Data Synthesis:
A Tool for Responsible Data Sharing

Khaled El Emam

16th June 2021

Agenda

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

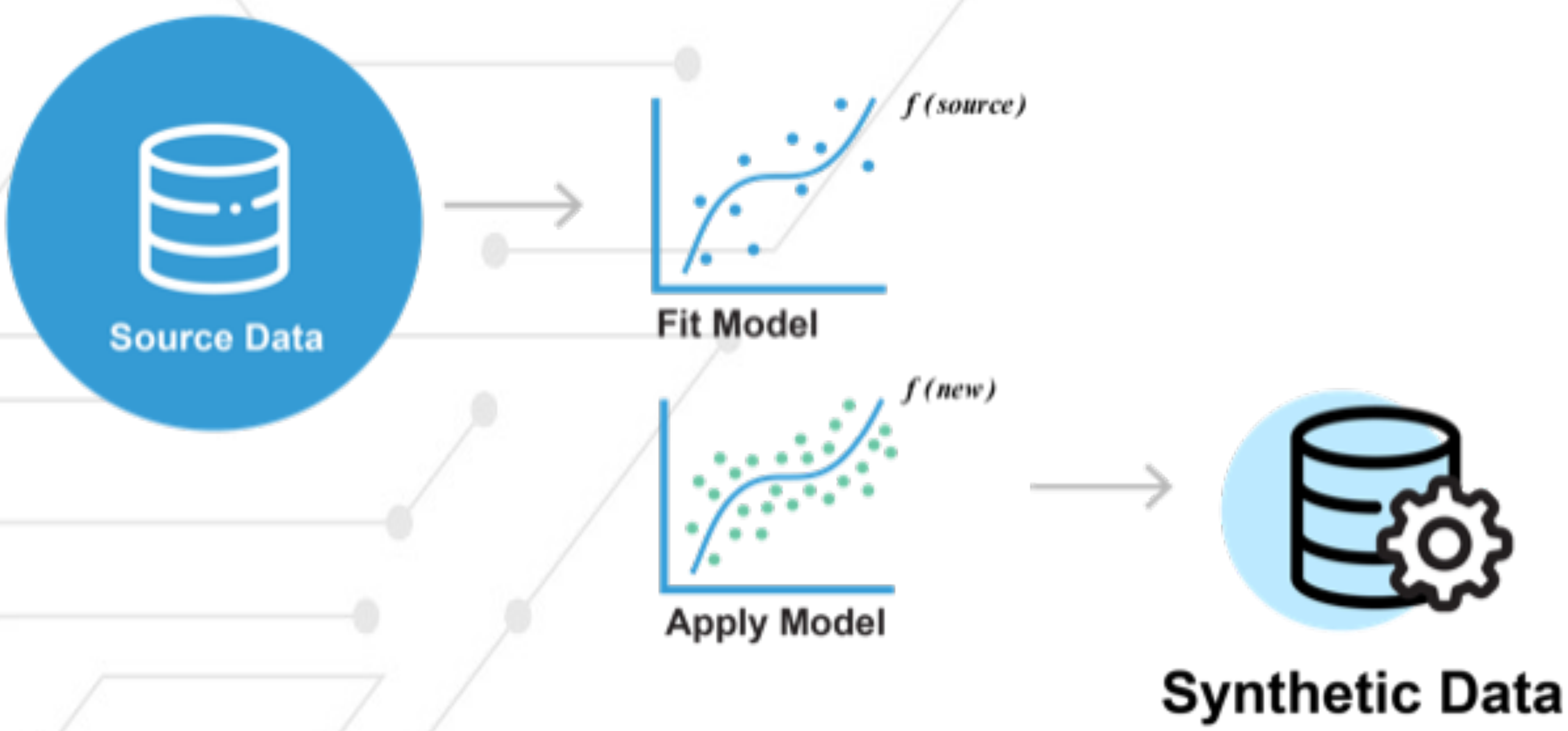
Privacy and Utility 2 An examination of privacy risks and the utility of synthetic data

Methods 3 A brief look at methods for the generation of synthetic data

Synthetic Data Uses

- Data Sharing and Data Access
  - AI and data science projects
  - Software testing
  - Proof of concept and technology evaluations
  - Open data/open science
  - Hackathons and data competitions/challenges
- Data Amplification and Data Augmentation
  - Amplifying small datasets
  - Correct bias

The Synthesis Process

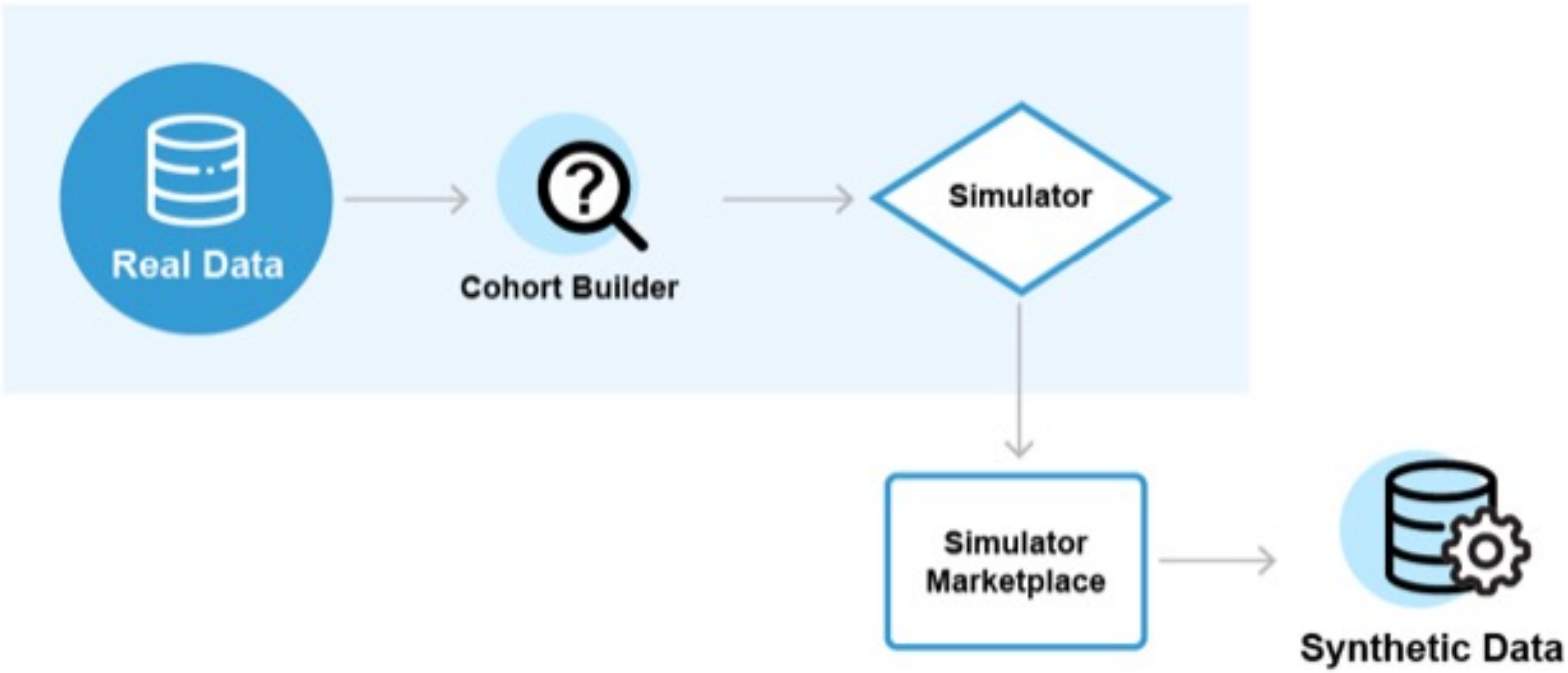


COU1A	AGECAT	AGELE70	WHITE	MALE	BMI
United States	2	1	1	1	33.75155
United States	2	1	1	0	39.24707
United States	1	1	1	0	26.5625
United States	4	1	1	1	40.58273
United States	5	0	0	1	24.42046
United States	5	0	1	0	19.07124
United States	3	1	1	1	26.04938
United States	4	1	1	1	25.46939

