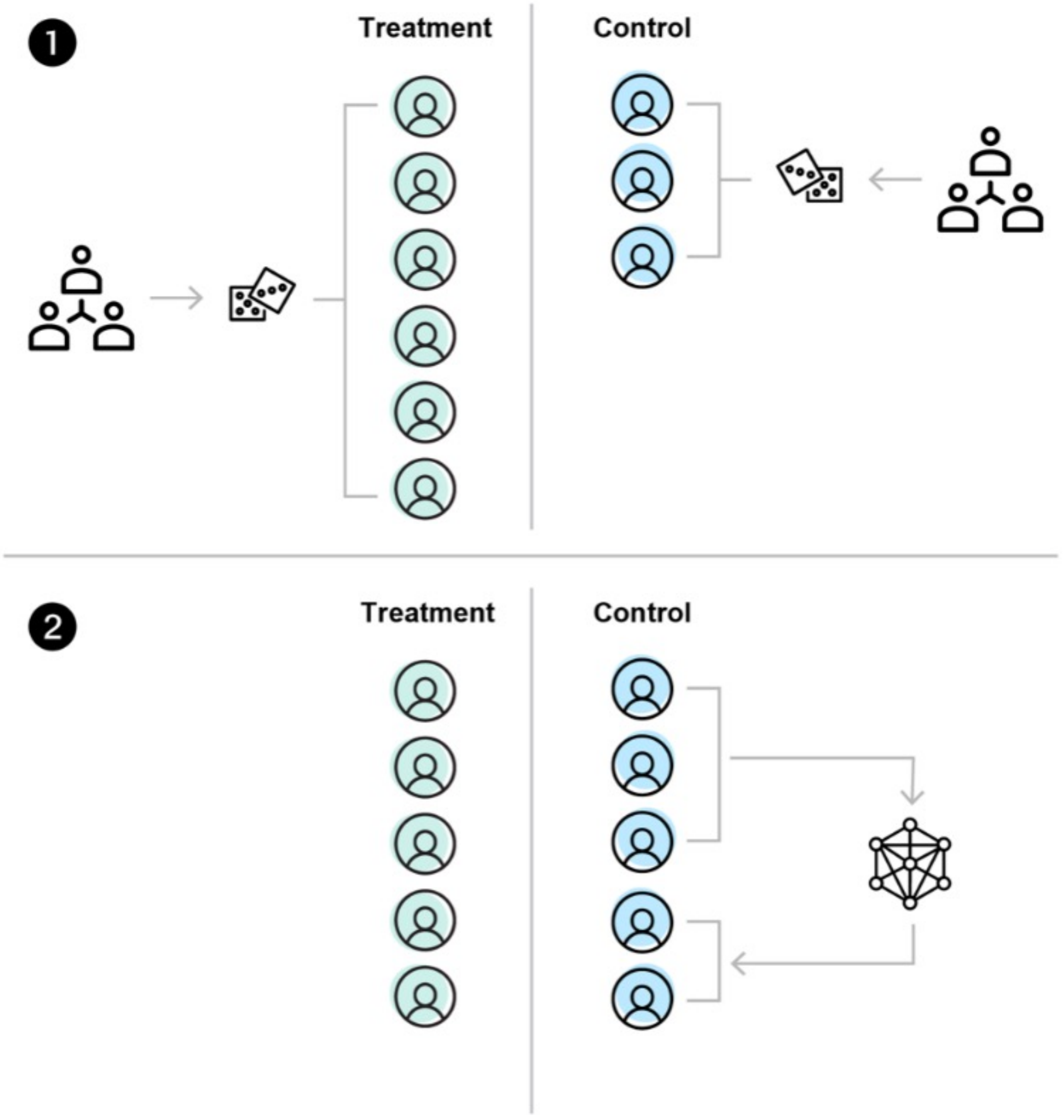


Data augmentation is when we use SDG to simulate virtual patients to add to an existing core dataset



