GENERATING SYNTHETIC LONGITUDINAL DATA

DEC. 1, 2021 | 11AM EDT

Presented by

Dr. Khaled El Emam,
CEO, Replica Analytics

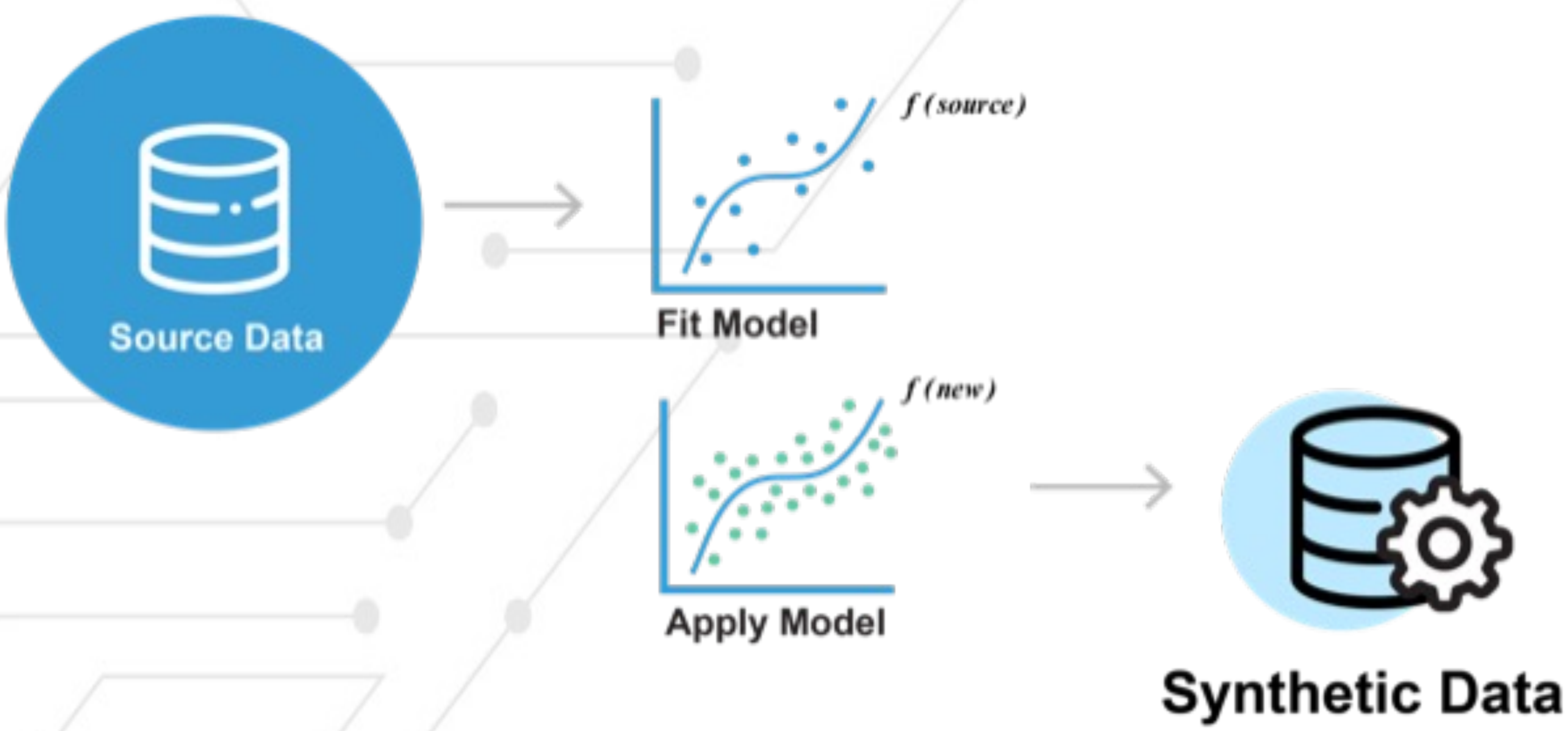
Lucy Mosquera,
Director of Data Science

Generating Synthetic Longitudinal Data

Khaled El Emam
Lucy Mosquera
1st *December 2021*

kelemam@replica-analytics.com
lmosquera@replica-analytics.com

The Synthesis Process



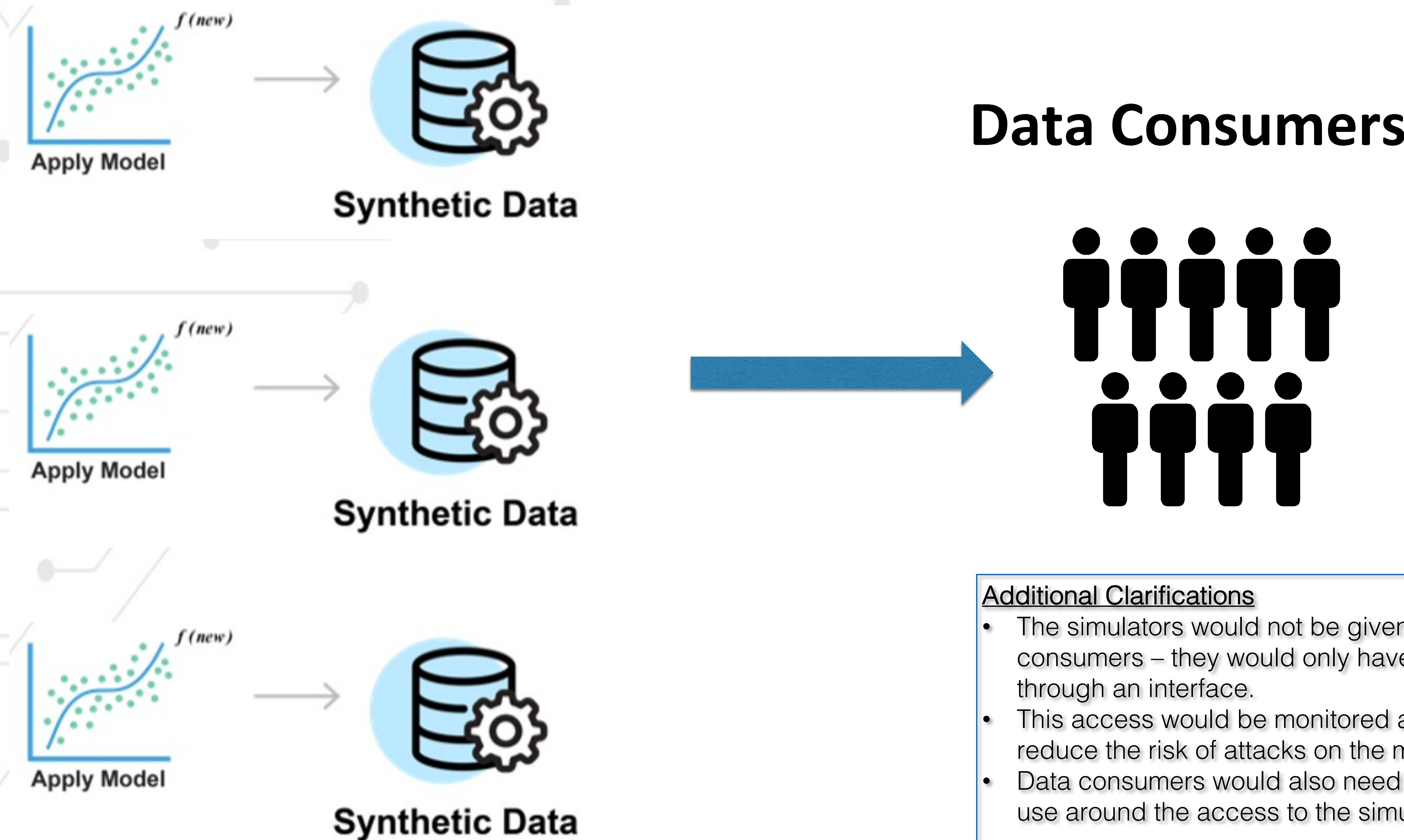
COU1A	AGECAT	AGELE70	WHITE	MALE	BMI
United States	2	1	1	1	33.75155
United States	2	1	1	0	39.24707
United States	1	1	1	0	26.5625
United States	4	1	1	1	40.58273
United States	5	0	0	1	24.42046
United States	5	0	1	0	19.07124
United States	3	1	1	1	26.04938
United States	4	1	1	1	25.46939