Electronic Health Information Laboratory

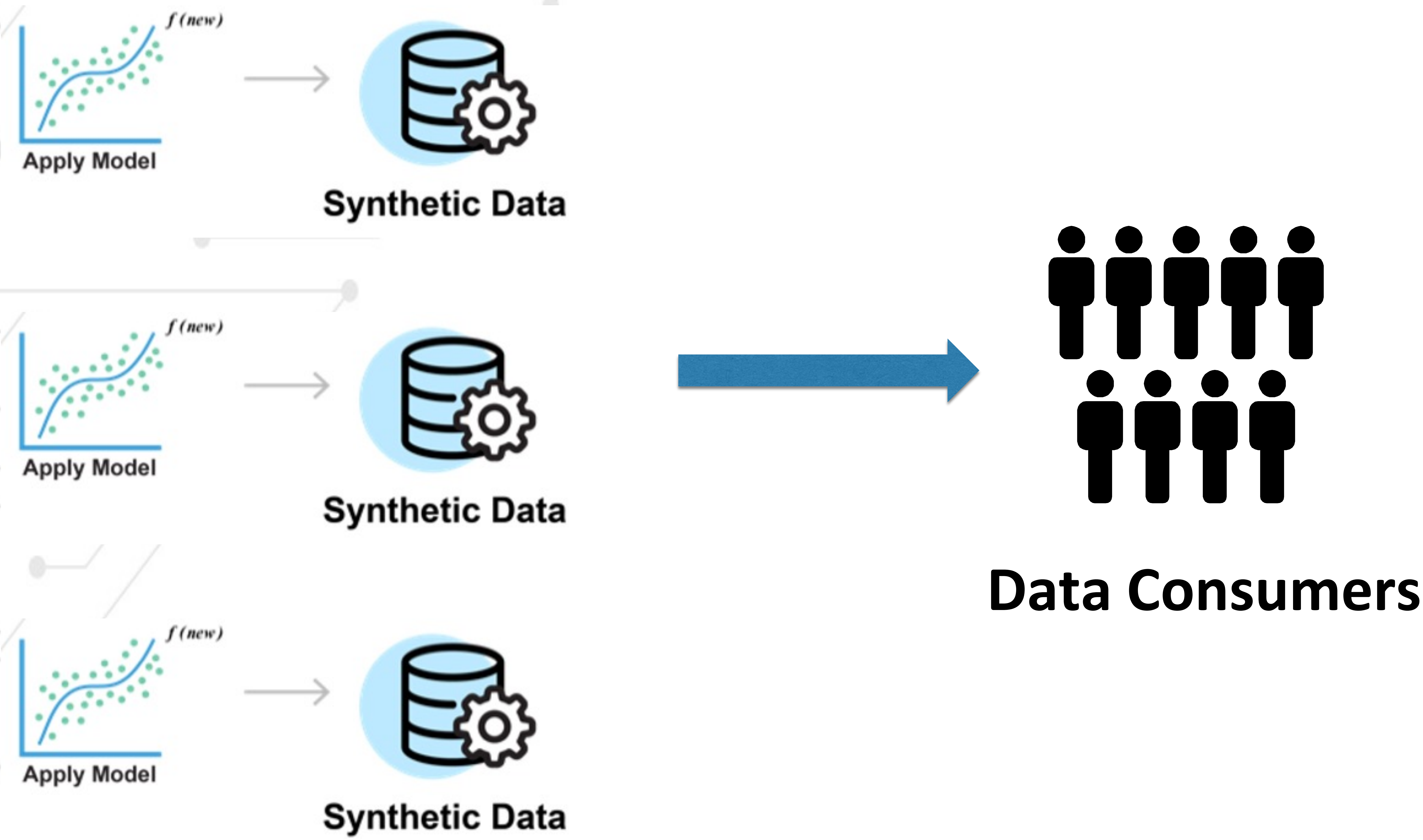
Replica Analytics

Data Simulator



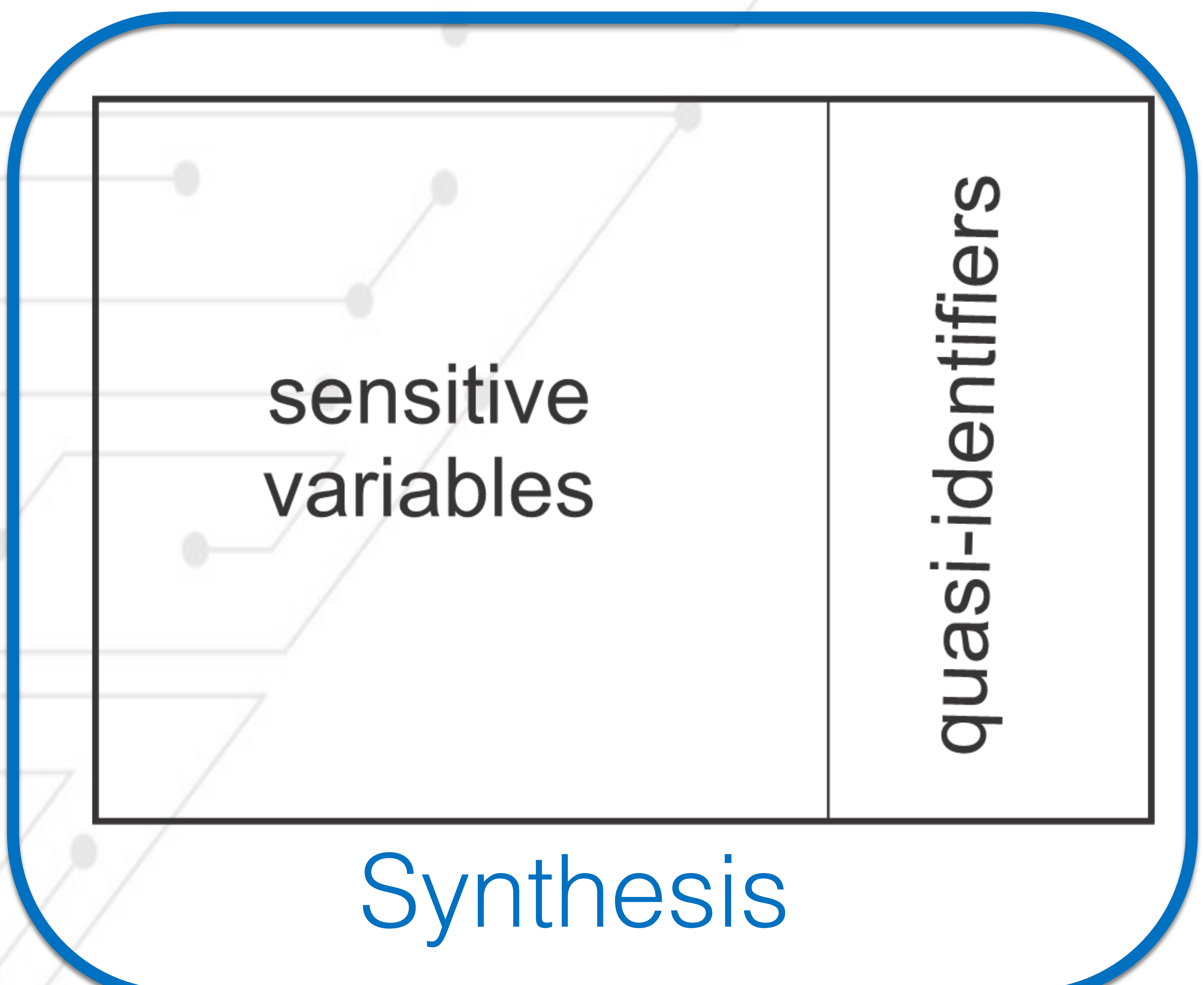
Allows generation of synthetic data without direct access to real data

