Data Synthesis
=
Future of Data Sharing
?

Khaled El Emam

13th September 2022

kelemam@replica-analytics.com

Agenda

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

Privacy & Utility 2 An overview of the evidence on privacy risks and utility of synthetic data

Regulatory Questions 3 Addressing some of the common questions that are asked by regulators

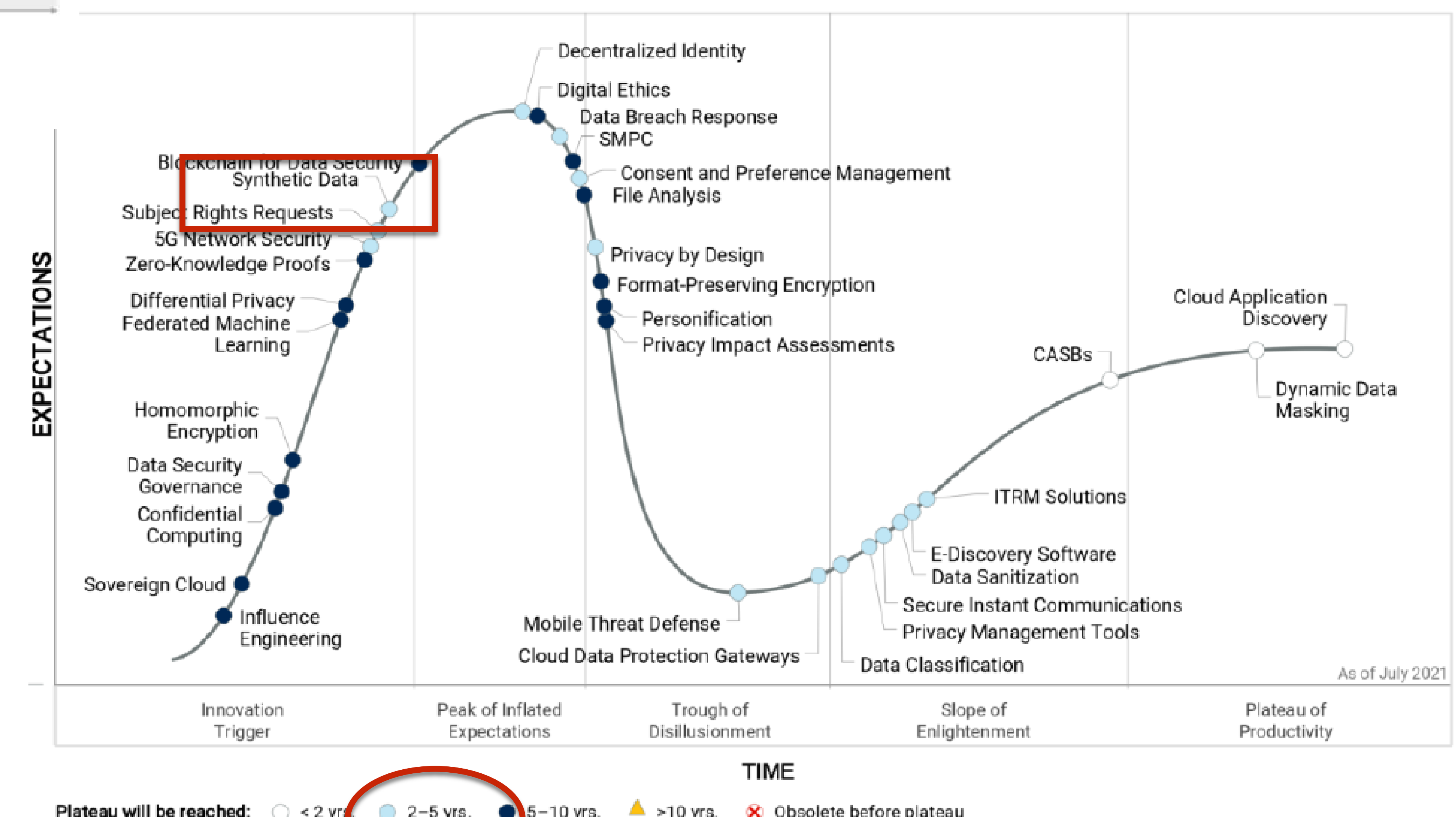
Implementation Questions 4 What are the next steps for implementing data synthesis in an organization

Replica Analytics
AN AETION COMPANY

The adoption of synthetic data has been accelerating quite rapidly

2020

2021



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

Top 10 Strategic Predictions for 2022 and Beyond

Data	Tracking	Behavior	Supervision	Talent
70% reduction in privacy sanctions	**40%** intentionally devalue personal data	**25%** neuromine at scale	**30%** teams without a boss	**30%** increase in talent across Africa
2025	2024	2027	2024	2026

Composability	Cyber Attack	Customers	Crypto	Digital
80% report better business performance	**G20** cyber attack breeds kinetic response	**75%** companies “break up” with customers	**NFTs** drive high value companies	**1 Billion** poorest people get internet
2024	2024	2025	2026	2027