Additional Clarifications

- 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

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

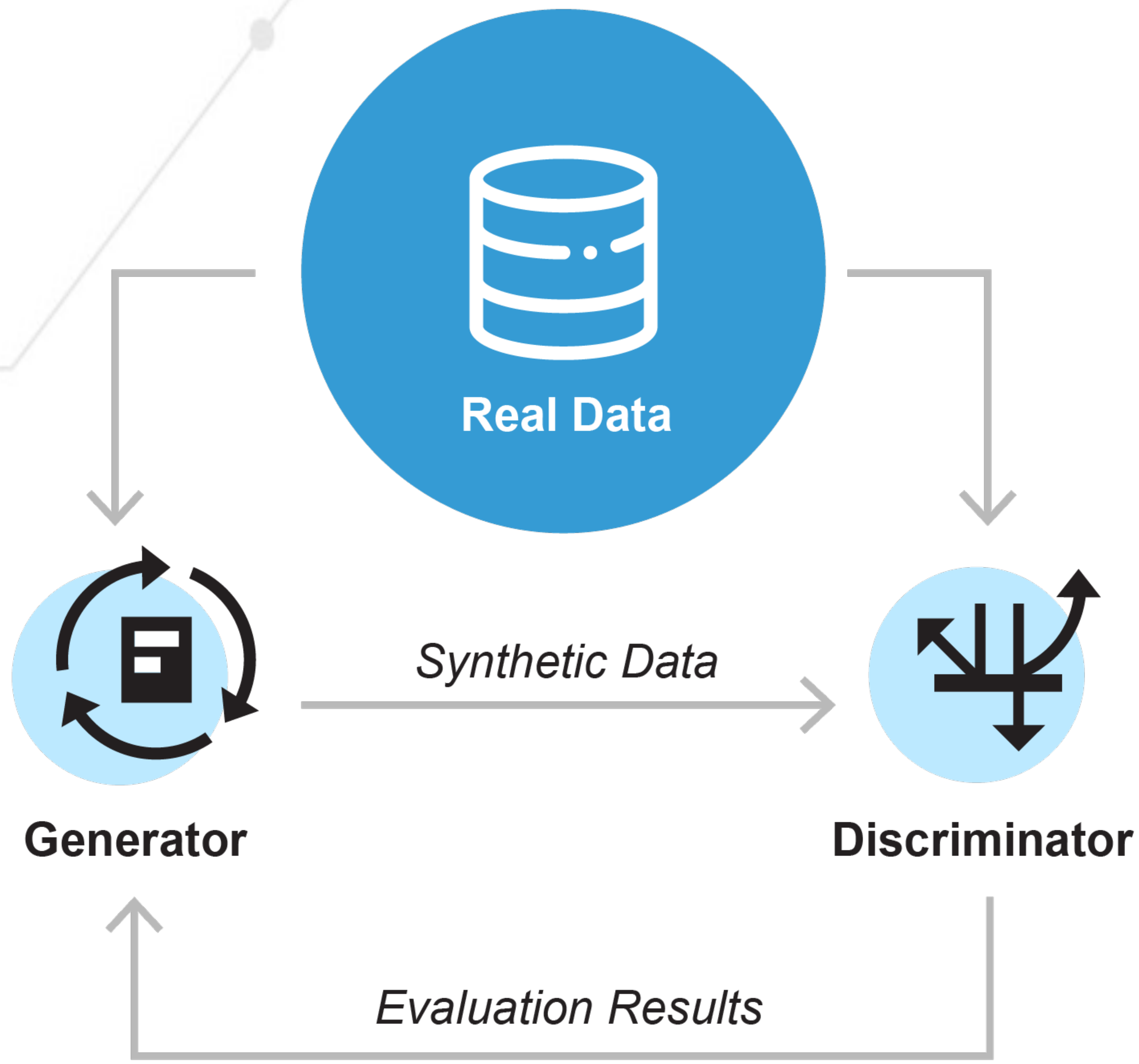


<u>Additional Clarifications</u>

- The simulators would not be given to the data consumers – they would only have access to them through an interface.
- This access would be monitored and throttled to reduce the risk of attacks on the models.
- Data consumers would also need to agree to terms of use around the access to the simulators.

Replica Analytics

Training a generative model often uses a discriminator



Replica Analytics

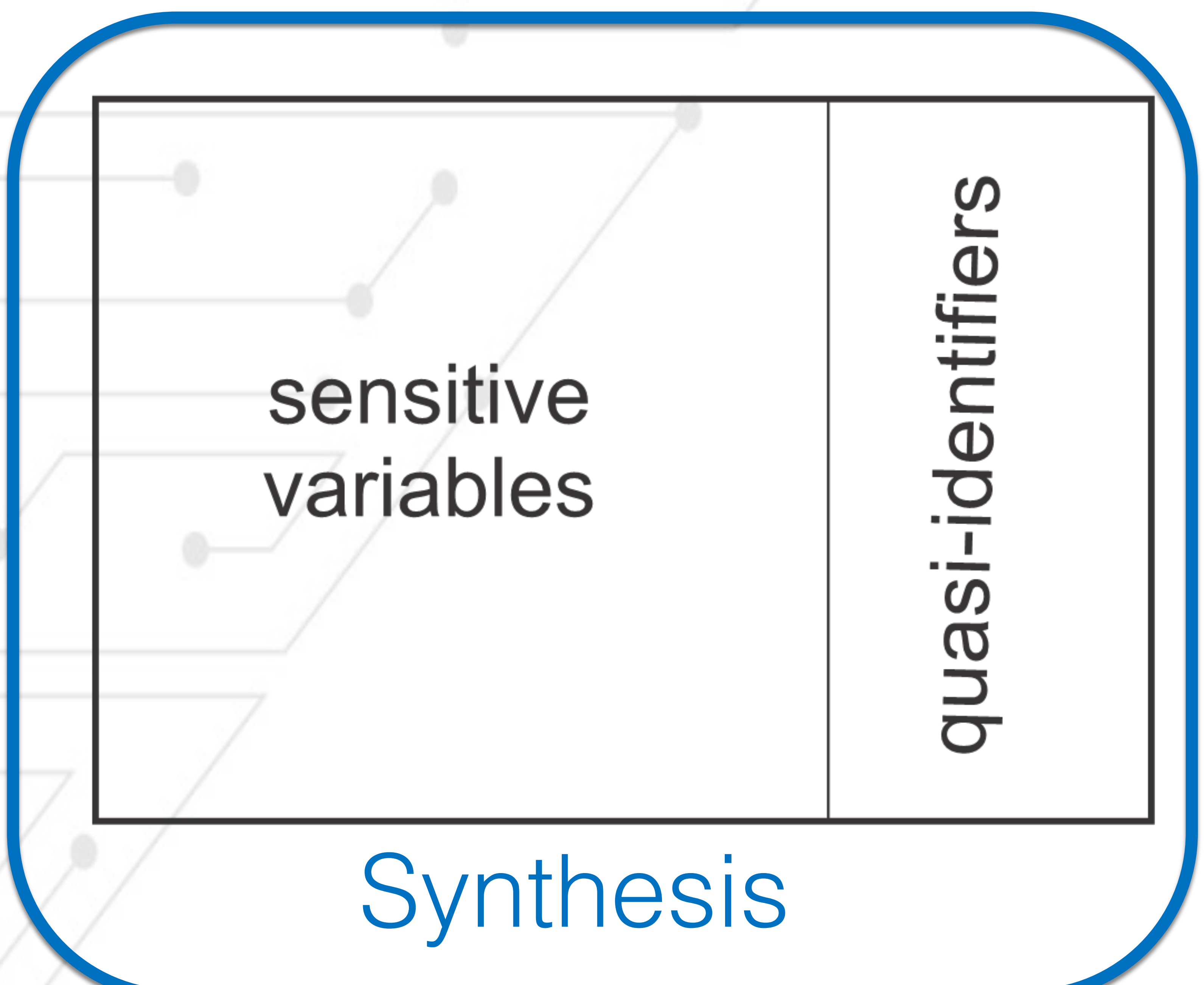
The synthesis of longitudinal data requires a different approach

- Features & Cohorts:
  - Define features on the raw longitudinal data and then synthesize the tabular feature dataset
  - Define a cohort on the raw longitudinal data and then synthesize the tabular cohort dataset
- Raw Longitudinal:
  - Fully vs partially synthetic data
  - For RWD we use a hybrid approach of sequential synthesis and recurrent neural network architectures to synthesize these – full synthesis