Simulator Exchange

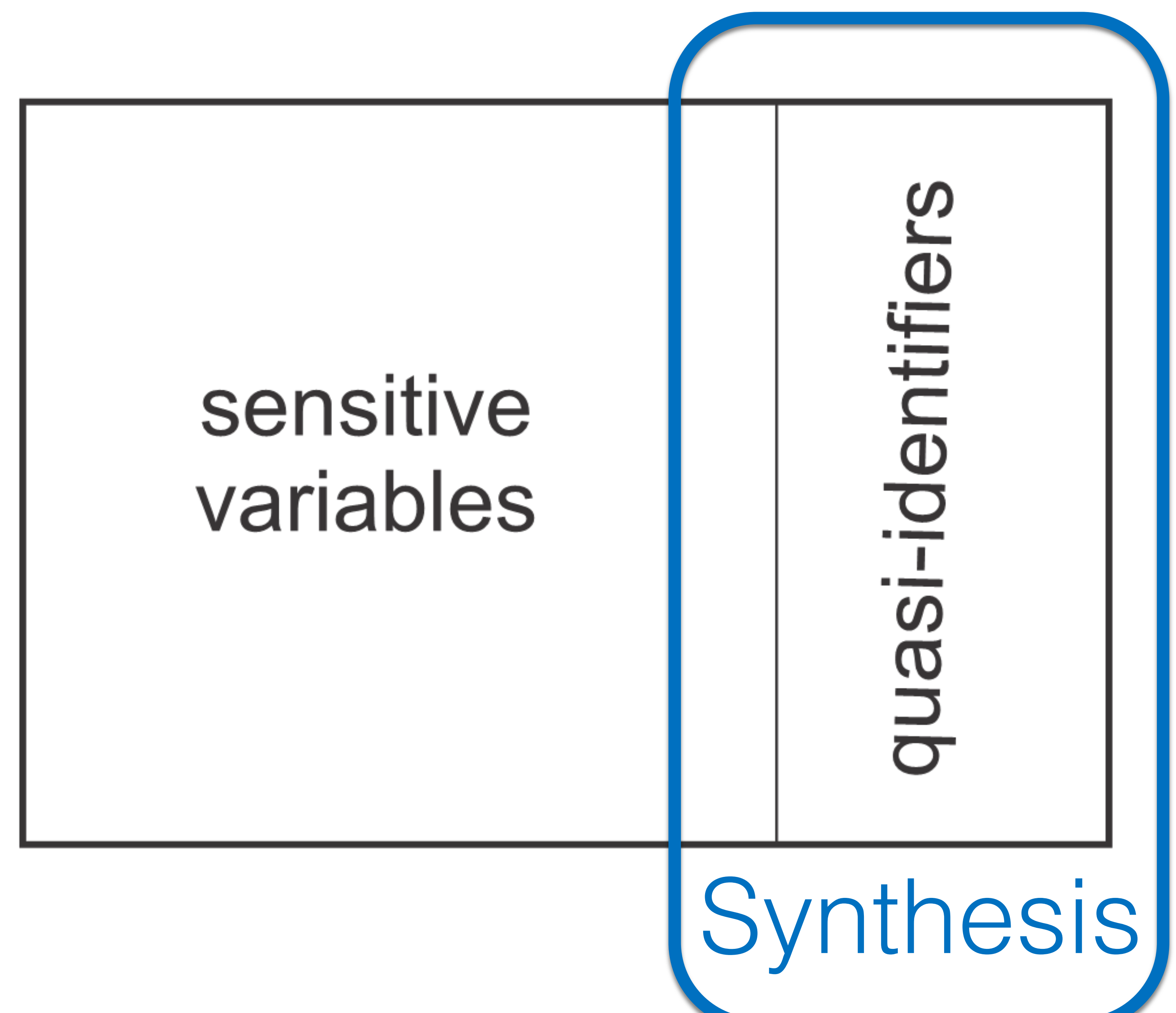


Two Synthesis Strategies

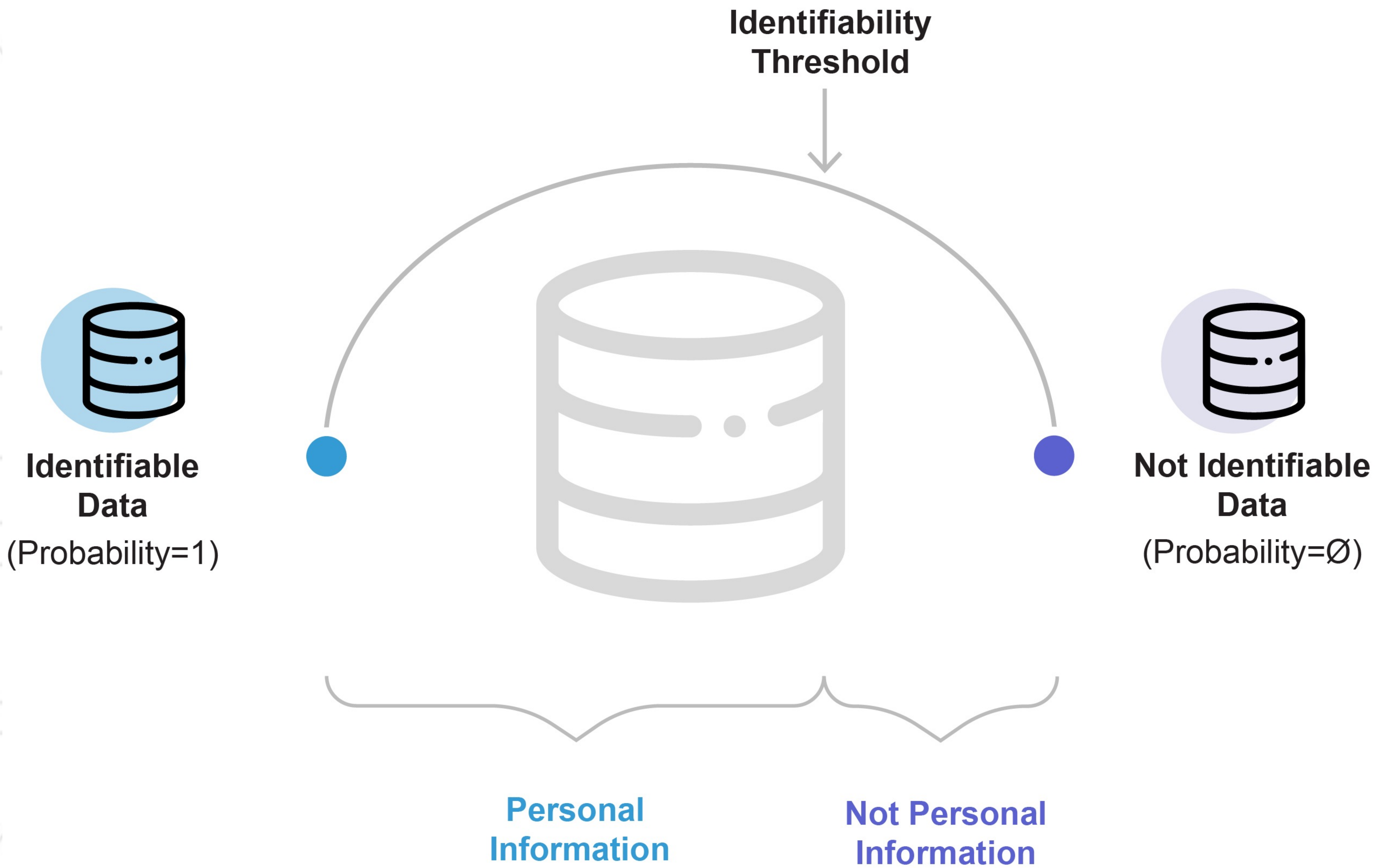
Full Synthesis
Synthesize all variables



Partial Synthesis
Synthesize quasi-identifiers