gartner.com

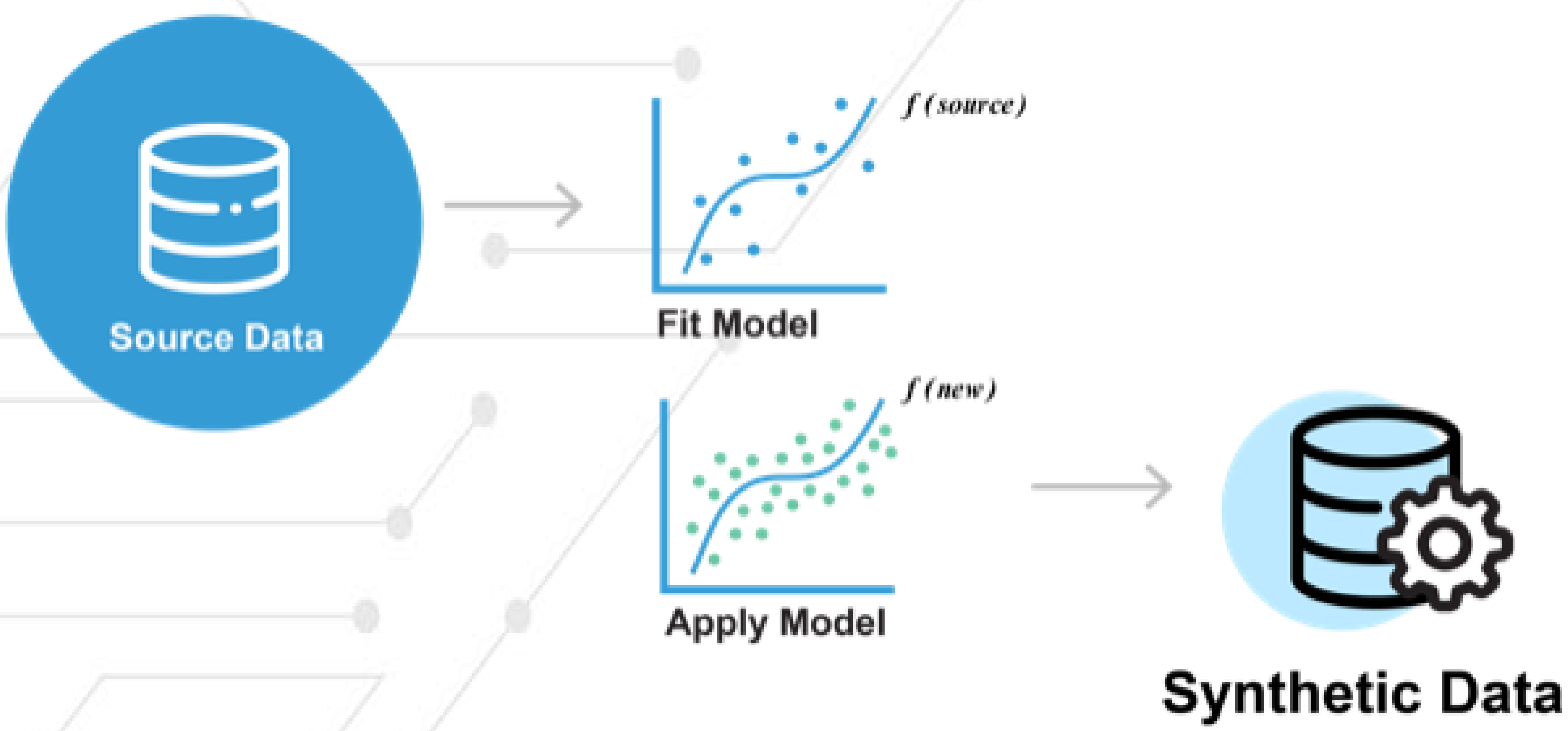
Source: Gartner

Gartner

Replica Analytics

AN AETION COMPANY

The Synthesis Process



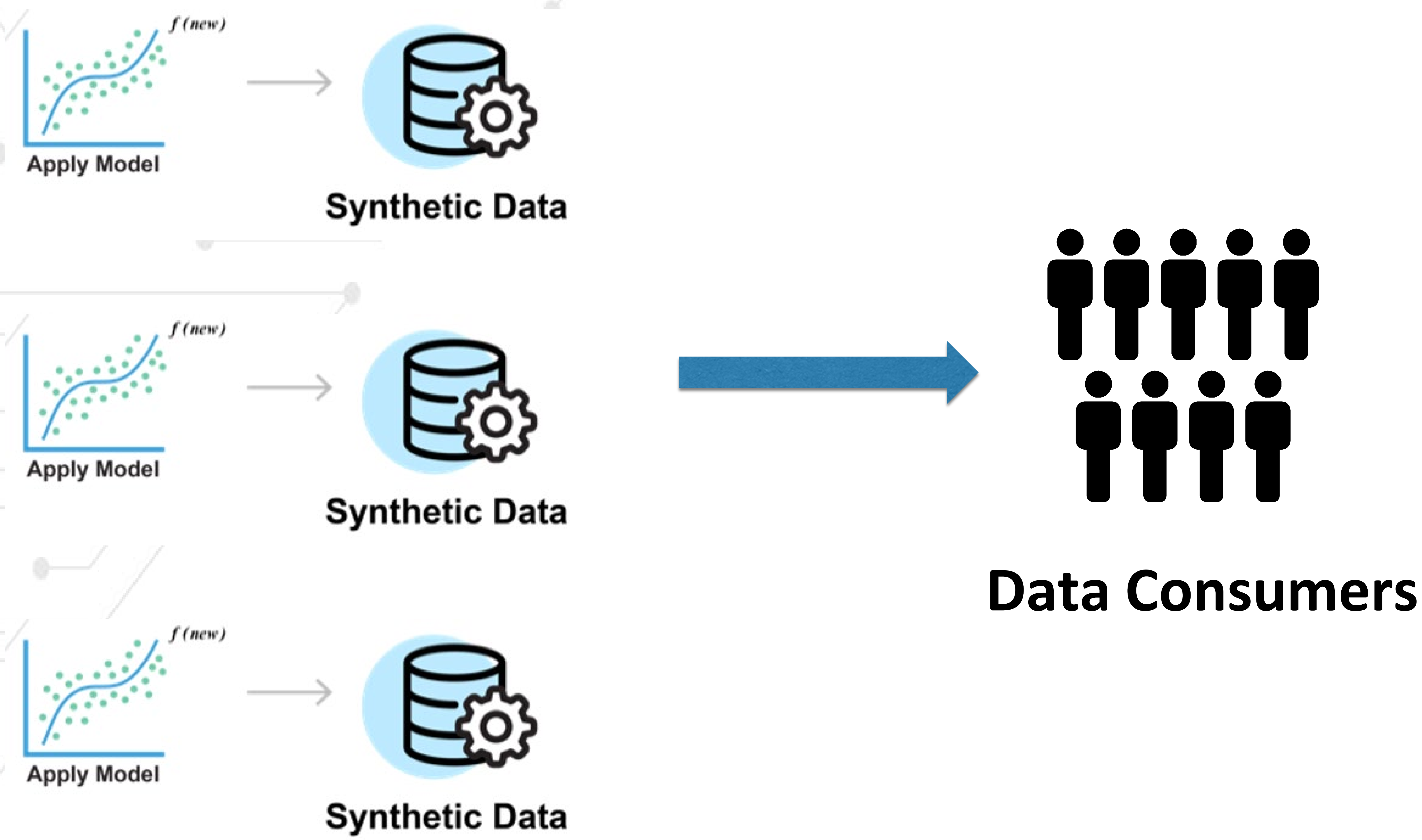
COU1A	AGECAT	AGELE70	WHITE	MALE	BMI
United States	2	1	1	1	33.75155
United States	2	1	1	0	39.24707
United States	1	1	1	0	26.5625
United States	4	1	1	1	40.58273
United States	5	0	0	1	24.42046
United States	5	0	1	0	19.07124
United States	3	1	1	1	26.04938
United States	4	1	1	1	25.46939

<u>Common Clarifications</u>

- The source datasets can be as small as 100 or 150 patients. We have developed generative modeling techniques that will work for small datasets.
- The source datasets can be very large – then it becomes a function of compute capacity that is available.
- It is not necessary to know how the synthetic data will be analyzed to build the generative models. The generative models capture many of the patterns in the source data.