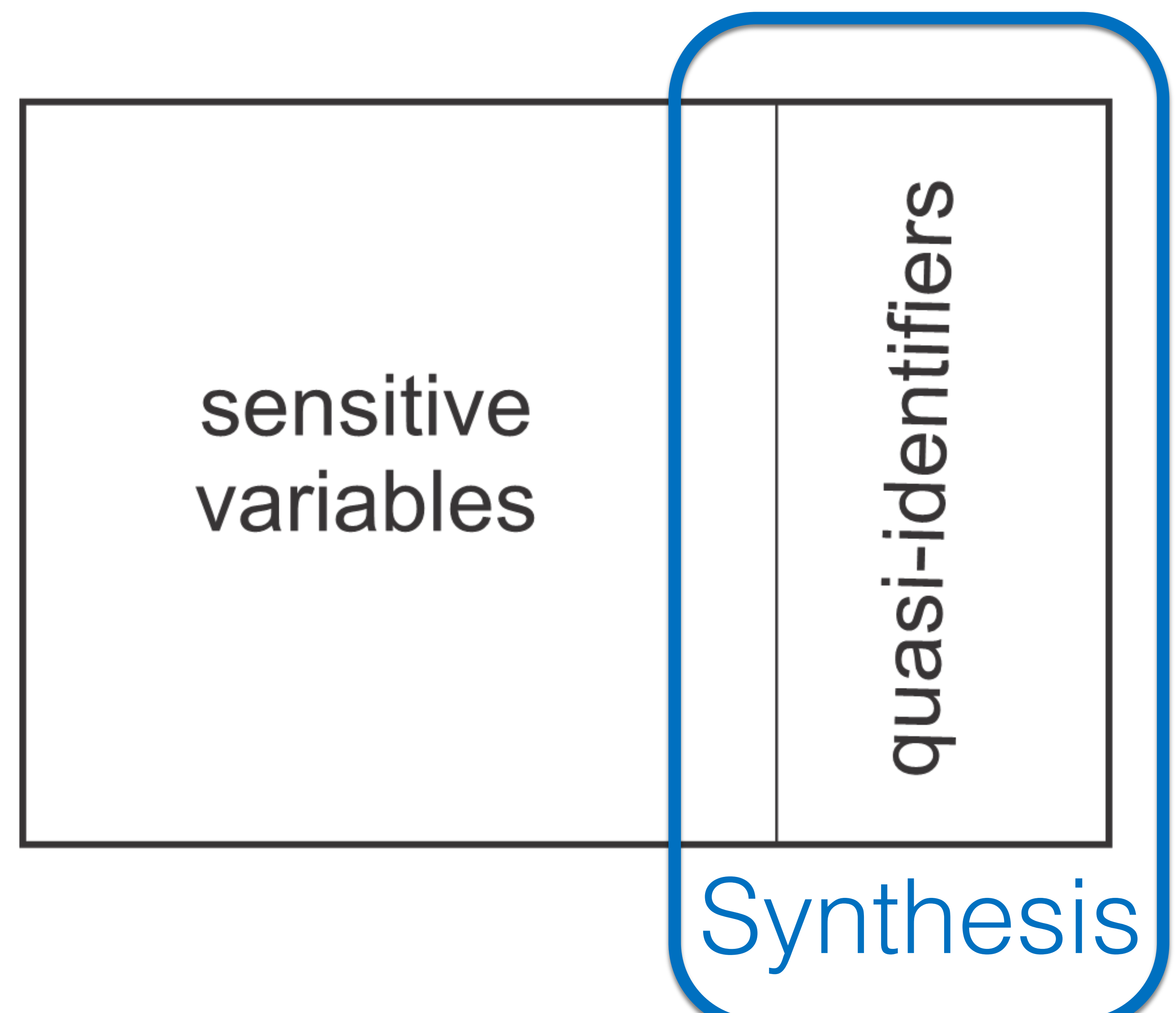
Replica Analytics

Two synthesis strategies for raw longitudinal data

Full Synthesis
Synthesize all variables

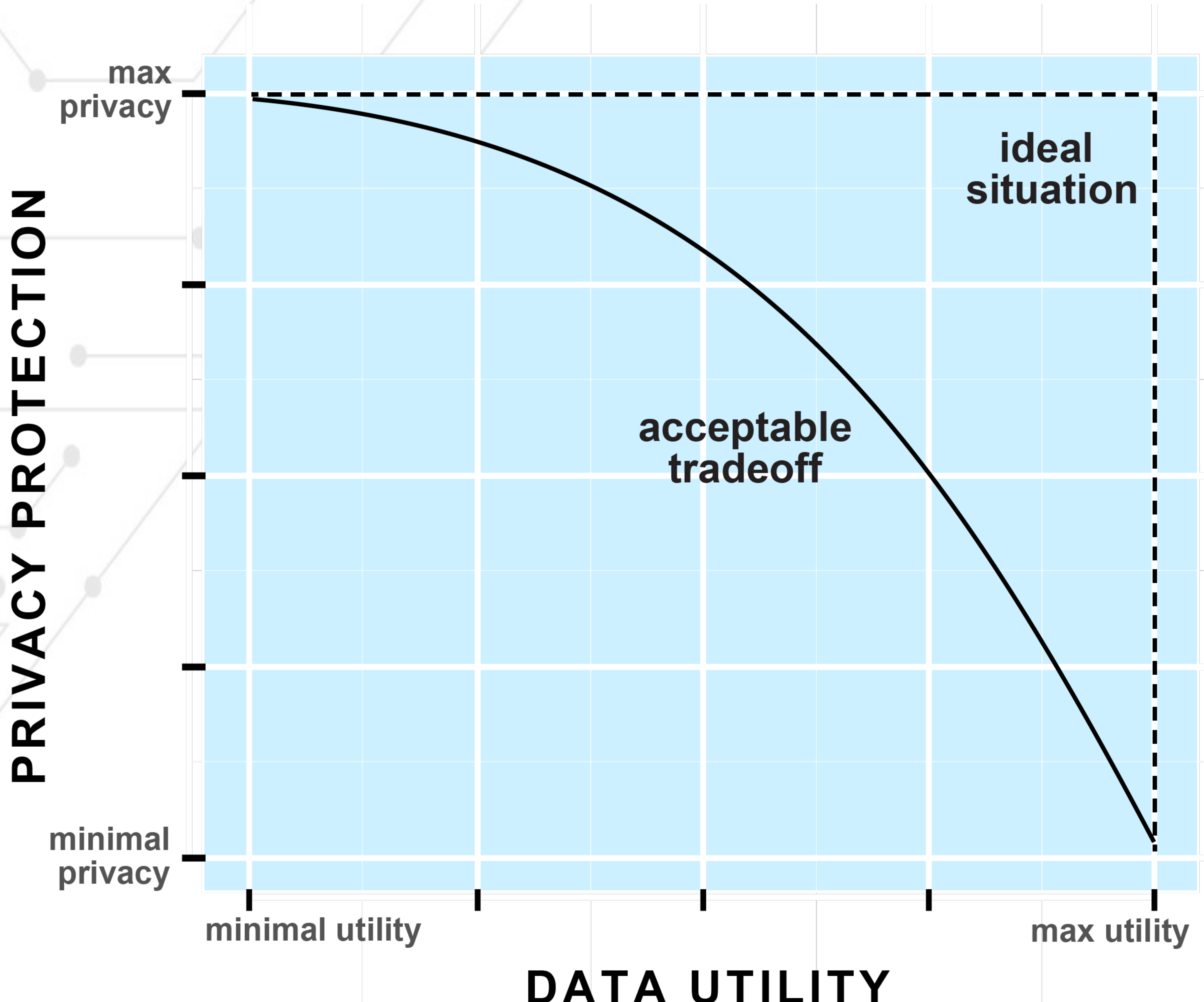


Partial Synthesis
Synthesize quasi-identifiers



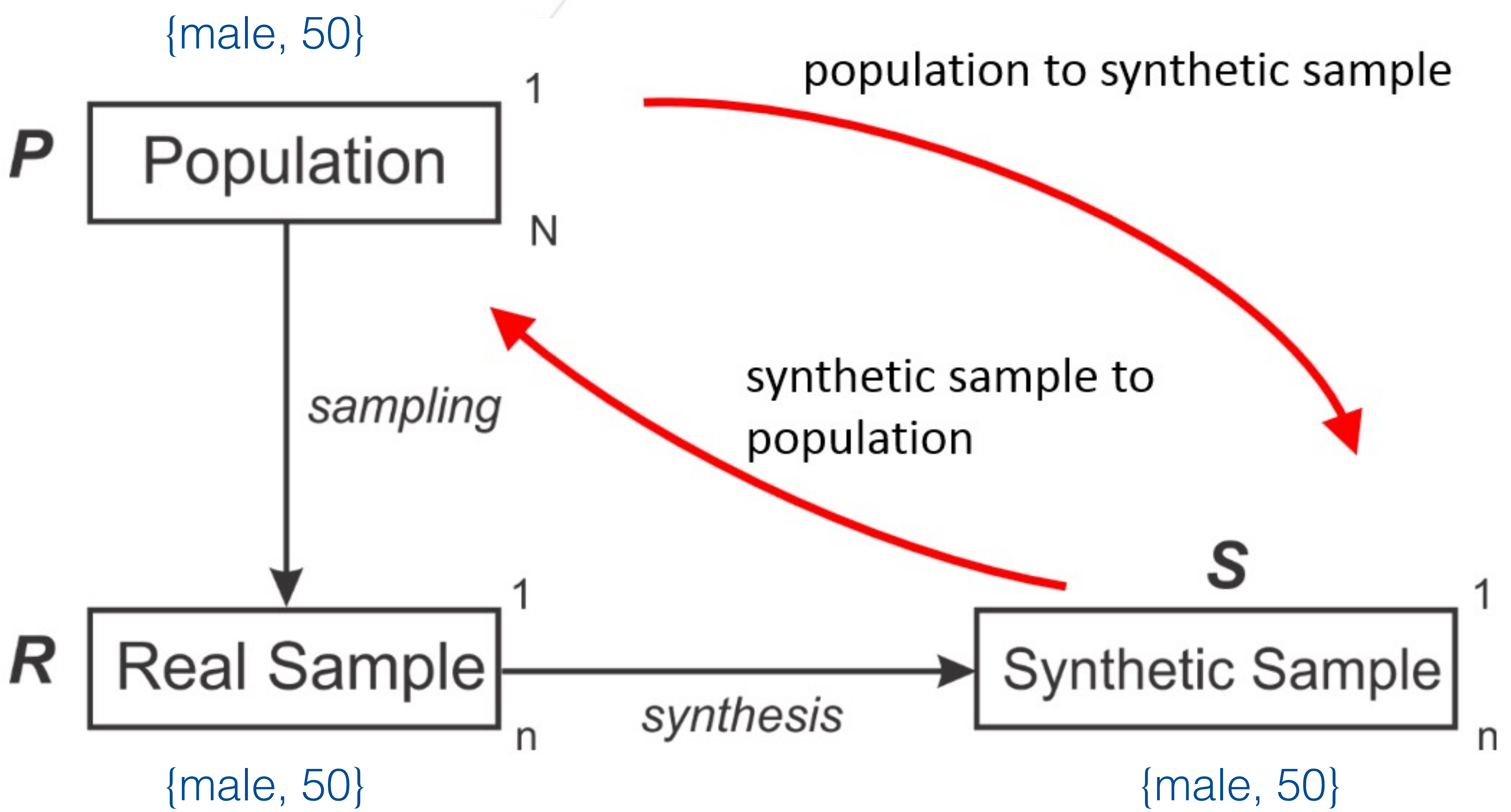