Identifiability Spectrum

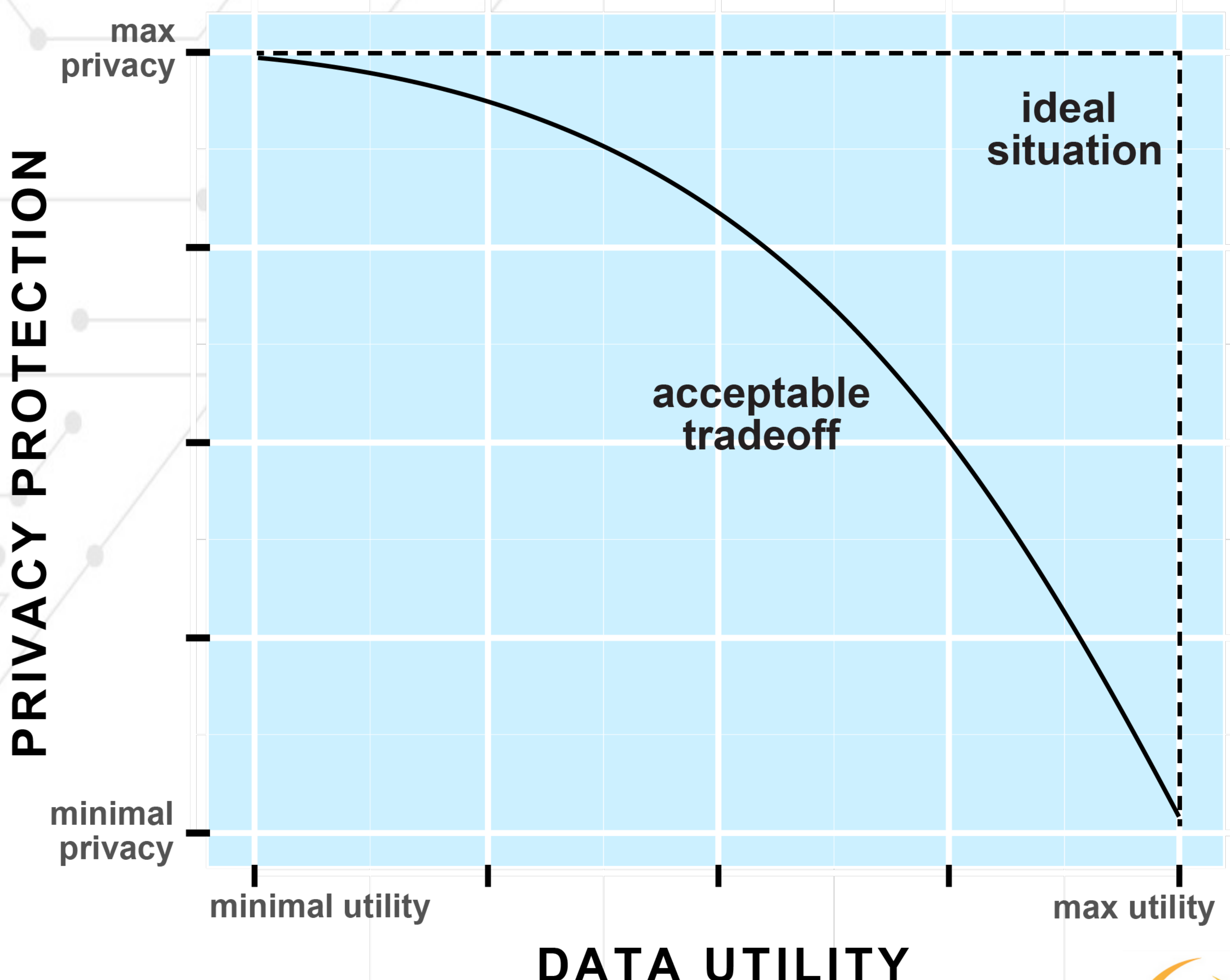


Privacy Risks

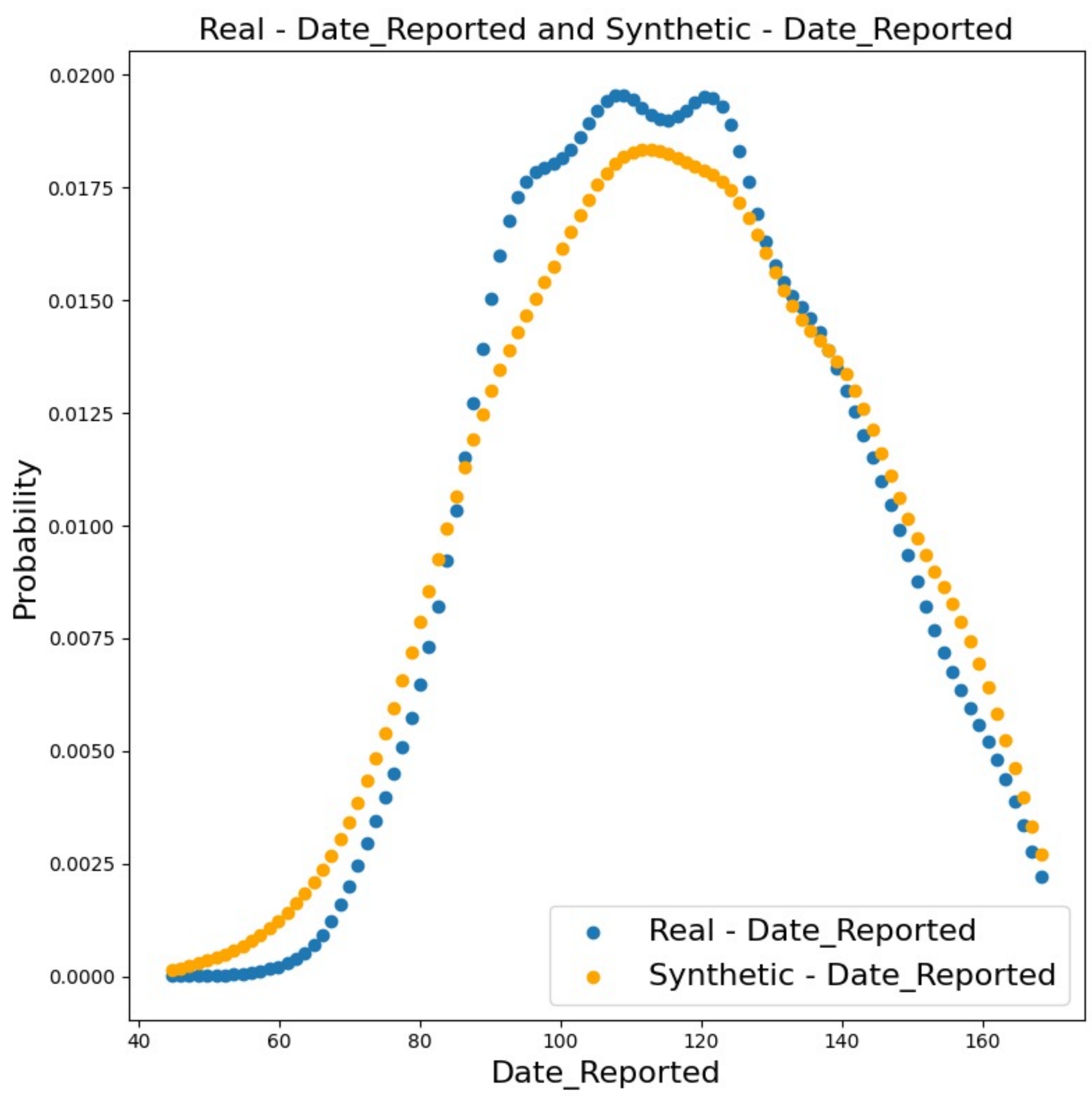
Dataset	Fully Synthetic Data	Original Data
Washington Hospital Data	0.0197	0.098
Canadian COVID Data	0.0086	0.034