Replica Analytics
AN AETION COMPANY

Simulator Exchange



Common use cases for synthetic data generation

Privacy

- Software testing
- Internal data reuse (analytics)
- External data sharing
- Vendor assessment
- Training / education

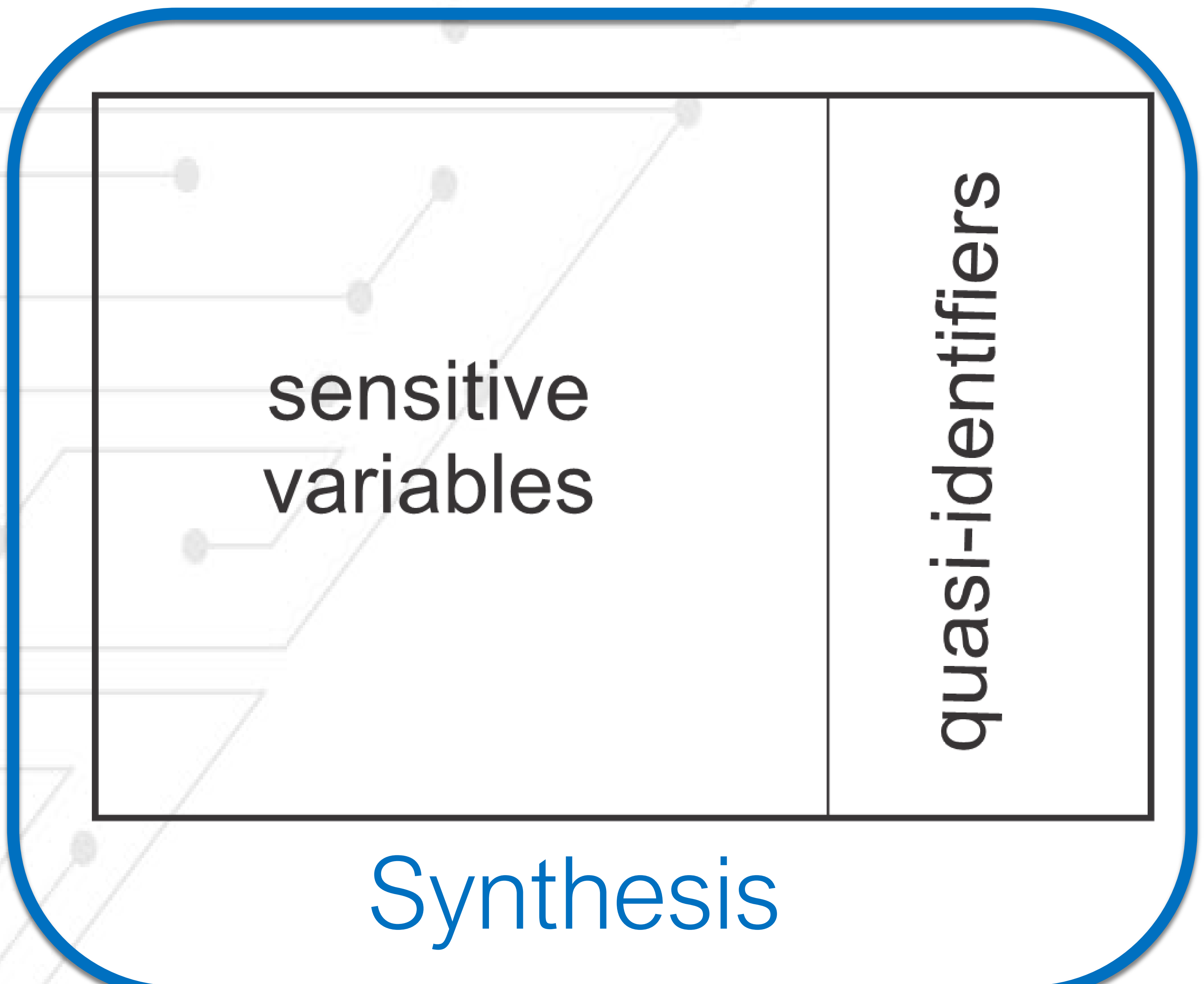
Data Enhancement

- Augmenting / amplifying small datasets (e.g., rare disease datasets)
- Compensating for under-represented groups in a dataset by simulating additional patients