Replica Analytics

Privacy-Utility Trade-off



Replica Analytics

Identity Disclosure Model



Replica Analytics

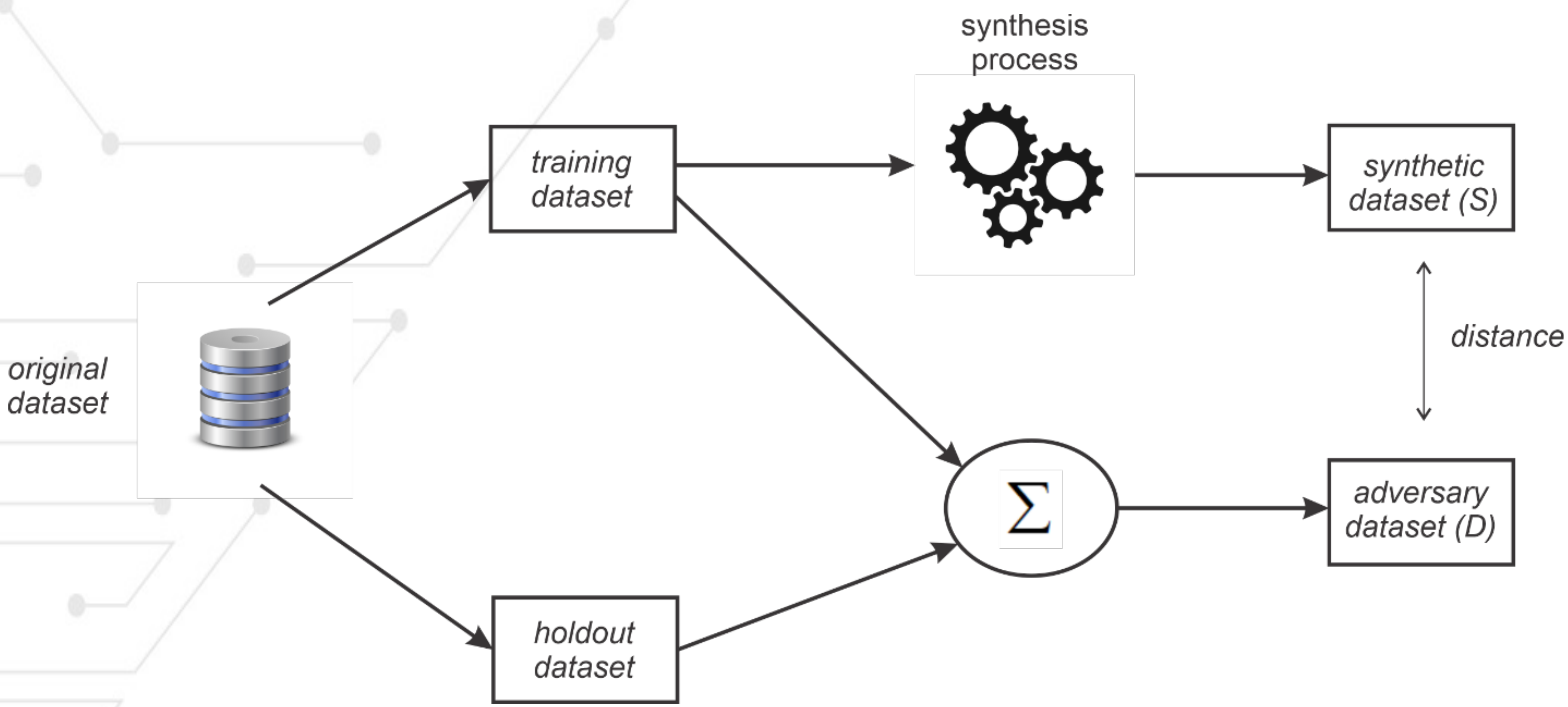
Evaluations of (re-)identification risks show that it is low in multiple studies across multiple datasets