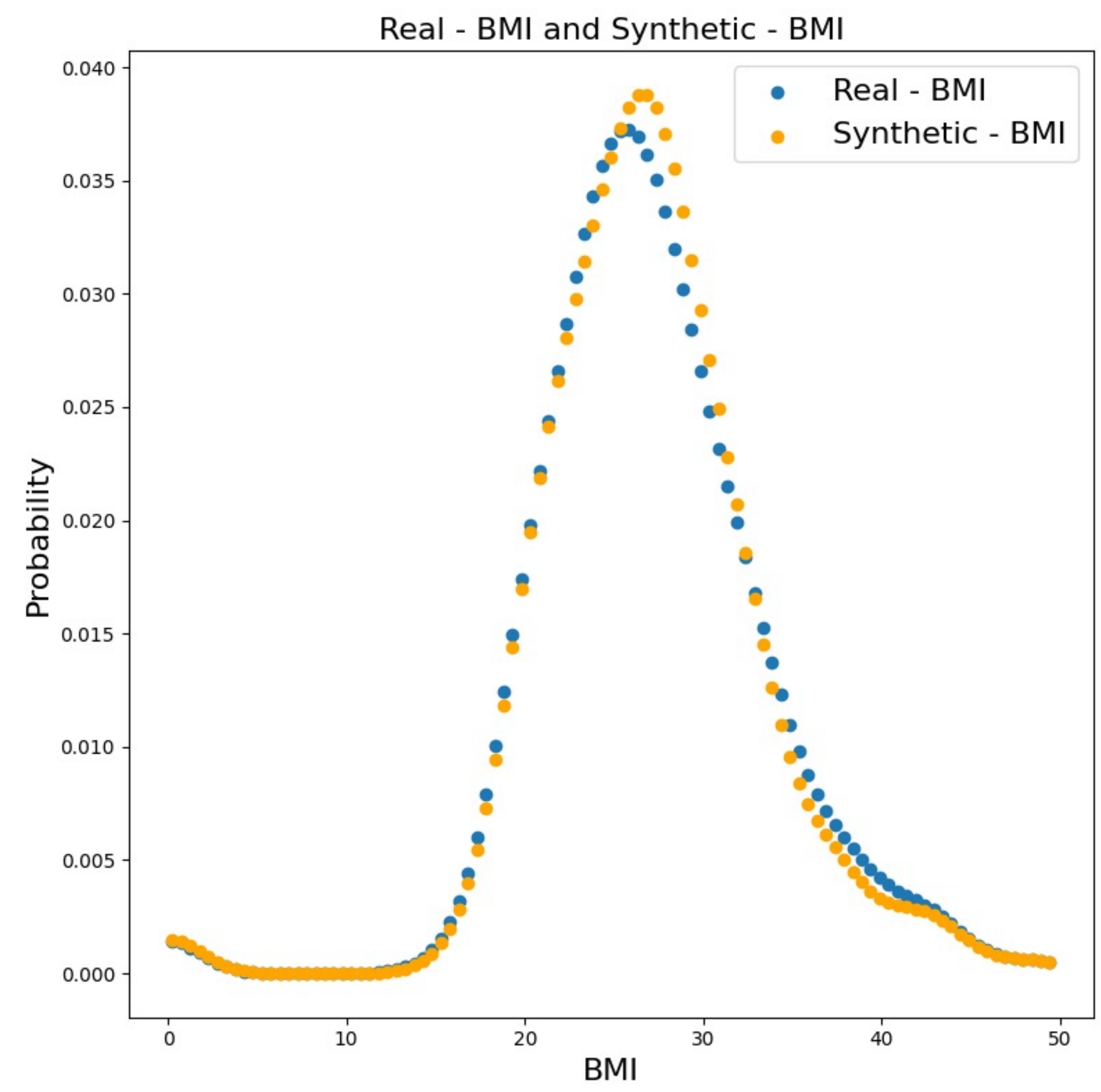
A commonly used risk threshold = 0.09

Privacy-Utility Tradeoff

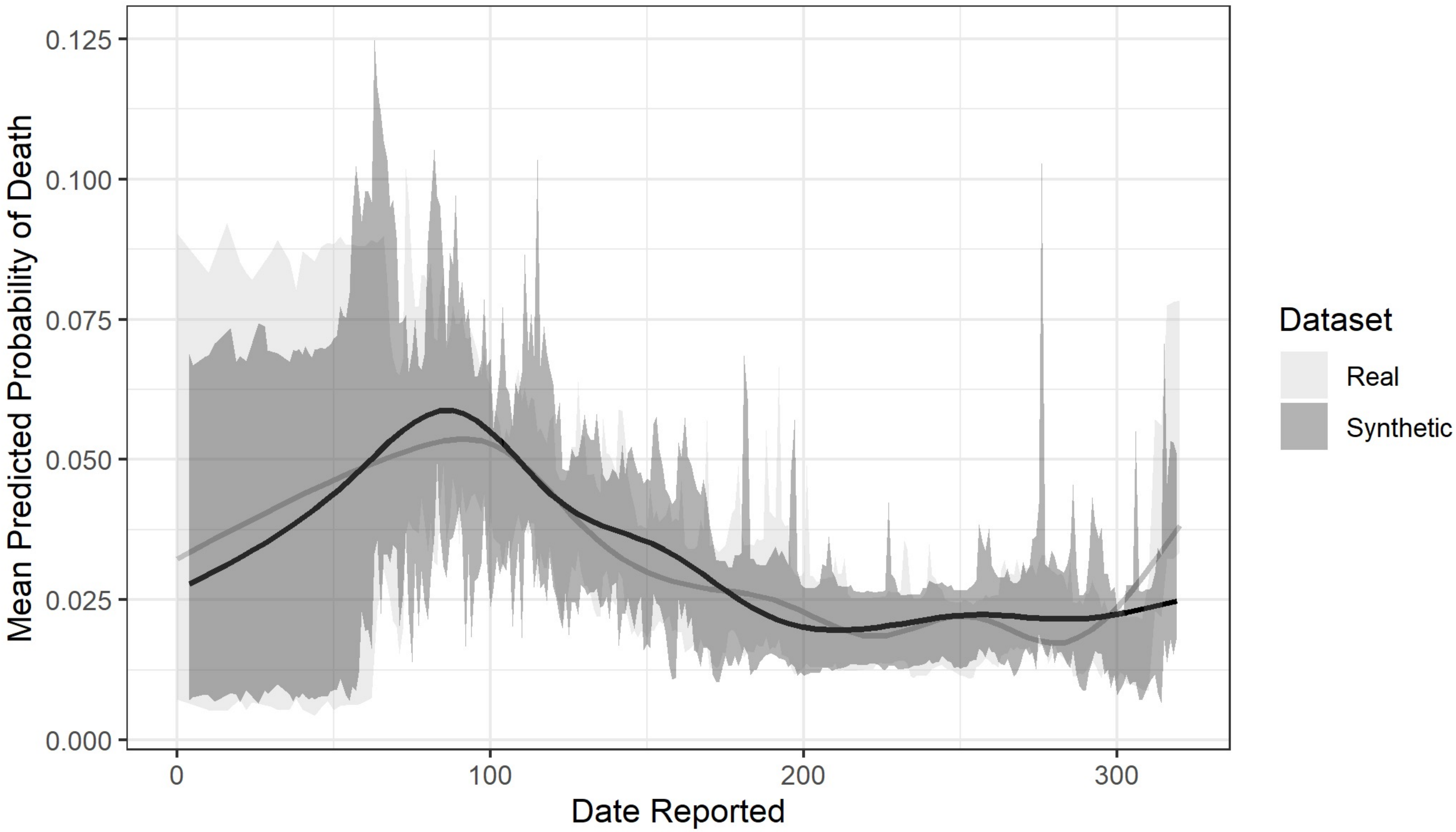


Distribution Comparisons





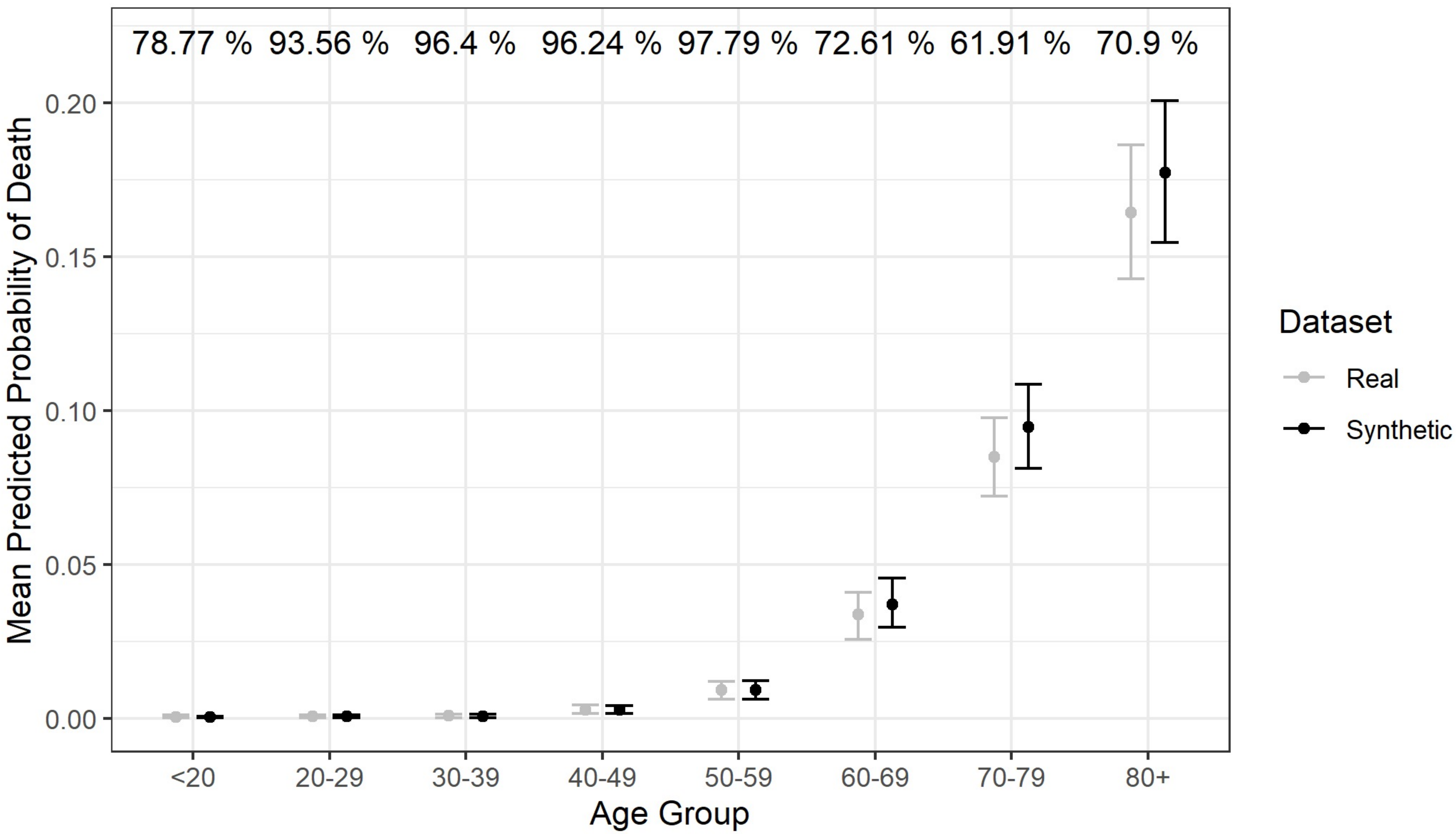
Mortality Over Time



Electronic Health Information Laboratory

Replica Analytics

Mortality By Age



Utility Framework

- An important concern of data users is the data utility
- Utility has multiple dimensions to it
- Synthetic data may be optimized on multiple utility dimensions simultaneously to meet the needs of multiple users, or on single dimensions to address the needs of limited users

IEEE SECURITY & PRIVACY
Editor: Khaled El Emam, kelemam@cheo.on.ca

Seven Ways to Evaluate the Utility of Synthetic Data

Khaled El Emam | Children's Hospital of Eastern Ontario Research Institute

Access to individual-level health data is going to be critical for managing the COVID-19 pandemic and enabling society to return to some form of (new) normal functioning. Broader data access is already starting to happen. At the same time, there has been growing alarm by the privacy community about the extent and manner of the level of data sharing that is going on with such sensitive information. In South Korea, broad data sharing has already resulted in some patients being reidentified and experiencing judgment and ridicule,[1,2] and some governments have begun to reduce the amount of information being shared about COVID-19 cases.[3–8] Data synthesis can provide a solution by enabling access to useful information while ensuring reasonable privacy protections.

There are already large-scale data-sharing efforts using synthetic data. For example, tabulations from the 2020 United States Census will be based on synthetic data. Public Health England has made a large cancer registry publicly available for analysts (the Simulacrum). Additional synthesis efforts are in the works by the National Institutes of Health (NIH) and NIH-funded projects.

Synthetic health data are generated from a model that is fit to a real data set as illustrated in Figure 1. Statistical machine learning and deep learning methods are typically used to fit this model. No specific advance knowledge of how the data will be used or analyzed is required to generate useful synthetic data. Once the model is fit, it is used to generate new data from that model. The generation is stochastic; therefore, a different data set is generated from the model each time.

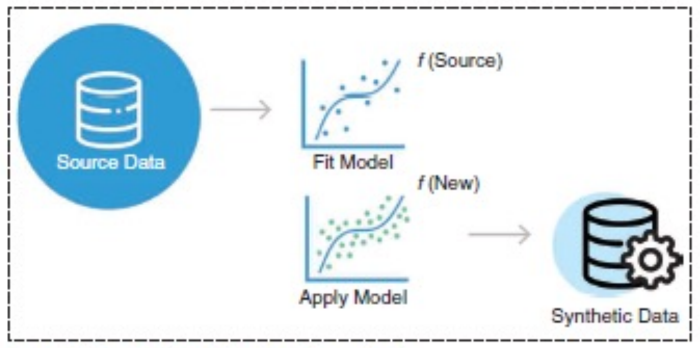


Figure 1. The basic workflow for data synthesis.

For data scientists to be comfortable using synthetic data, especially to build models that would influence public health and clinical decisions, there needs to be evidence demonstrating the utility of that data. In this article, we summarize the seven ways that the utility of synthetic data has been assessed thus far, and we close with some recommendations on their application.

Utility Assessment Methods

The following are seven methods for assessing the utility of synthetic data. In these descriptions, we will refer to the real data as the source and the synthetic data as the generated data set. The assumption is made that the objective is to make individual-level patient data broadly available, as opposed to, for example, releasing aggregate statistics or summary tables.

Utility assessment is performed by the entity performing the data synthesis before making the data available more broadly. Typically, the results of the utility assessments are documented and shared with the data users.

Replication of Studies

The default approach to assess utility is to perform an analysis on the real data and then replicate that on the synthetic data. If the same conclusions are drawn from the two different analyses, then the synthetic data are deemed to have high utility. The analysis that is chosen must be

Digital Object Identifier 10.1109/MSEC.2020.2992821
Date of current version: 9 July 2020