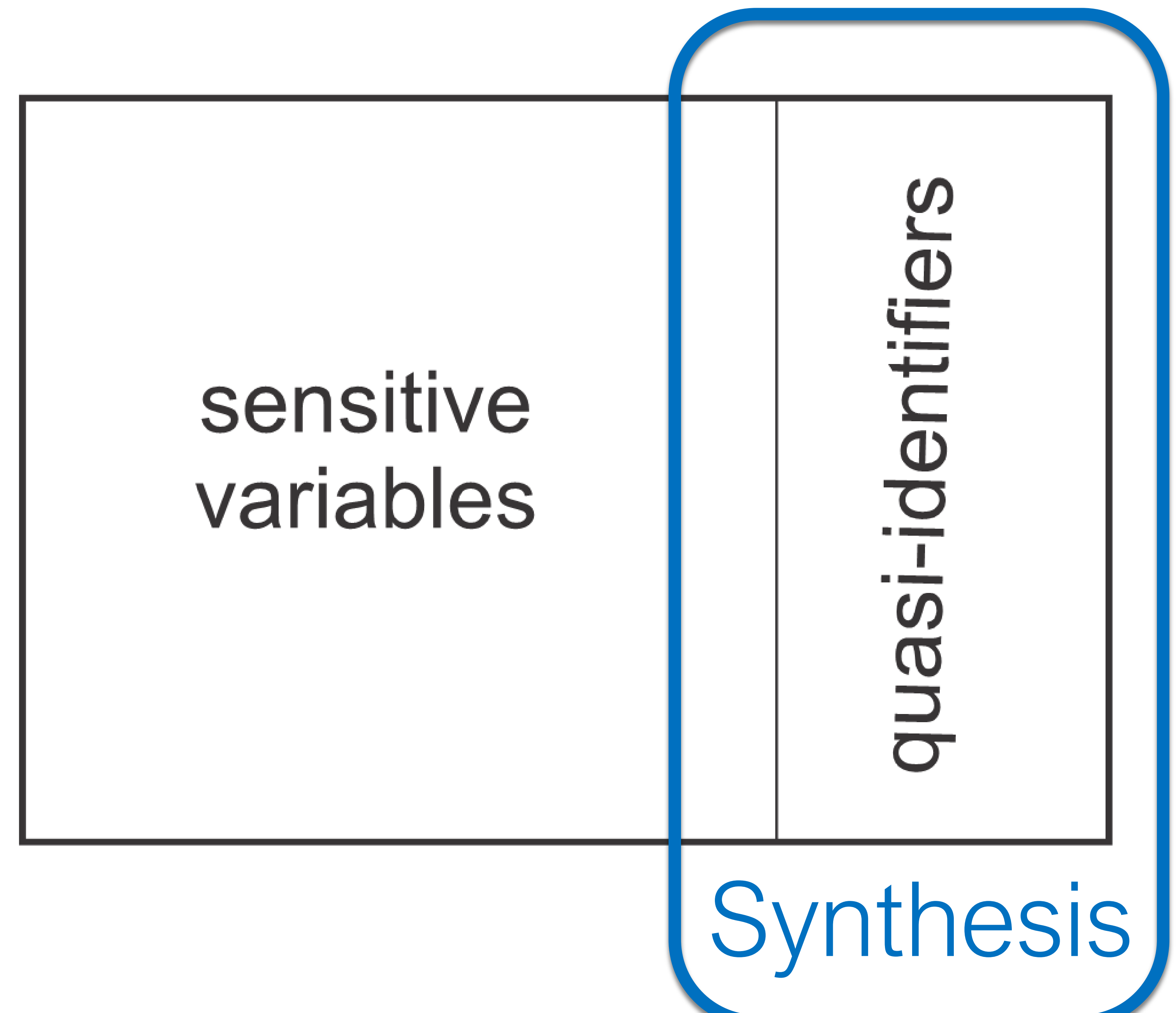
Replica Analytics
AN AETION COMPANY

Two Synthesis Strategies

Full Synthesis
Synthesize all variables



Partial Synthesis
Synthesize quasi-identifiers



Replica Analytics
AN AETION COMPANY

Operating models for secondary analysis using synthetic data

- Sharing synthetic data and conclusions are drawn from the analysis of synthetic datasets
- Make synthetic data available for exploratory analysis and if there are interesting results, make a request for the full dataset (which may be a long and complicated process, but at least there is confidence that there are interesting results)
- Perform the analysis on the synthetic data and then submit the analysis code (R, SAS, Python, …) to be executed on the real dataset behind a firewall

Replica Analytics

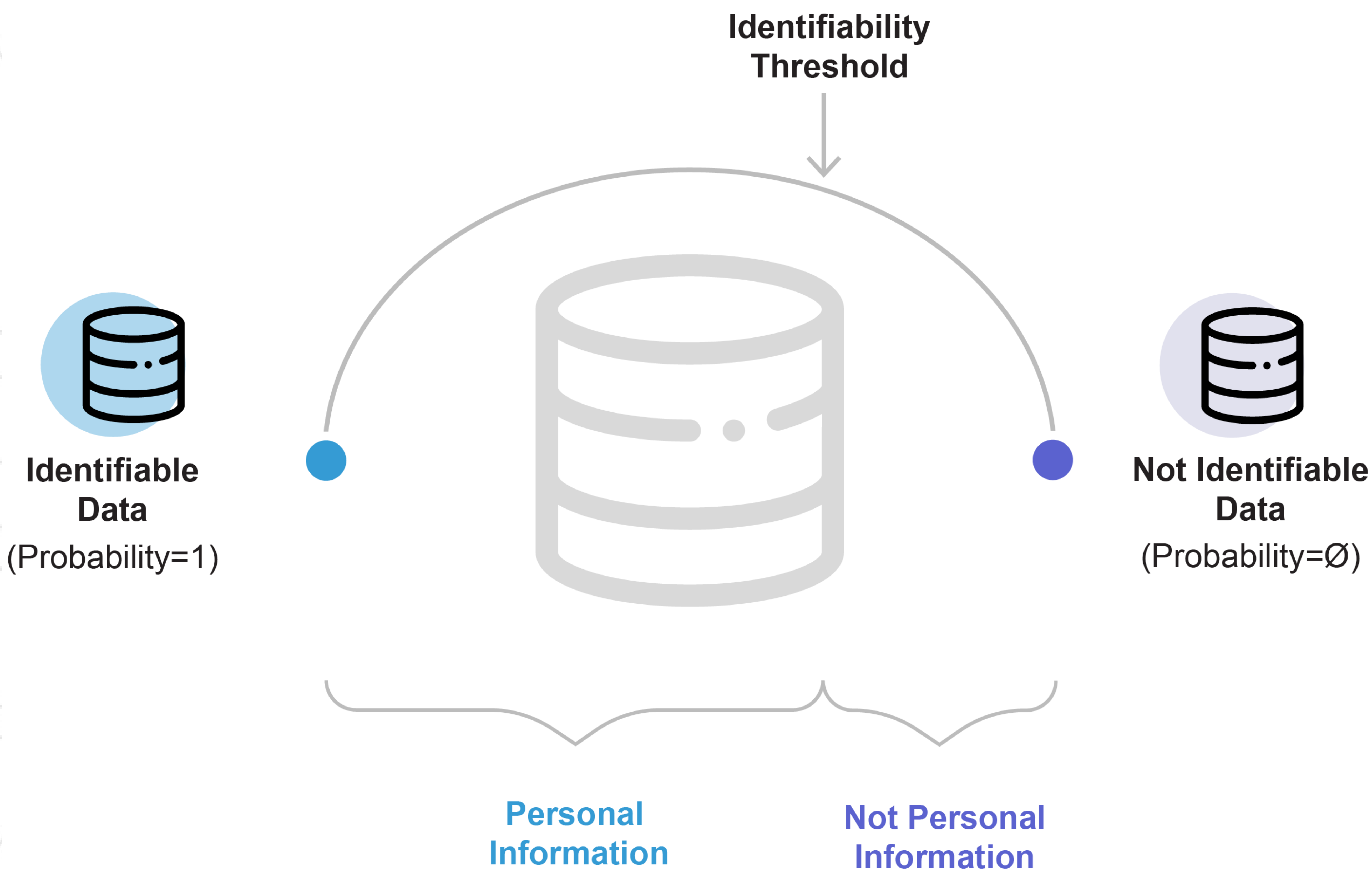
AN AETION COMPANY

Additional risks that may be relevant depending on the privacy enhancing technology that is being used