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

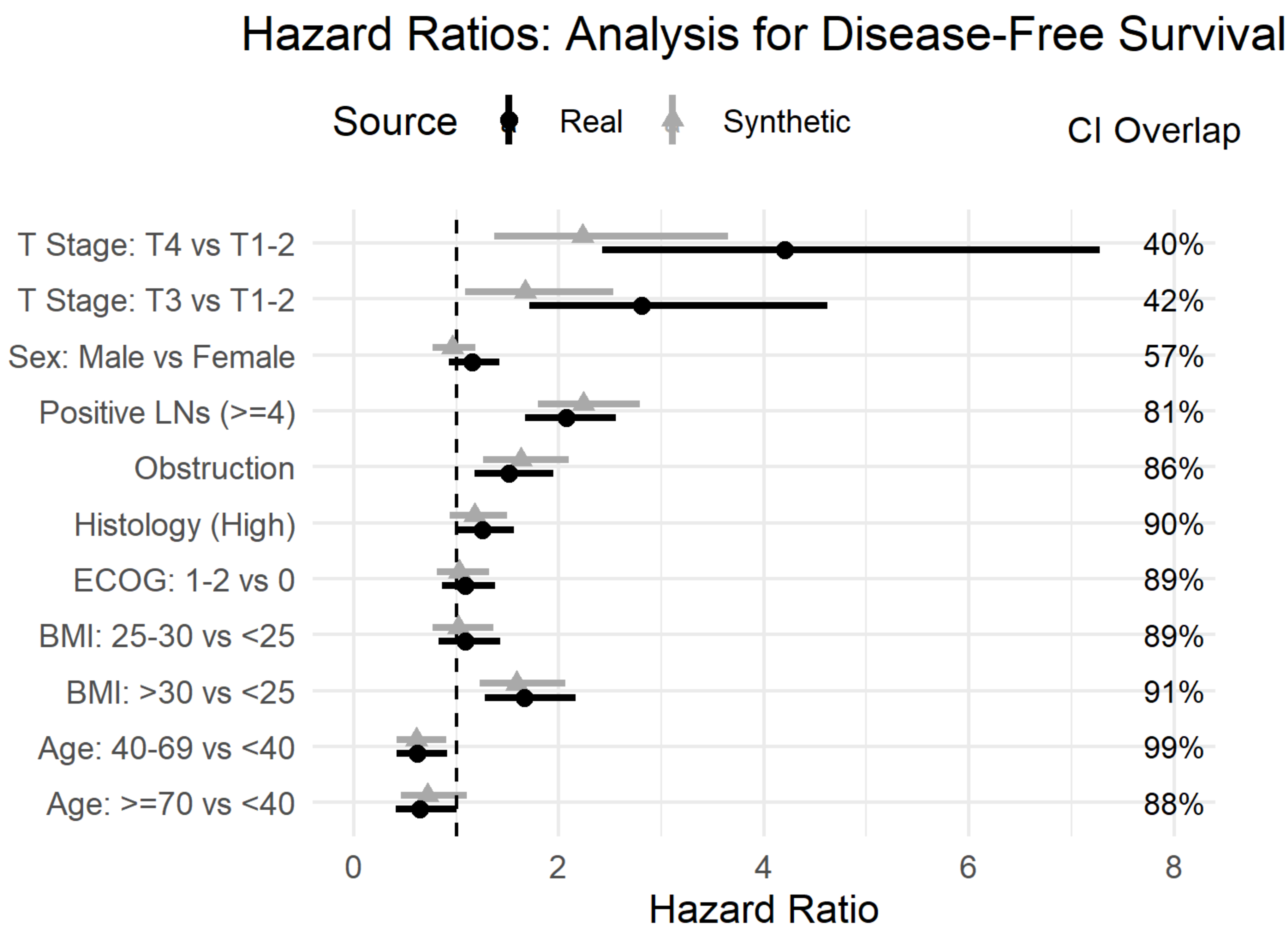
A commonly used risk threshold = 0.09

Replica Analytics

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

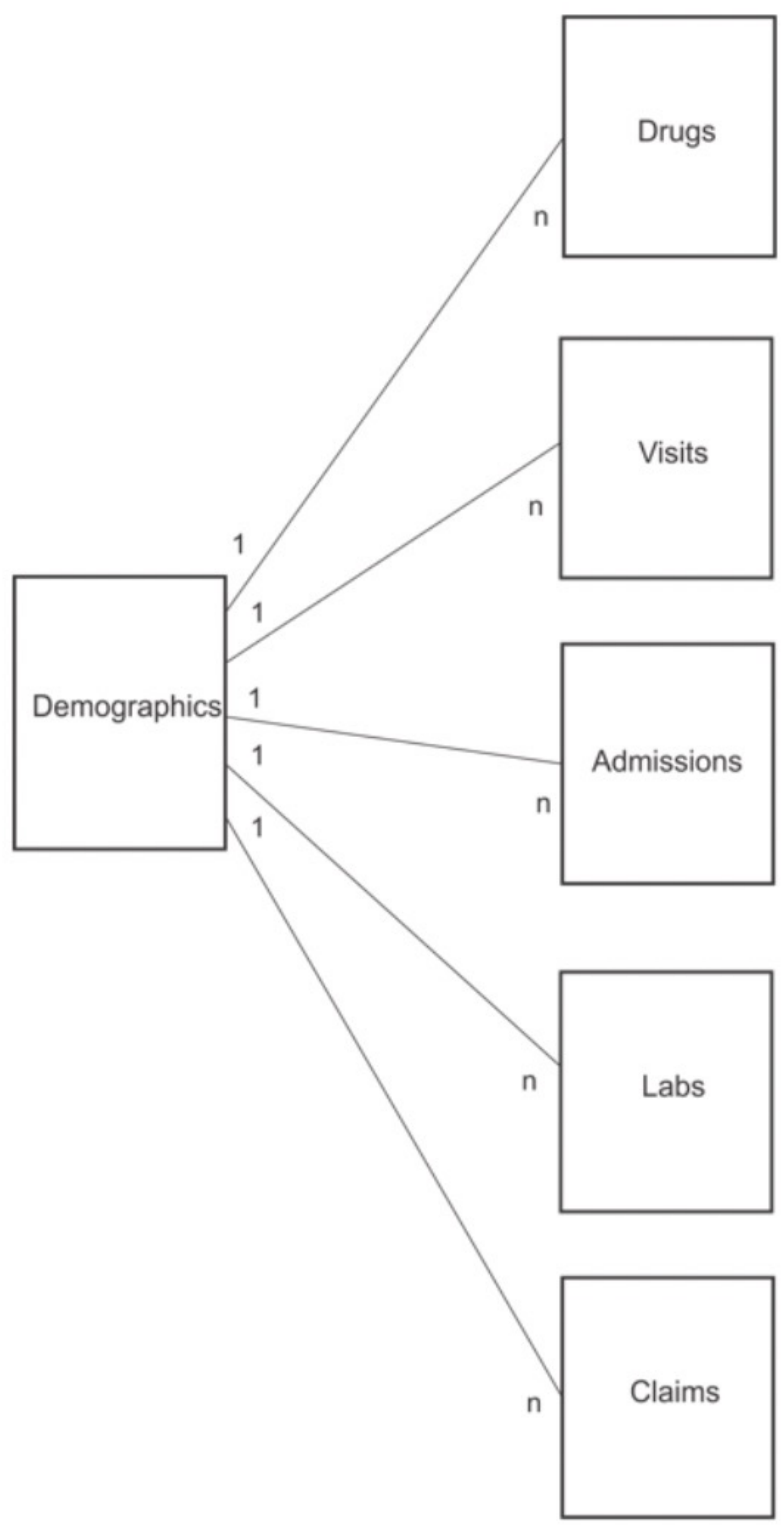


Comparing real and synthetic data: Adjusted model of impact of bowel obstruction on DFS



Replica Analytics

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)

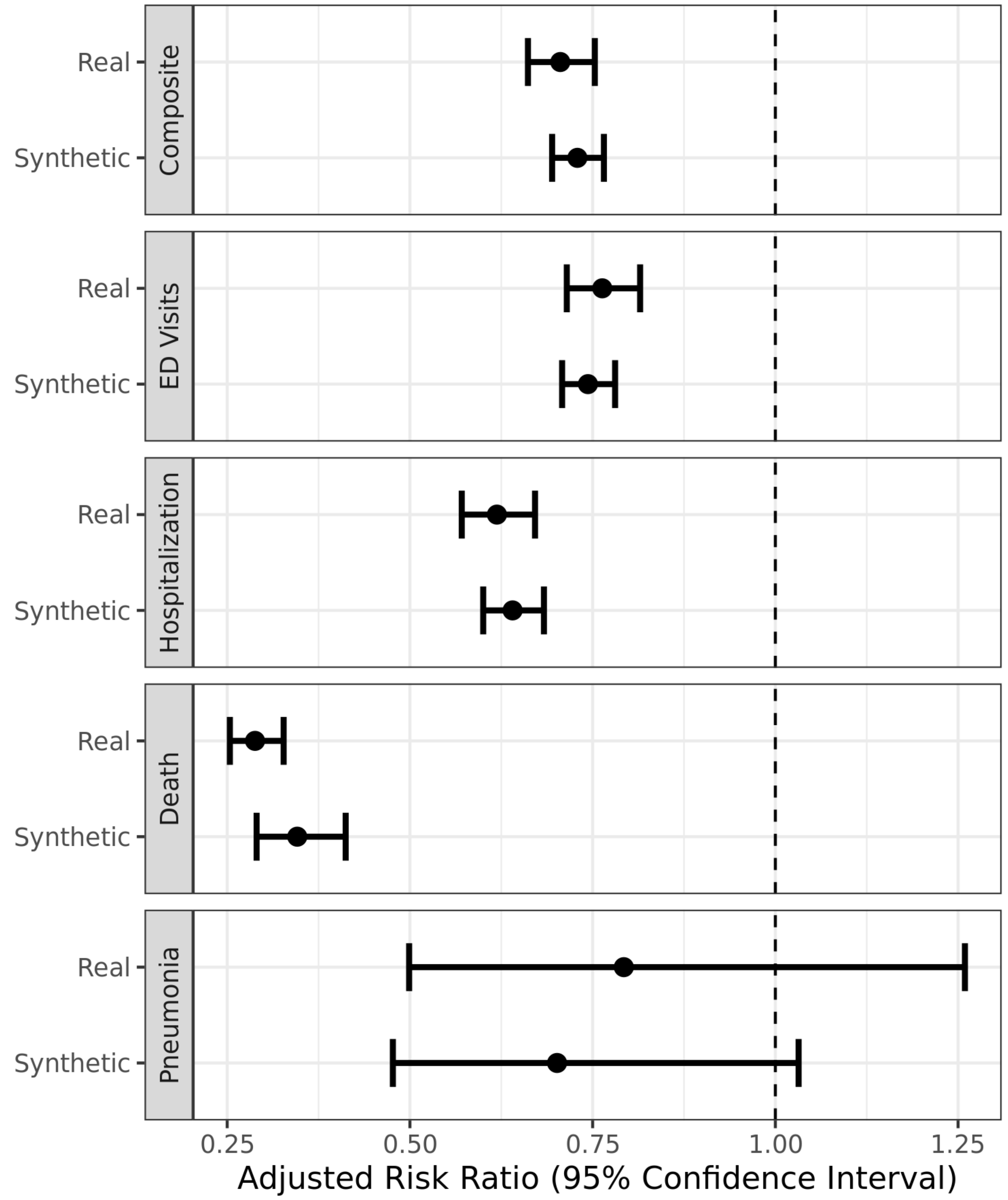
Visits (ED)
Relative admission time in days
Problem code 1
Problem code 2
Resource intensity weights