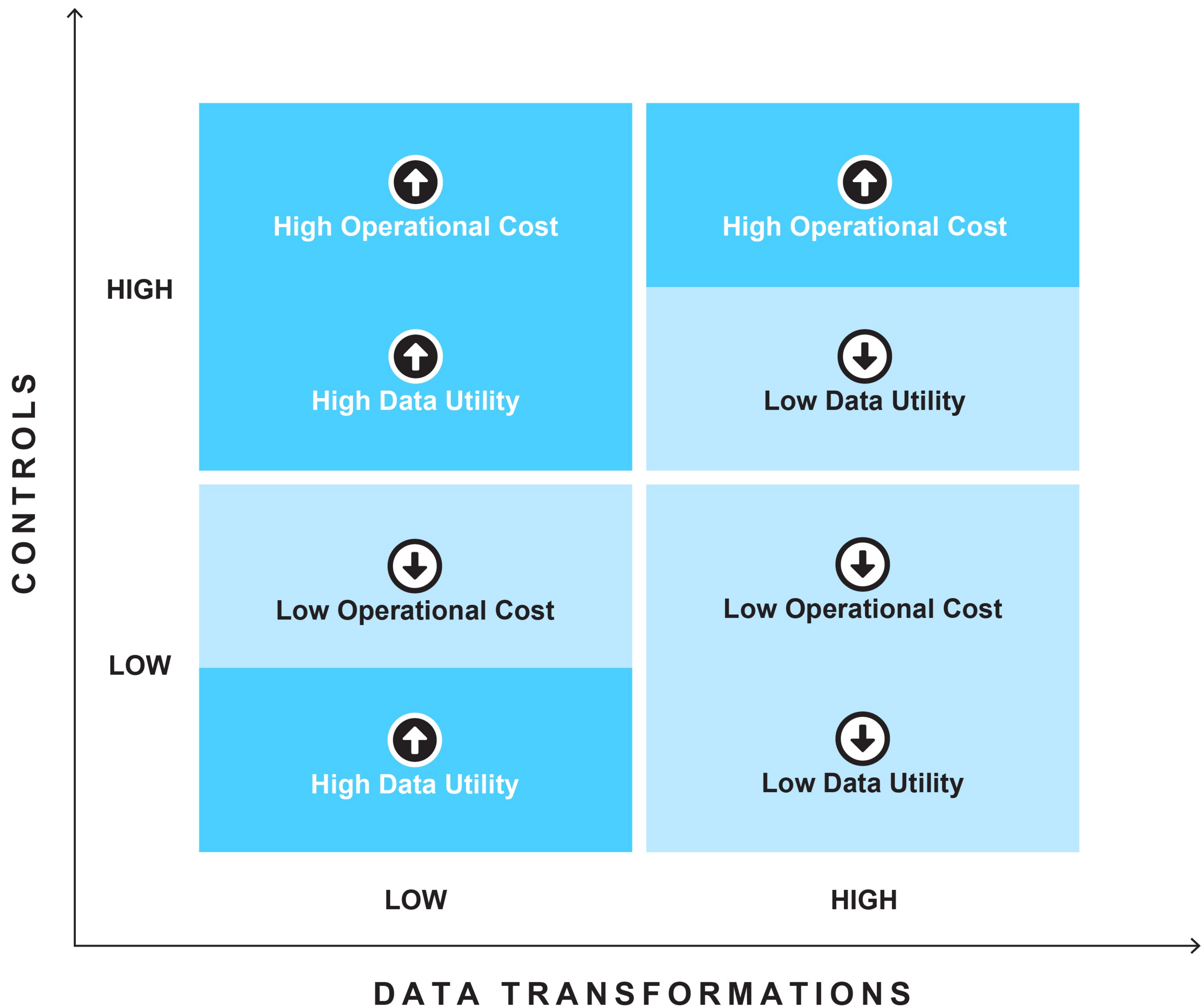
Risk-based Approach

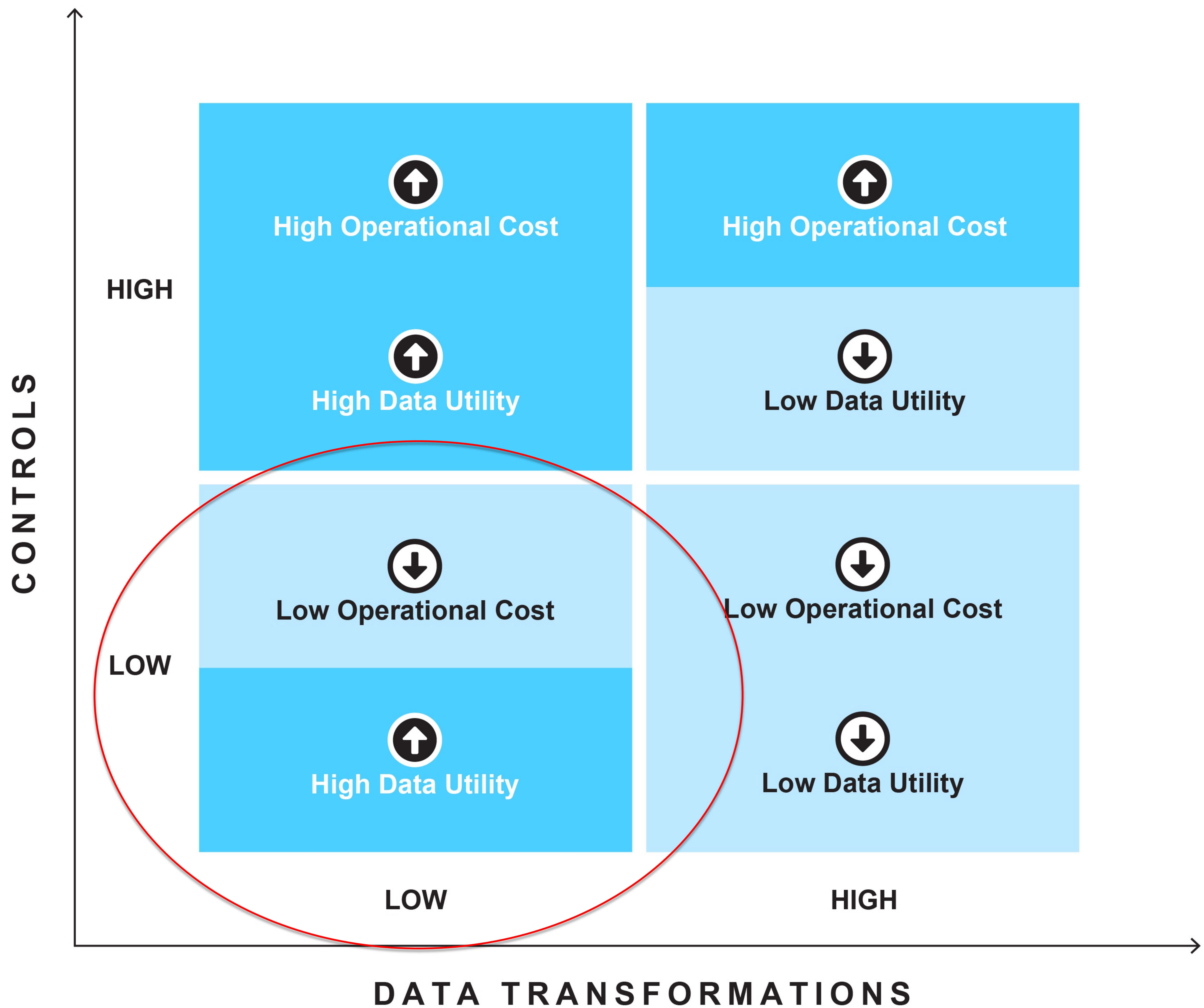
Risk-based Approach

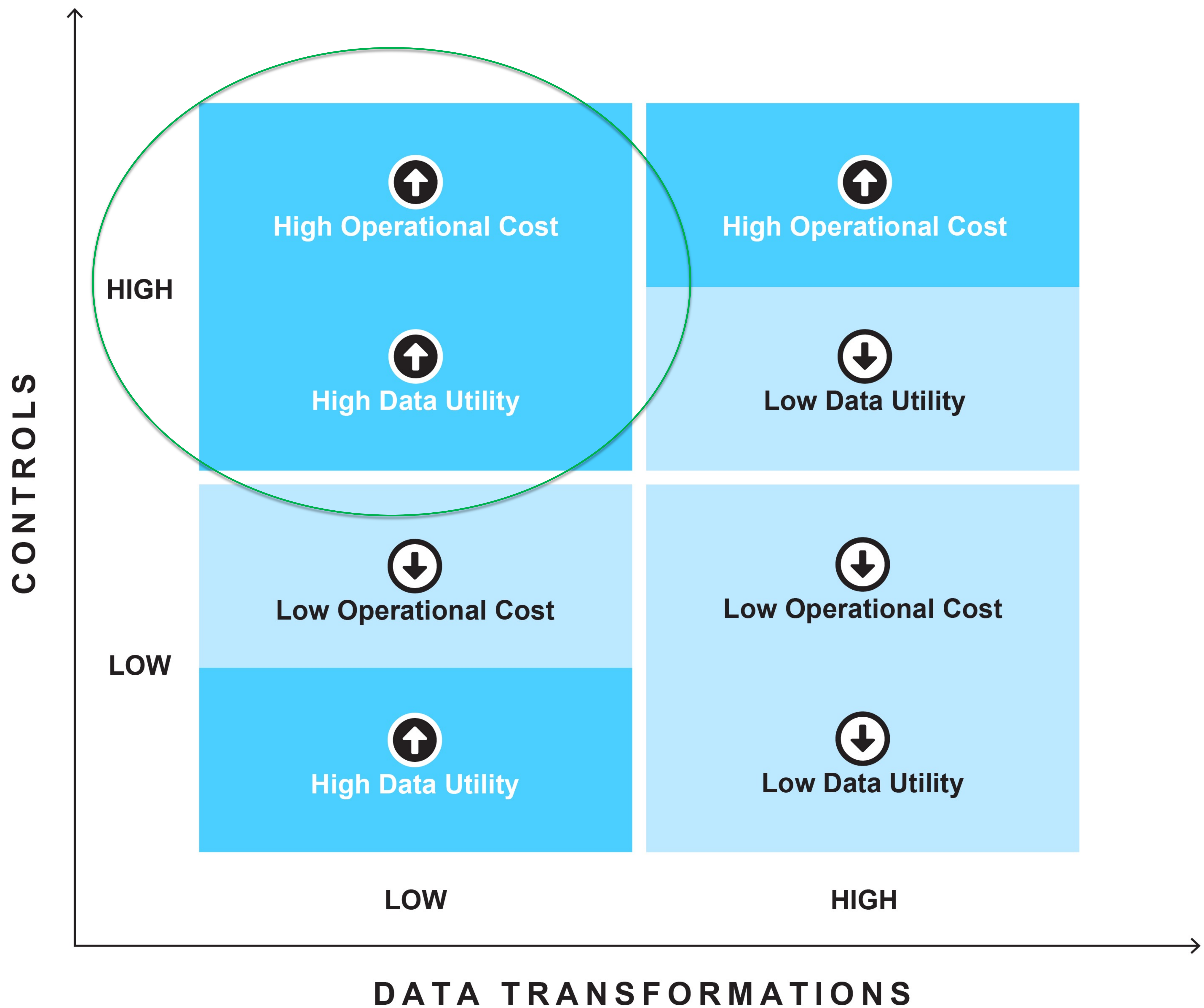
- Generalization
- Suppression
- Addition of noise
- Microaggregation