- Identity disclosure – generally low for synthetic data
- Attribution disclosure – needs to be evaluated for synthetic data
- Membership disclosure – needs to be evaluated for synthetic data

Replica Analytics
AN AETION COMPANY

Identifiability Spectrum



Replica Analytics
AN AETION COMPANY

Example of evaluating attribution disclosure

Dataset	Fully Synthetic Data	Original Data
Washington Hospital Data	0.0197	0.098
Canadian COVID-19 Data	0.0086	0.034

A commonly used risk threshold = 0.09

Replica Analytics
AN AETION COMPANY

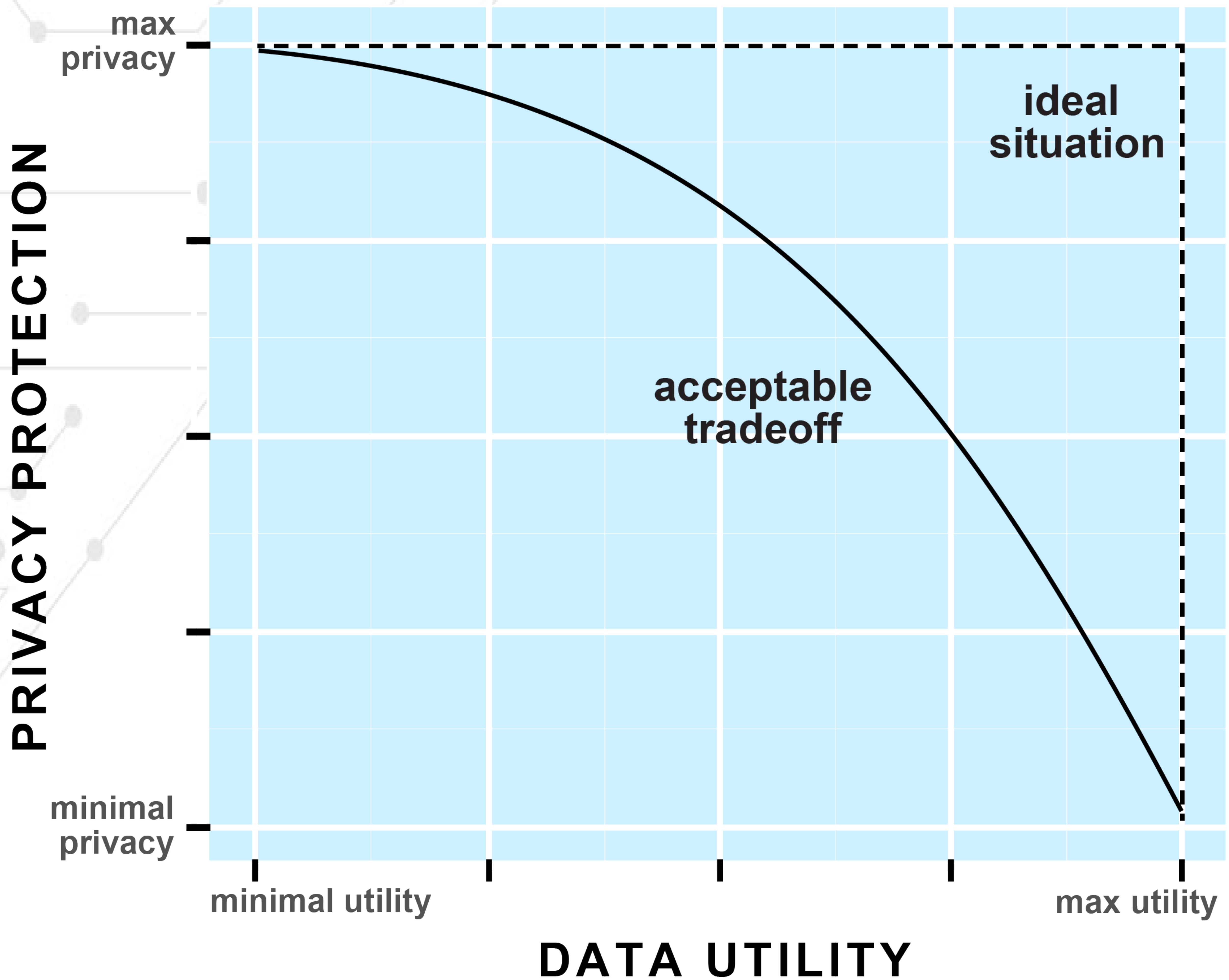
Example of evaluating membership disclosure

Dataset	Dataset size	Risk
Trial #1 (NCT00041197): National Cancer Institute	773	-1.42
Trial #2 (NCT01124786): Clovis Oncology	367	-0.0137
Trial #3 (NCT00688740): Sanofi	746	-0.034
Trial #4 (NCT00113763): Amgen	370	-0.0137
Trial #5 (NCT00460265): Amgen	520	-0.0947
Trial #6 (NCT00119613): Amgen	479	-0.0322
Trial #7 (N0147)	1543	0.052