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

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

Replica Analytics

One way to classify utility metrics is as broad and narrow

broad metrics → narrow 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.

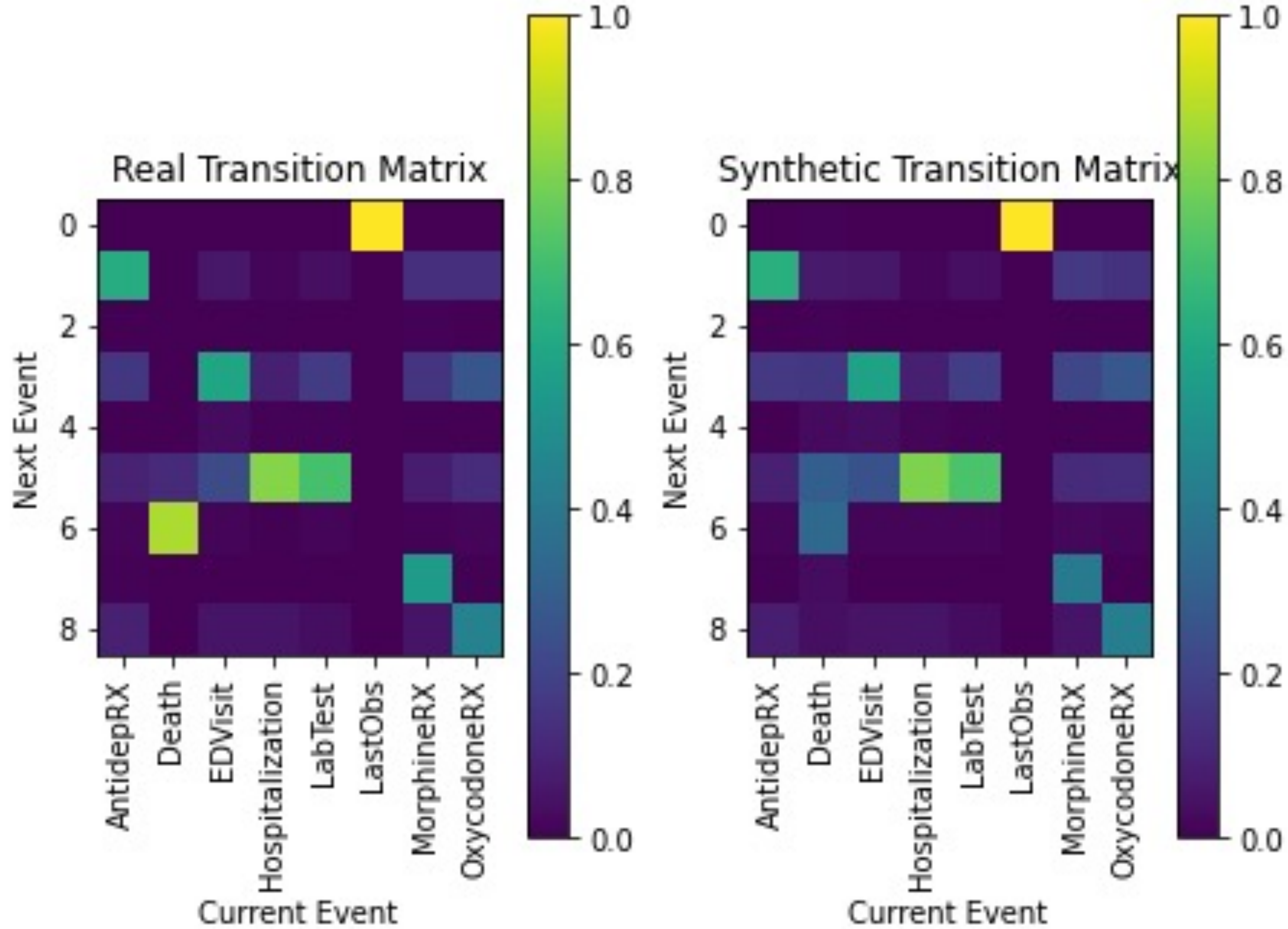
These are workload-specific and are what is of most interest to the data users. However, all the possible workloads will not be known in advance and therefore we have to consider representative workloads when developing and evaluating utility metrics.