Risk-based Approach

- Security controls
- Privacy controls
- Contractual controls


CONTROLS
HIGH
LOW
High Operational Cost
High Data Utility
High Operational Cost
Low Data Utility
Low Operational Cost
High Data Utility
Low Operational Cost
Low Data Utility
LOW
HIGH
DATA TRANSFORMATIONS


CONTROLS
HIGH
LOW
High Operational Cost
High Data Utility
High Operational Cost
Low Data Utility
Low Operational Cost
High Data Utility
Low Operational Cost
Low Data Utility
LOW
HIGH
DATA TRANSFORMATIONS



CONTROLS
HIGH
LOW
High Operational Cost
High Data Utility
High Operational Cost
Low Data Utility
Low Operational Cost
High Data Utility
Low Operational Cost
Low Data Utility
LOW
HIGH
DATA TRANSFORMATIONS

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

Electronic Health Information Laboratory

Replica Analytics

Skill Set

- The skills needed to create de-personalized datasets are very specialized, take time to develop, and generally difficult to find cost-effectively
- This limits the ability to scale
- Synthesis requires minimal skills in practice – it is a computationally intensive process

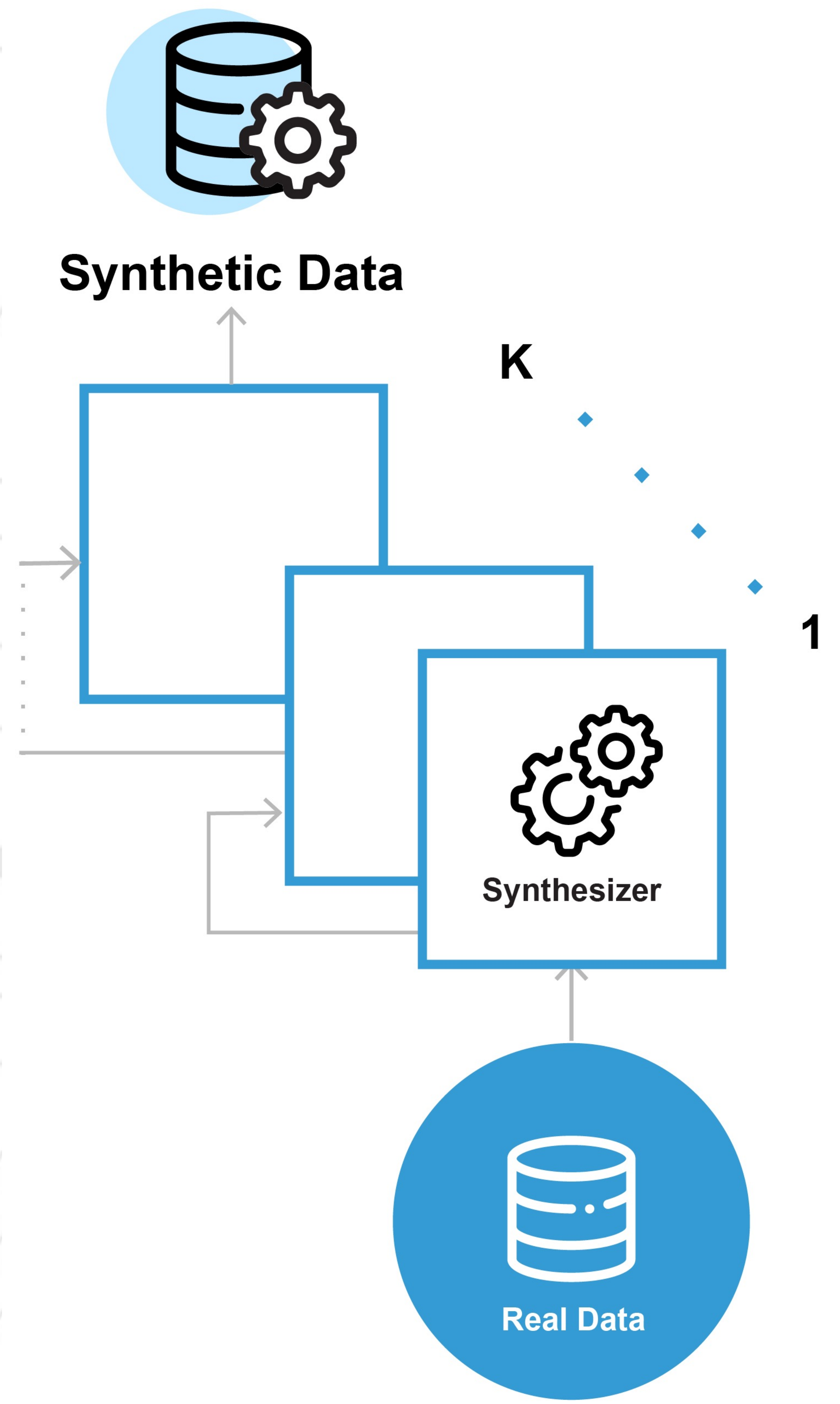
Electronic Health Information Laboratory