A commonly used risk threshold = 0.2

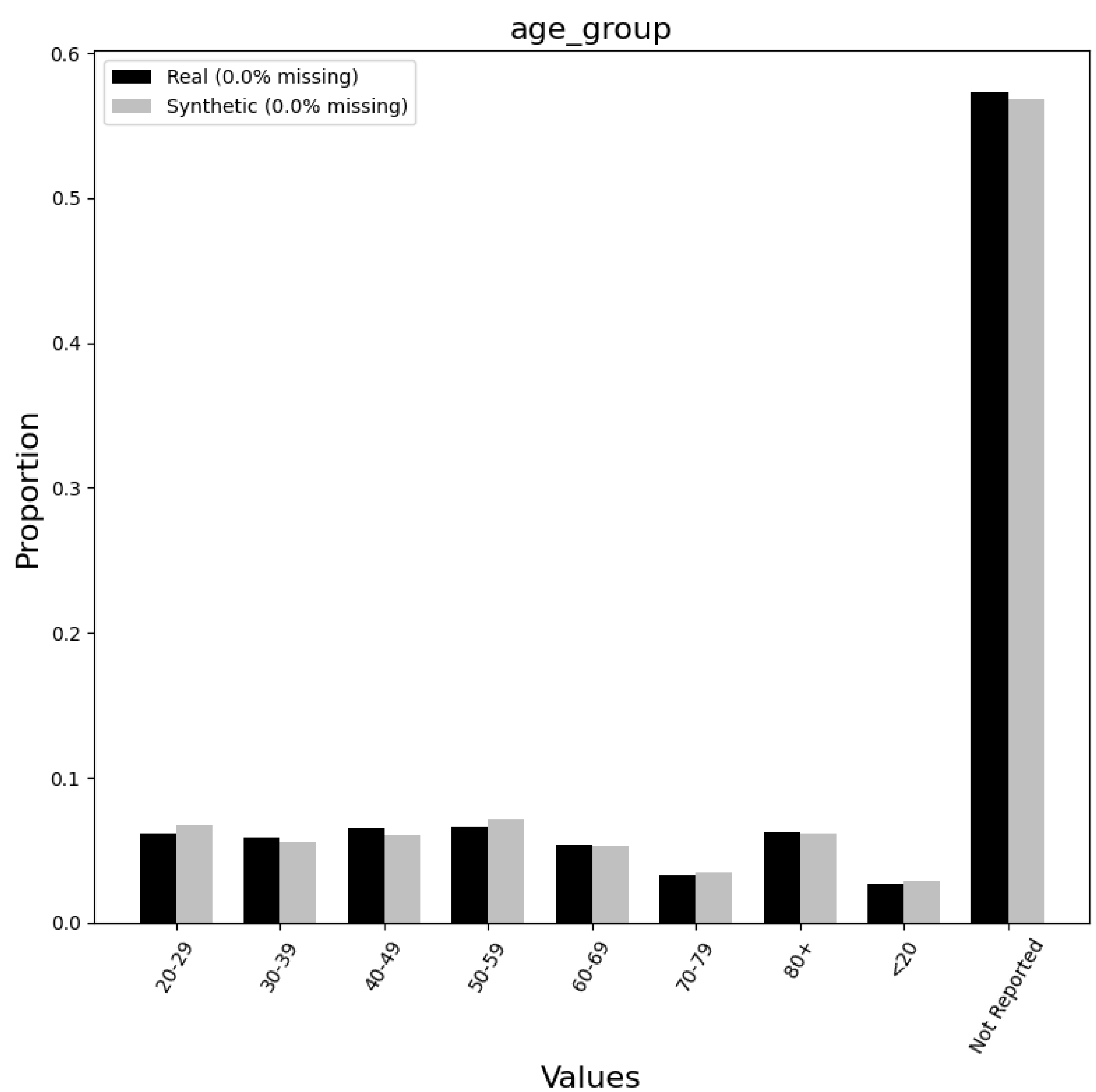
Replica Analytics
AN AETION COMPANY

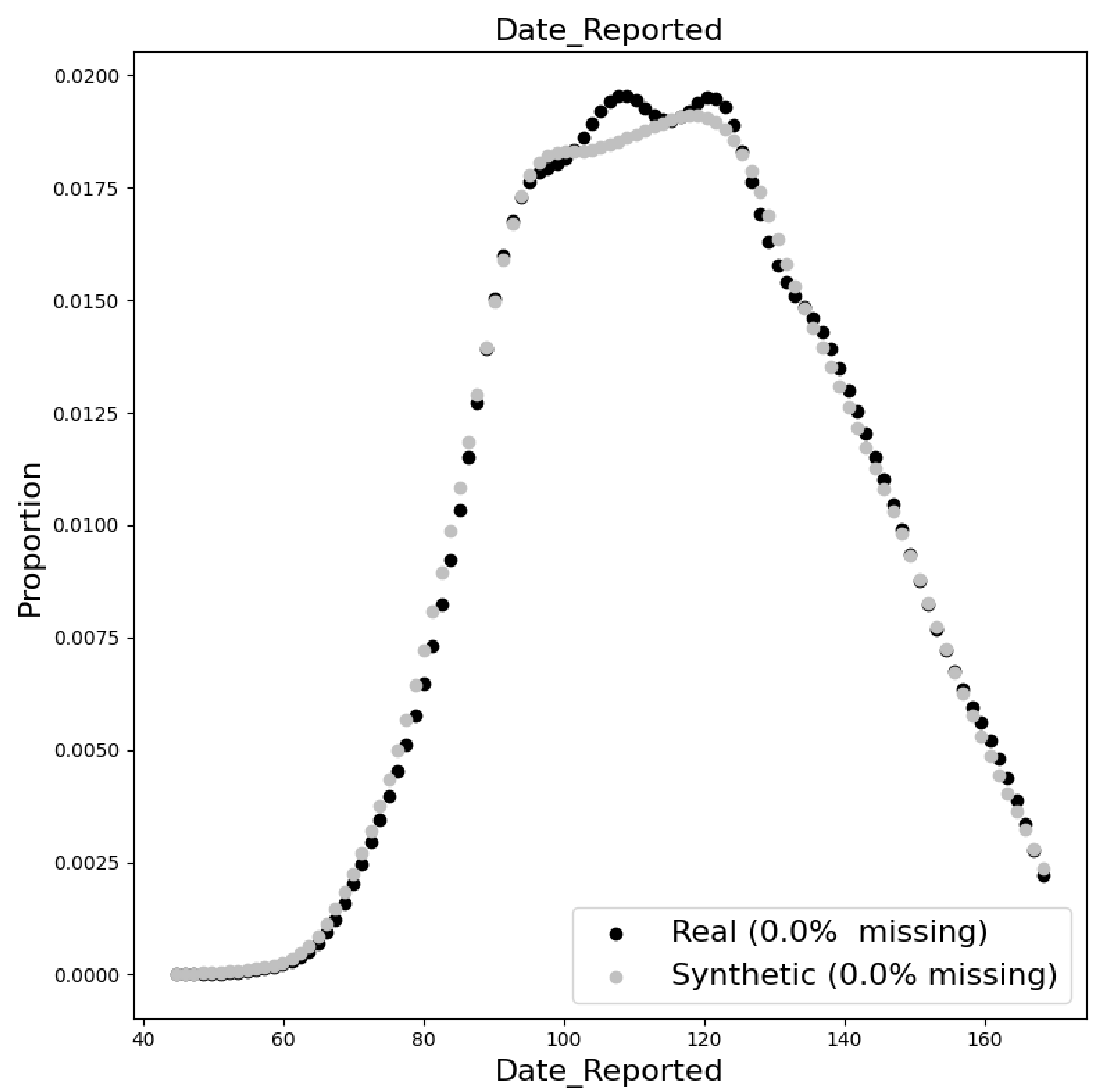