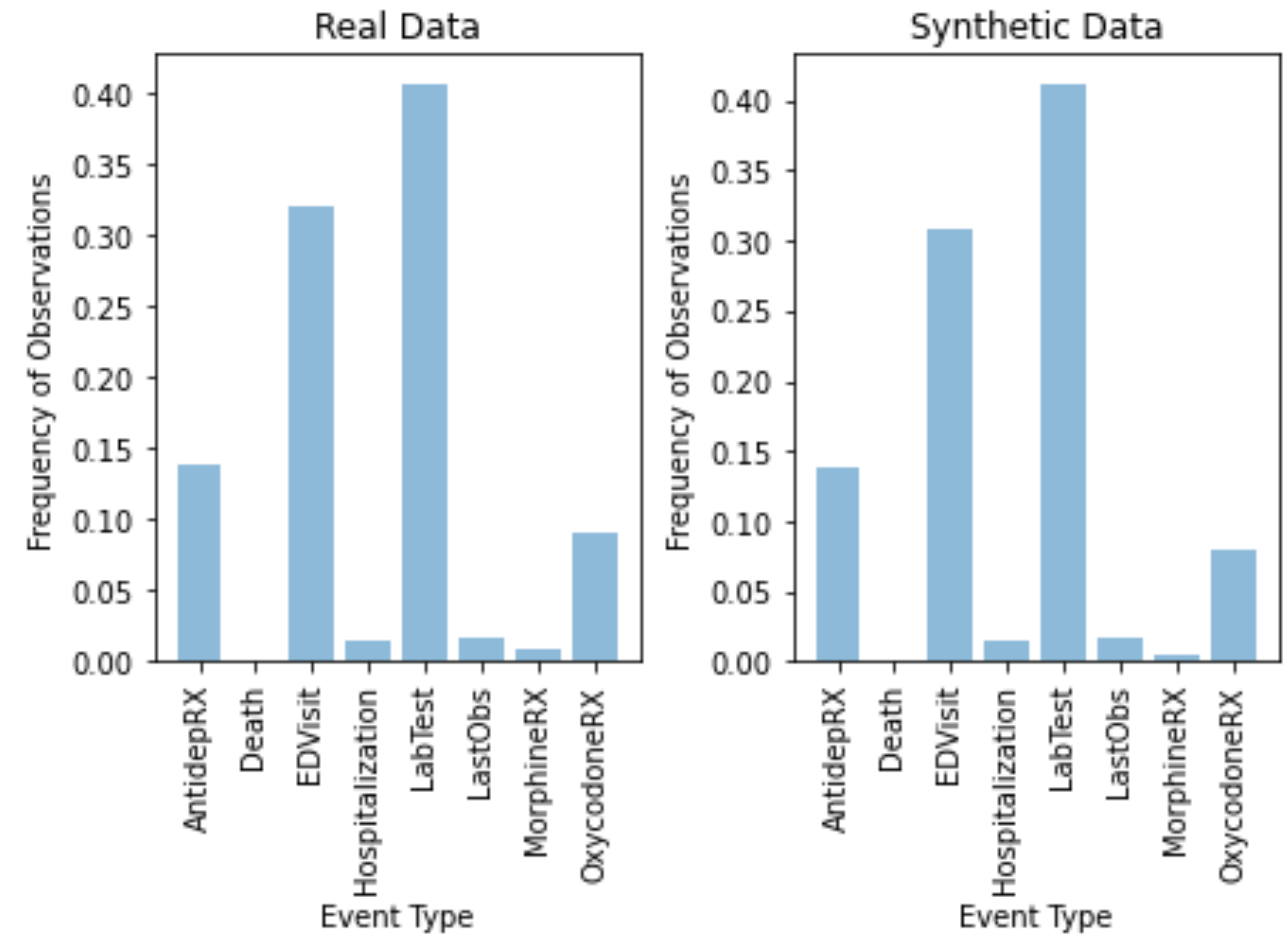
Replica Analytics

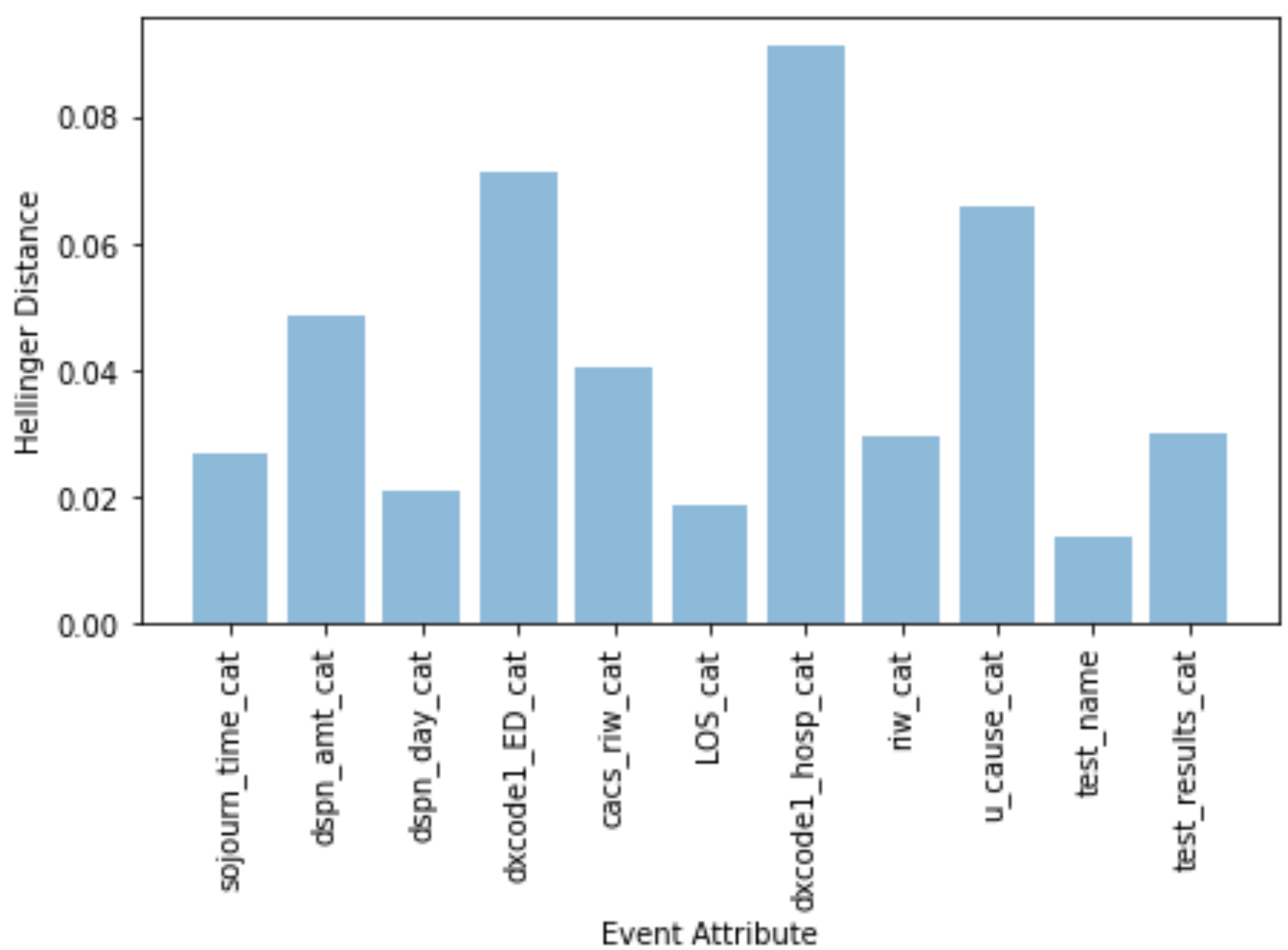
Examples of Broad Metrics

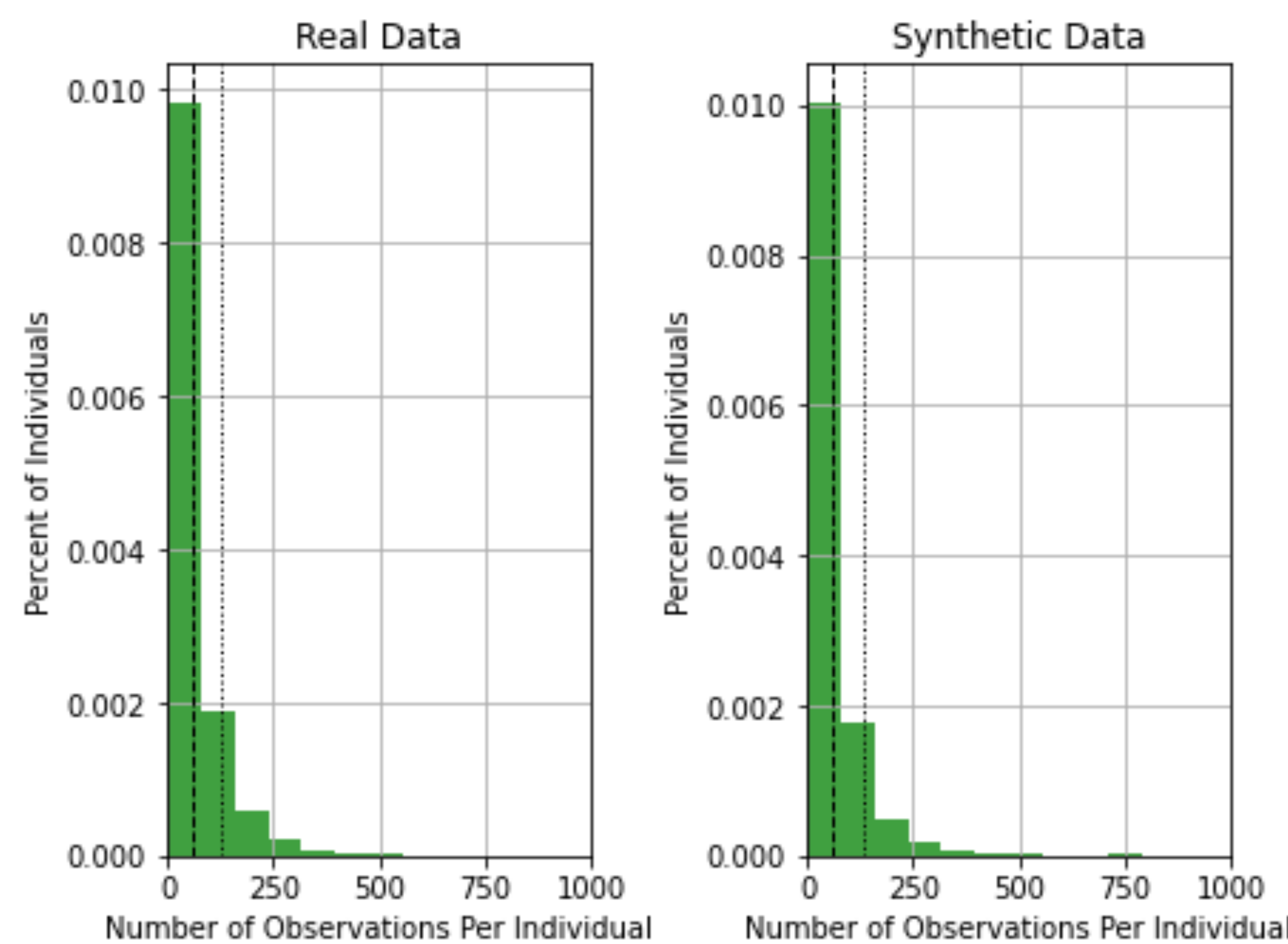
- Comparison of the number of events per patient
  - Number of certain types of events (e.g., prescriptions) per patient
  - Limit the above to a certain time interval
- Comparison of the overall frequency of events
- Comparisons of event distributions across classes of events using univariate distribution comparison metrics
- Evaluation of the k-order transition matrices among events or classes of events