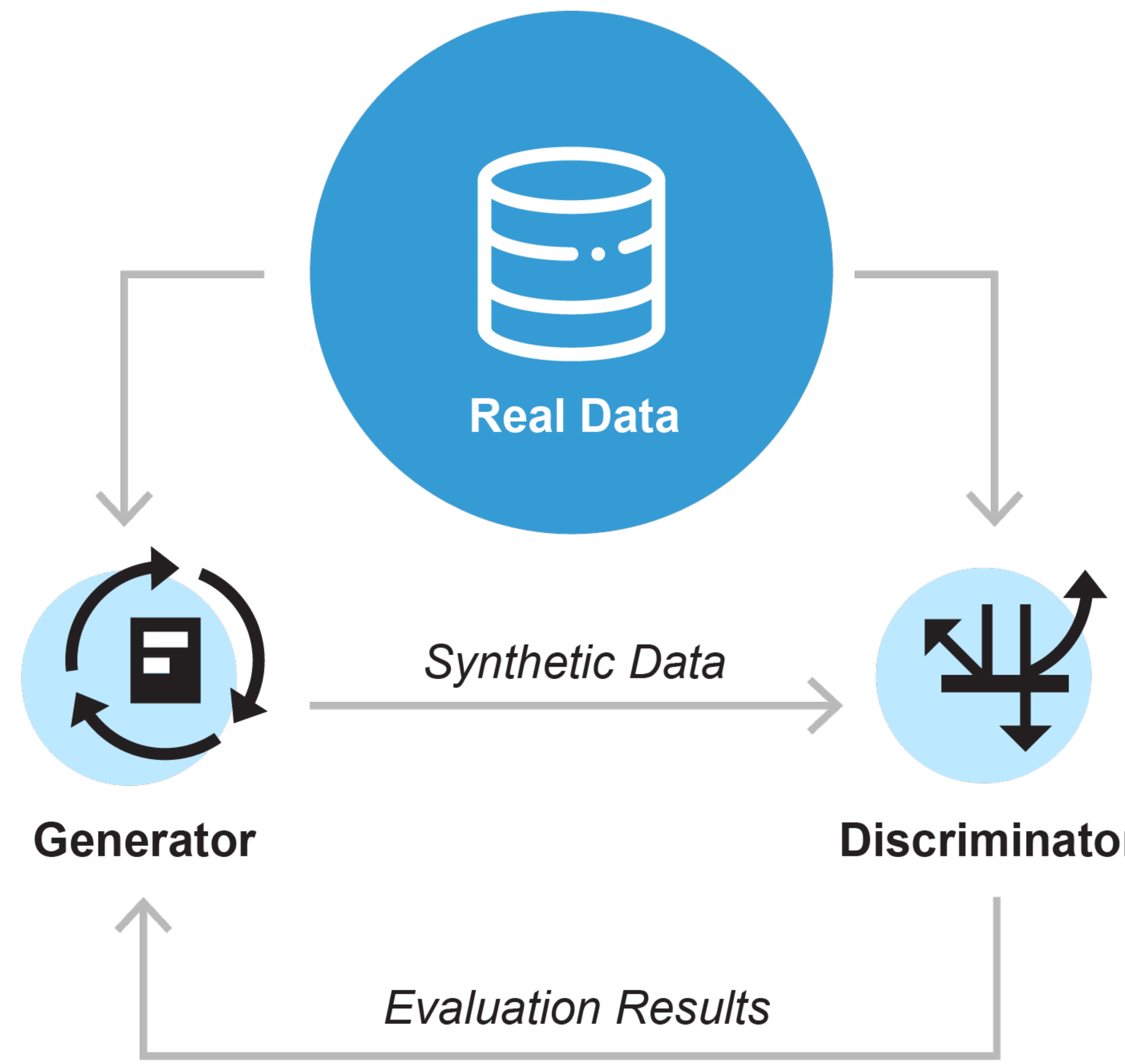
Replica Analytics

Regulatory Questions

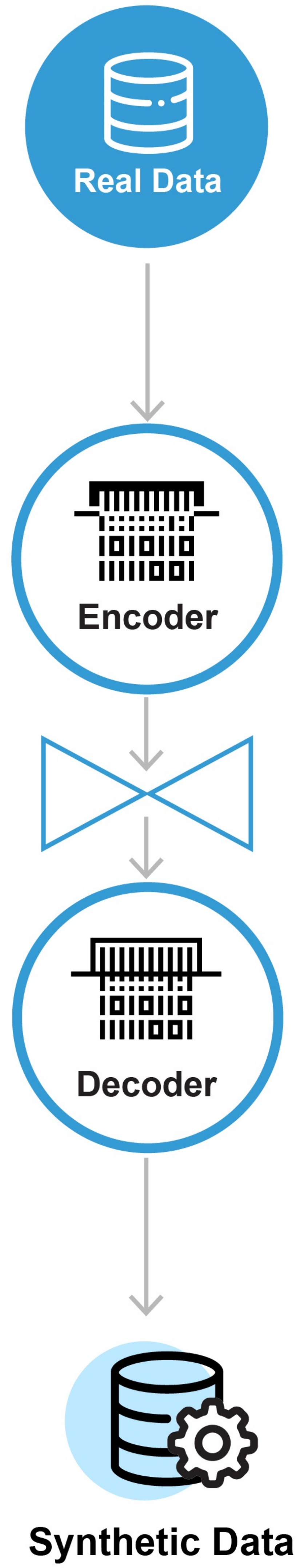
- Is synthetic data considered non-identifiable information ?
- Does the act of converting identifiable information into non-identifiable synthetic information require additional consent or authorization ?
- Can a data custodian outsource the creation of synthetic data ?
- Can synthetic data be used for any purpose ?

Sequential Synthesis

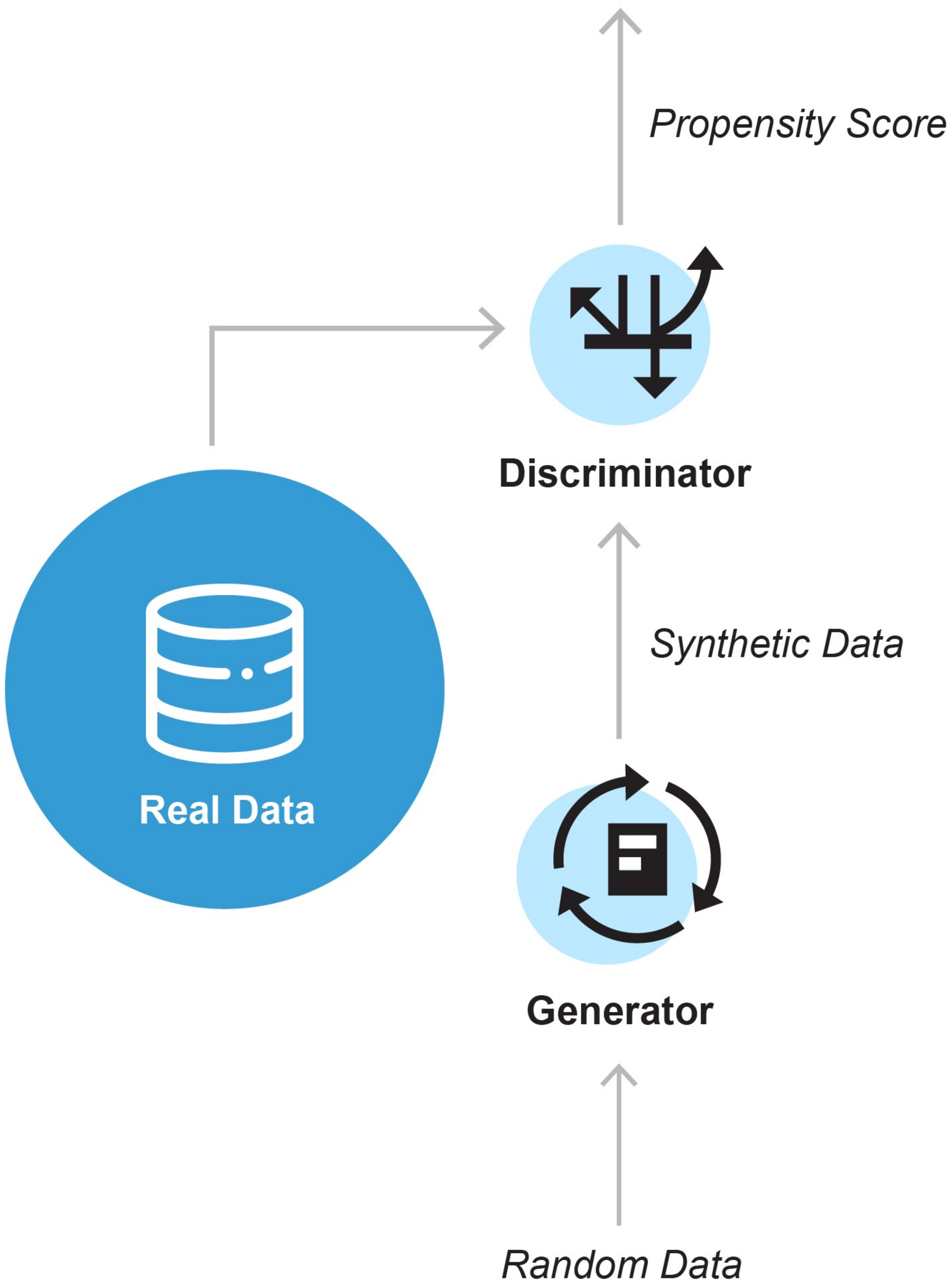


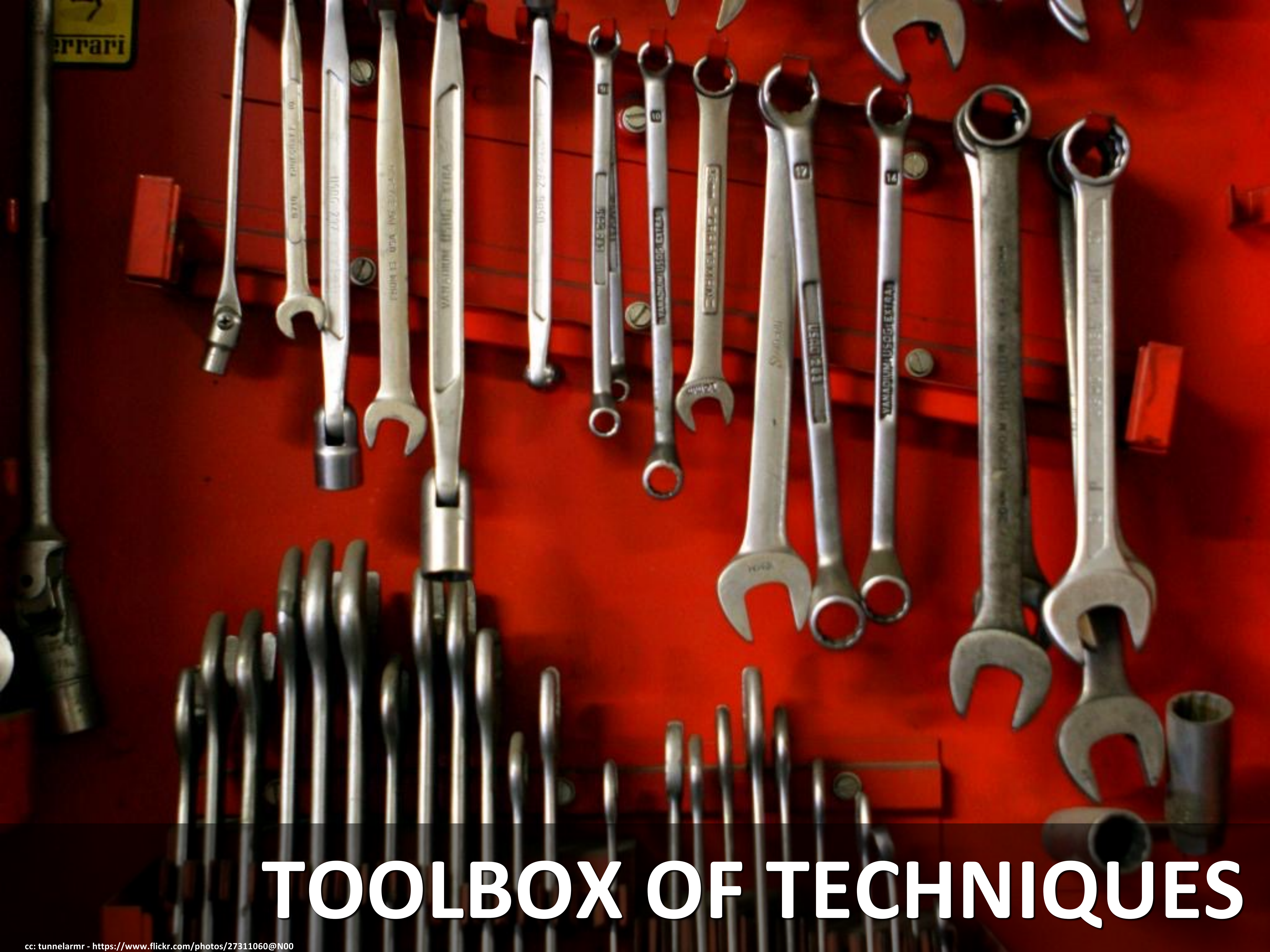


Variational Auto Encoder (VAE)



Generative Adversarial Network (GAN)





TOOLBOX OF TECHNIQUES


QUESTIONS

References

- 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. [Online]. Available: https://www.jmir.org/2020/11/e23139.
- 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.

Electronic Health Information Laboratory

Replica Analytics