Privacy-Utility Trade-off



Replica Analytics
AN AETION COMPANY

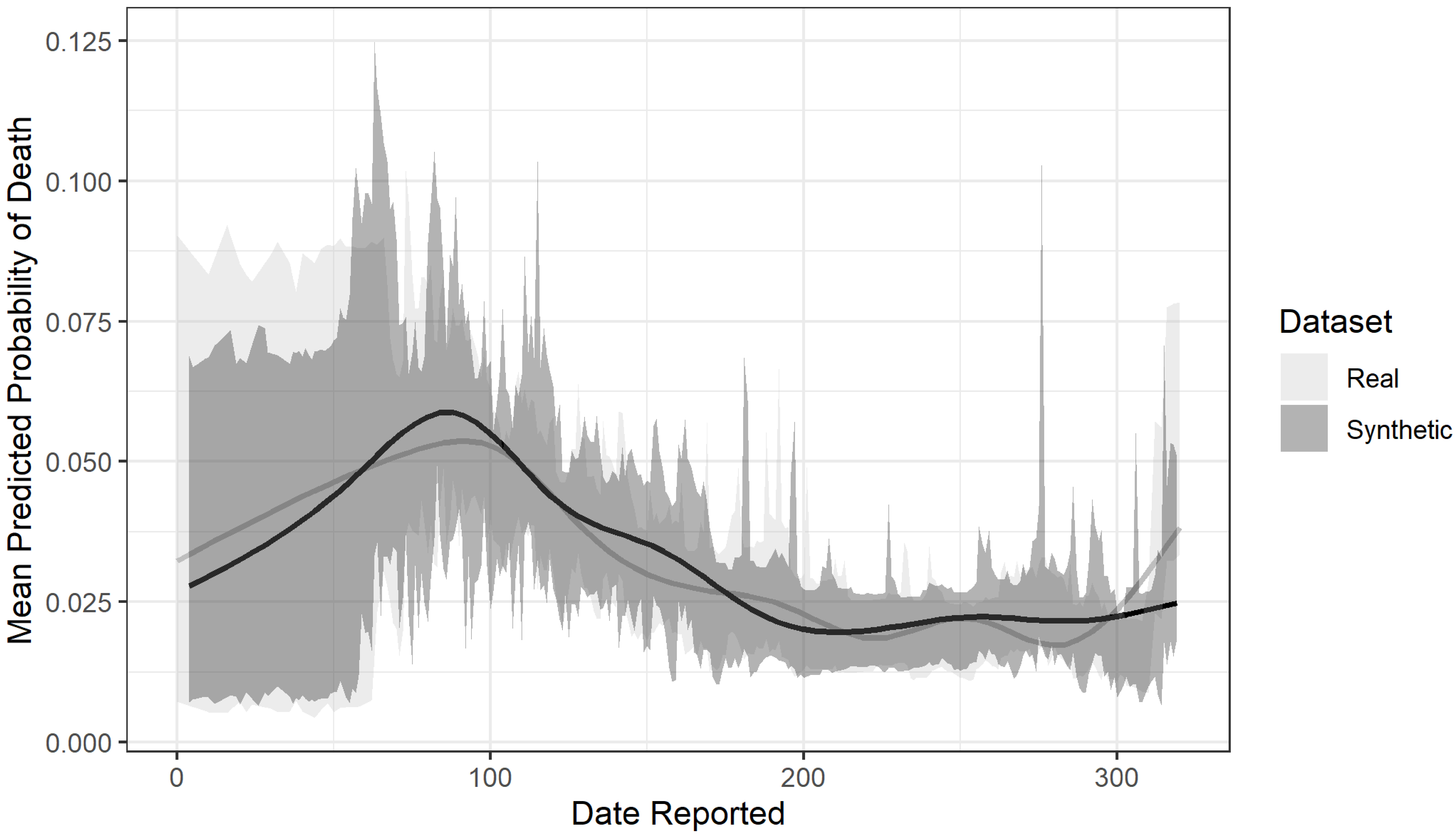
The distributions of real and synthetic datasets look similar