Virtual Patients - I

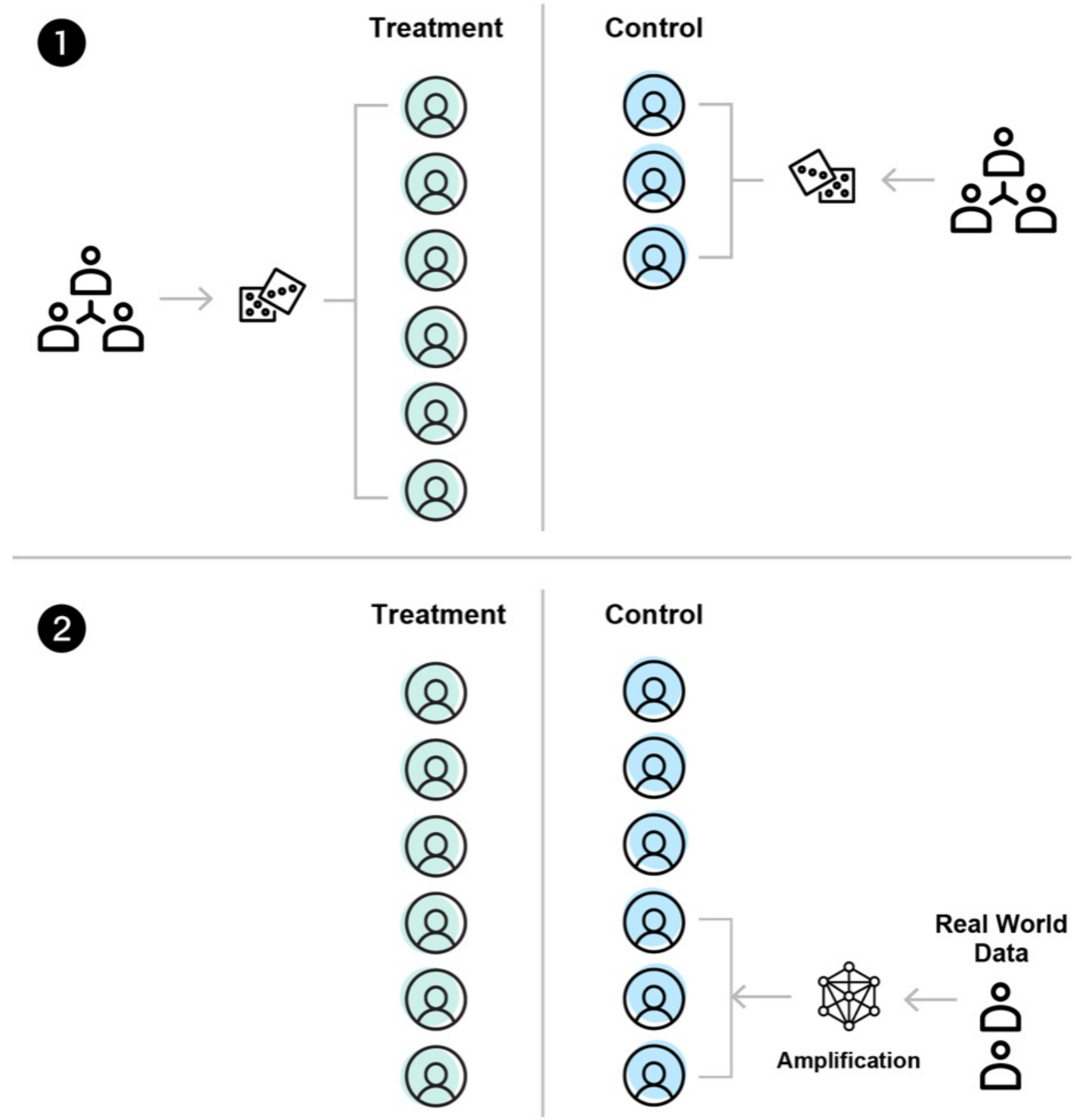
Virtual patients can be simulated to reduce recruitment or to rescue studies with low recruitment or high attrition



Replica Analytics
AN AETION COMPANY

Virtual Patients - II

Real-world data can be amplified to create synthetic external controls, especially when there are insufficient RWD or RWD diversity



Replica Analytics
AN AETION COMPANY


QUESTIONS