Replica Analytics
AN AETION COMPANY

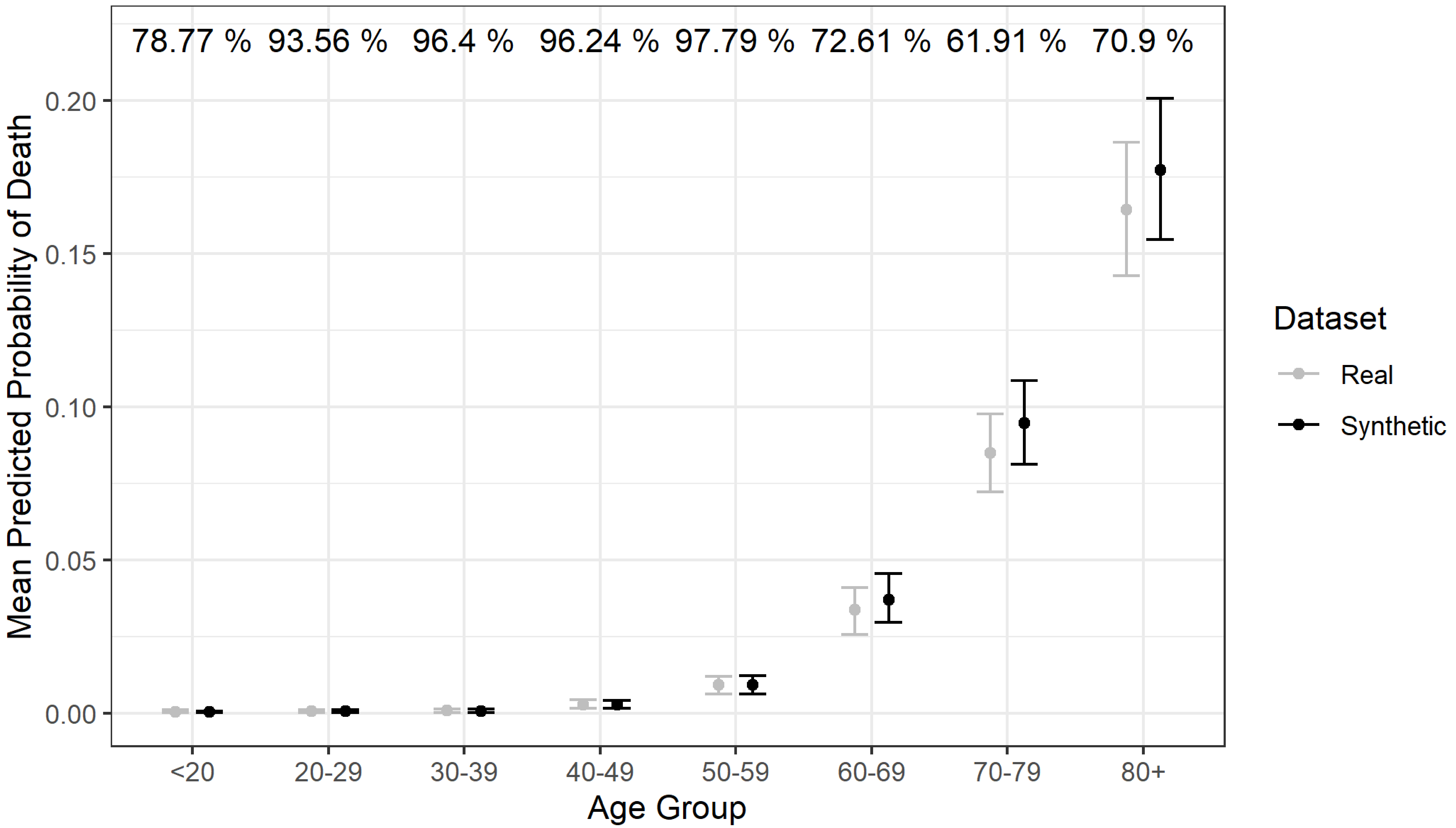
Comparing Real and Synthetic Data: Mortality Over Time



Replica Analytics

AN AETION COMPANY

Comparing Real and Synthetic Data: Mortality By Age



Replica Analytics
AN AETION COMPANY

There is rapid adoption and consequent interest in learning more about synthetic data generation by regulators

- CNIL allowing synthetic data generation as a form of data anonymization
- Norwegian DPA suggesting synthetic data for software testing
- EDPS organizing an IPEN event on synthetic data
- Canadian OPC funding a project on regulating synthetic data through contributions program

Replica Analytics
AN AETION COMPANY

Risk-based Approach

Data Transformations + **Controls**

- Generalization
- Suppression
- Addition of noise
- Microaggregation

- Security controls
- Privacy controls
- Contractual controls

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

Skill Set

- Synthesis requires minimal skills in practice – it is a largely automated process
- On the other hand the skills needed to create non personal datasets using other methods are very specialized, take time to develop, and generally difficult to find cost-effectively

Replica Analytics
AN AETION COMPANY

Acceptance of Synthetic Data

- Privacy Regulators
  - Identifiability not the appropriate measure of risk, with some exceptions
  - Still new but indications are that this can be treated differently than previous approaches
- Data Scientists
  - Main concern is data utility – case studies will address that concern
  - Results thus far are promising

Replica Analytics
AN AETION COMPANY


QUESTIONS
cc: an untrained eye - https://www.flickr.com/photos/

Thank you

- Replica Analytics develops the Replica Synthesis software – generator of privacy protective synthetic health data and simulator exchange
  - For more information on our synthetic data solutions:
    - Visit our website www.replica-analytics.com
    - Message us via the website contact page

Replica Analytics
AN AETION COMPANY

Synthetic Data Generation References

- Y. Jiang, L. Mosquera, B. Jiang, L. Kong, and K. El Emam, “Measuring re-identification risk using a synthetic estimator to enable data sharing,” PLoS ONE, vol. 17, no. 6, p. e0269097, Jun. 2022.
- K. El Emam, L. Mosquera, X. Fang, and A. El-Hussuna, “Utility Metrics for Evaluating Synthetic Health Data Generation Methods: Validation Study,” JMIR Medical Informatics, vol. 10, no. 4, p. e35734, Apr. 2022.
- S. James, C. Harbron, J. Branson, and M. Sundler, “Synthetic data use: exploring use cases to optimise data utility,” Discov Artif Intell, vol. 1, no. 1, p. 15, Dec. 2021, doi: 10.1007/s44163-021-00016-y.
- Z. Azizi, C. Zheng, L. Mosquera, L. Pilote, K. El Emam: “Replicating Secondary Studies Using Synthetic Clinical Trial Data”, *BMJ Open*, 11:e043497, 2021.
- K. El Emam, L. Mosquera, E. Jonker, H. Sood: “Evaluating the Utility of Synthetic COVID-19 Case Data”, *JAMIA Open*, 14(1):ooab012, January 2021.
- K. El Emam, L. Mosquera, and C. Zheng, “Optimizing the synthesis of clinical trial data using sequential trees,” *JAMIA*, 28(1): 3-13, 2021.
- K. El Emam, L. Mosquera, and J. Bass, “Evaluating Identity Disclosure Risk in Fully Synthetic Health Data: Model Development and Validation,” *JMIR*, vol. 22, no. 11, Nov. 2020.
- K. El Emam, L. Mosquera, and R. Hoptroff, Practical Synthetic Data Generation: Balancing Privacy and the Broad Availability of Data. O’Reilly, 2020.
- K. El Emam, “Seven Ways to Evaluate the Utility of Synthetic Data,” *IEEE Security and Privacy*, July/August, 2020.

Replica Analytics

AN AETION COMPANY