Replica Analytics

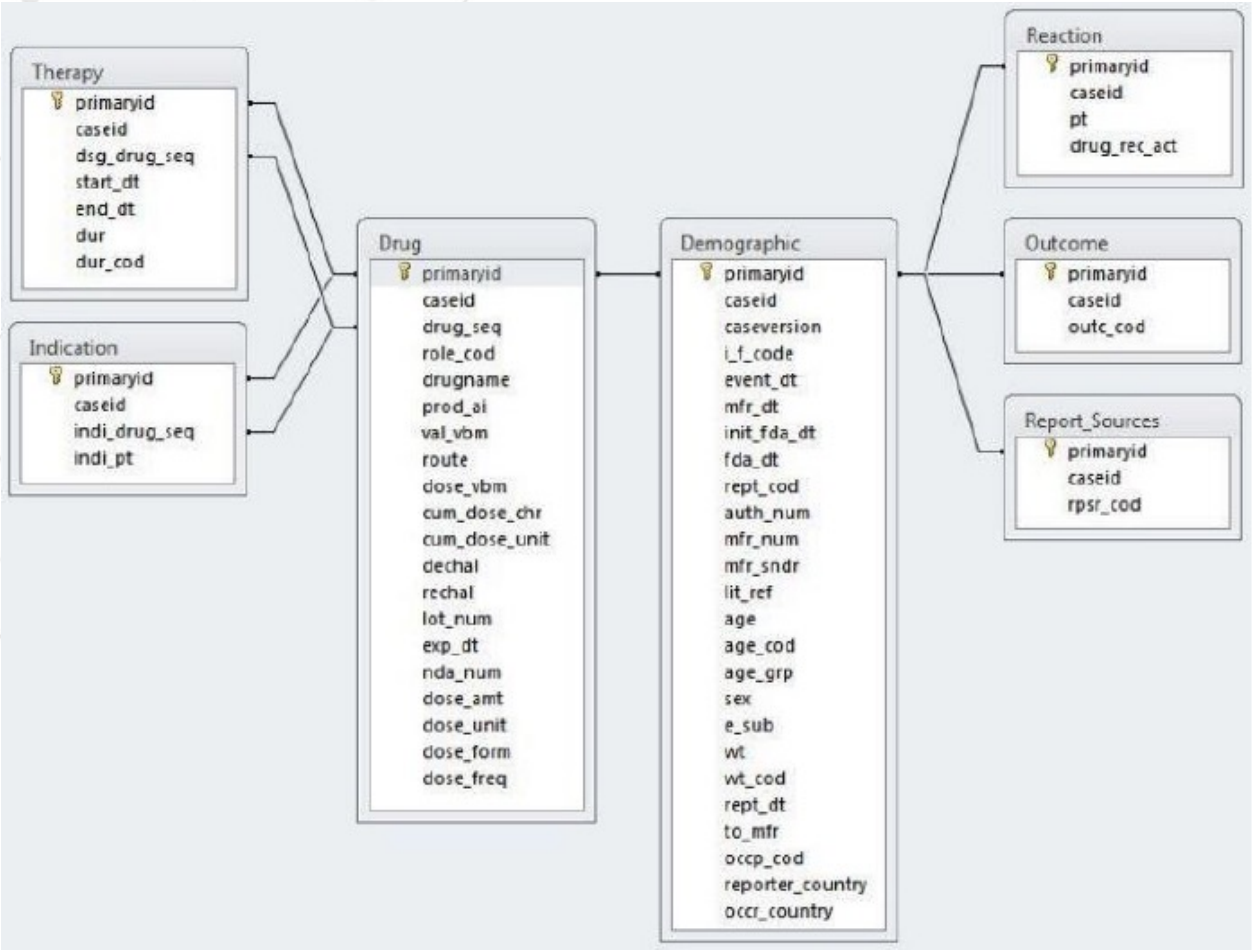

Real Transition Matrix
Synthetic Transition Matrix
Next Event
Current Event
AntidepRX
Death
EDVisit
Hospitalization
LabTest
LastObs
MorphineRX
OxycodoneRX

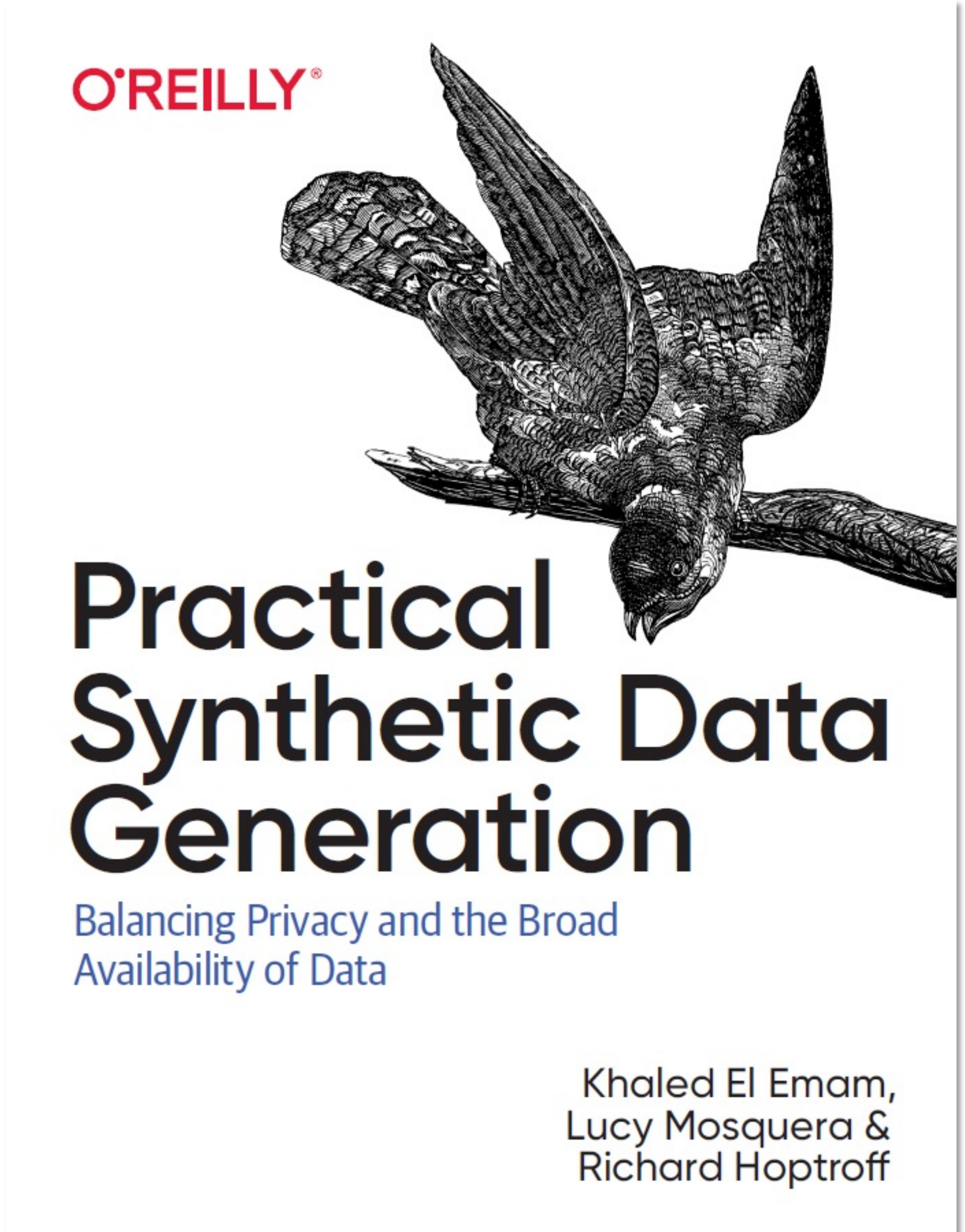

Real Data
Synthetic Data
Frequency of Observations
Event Type
AntidepRX
Death
EDVisit
Hospitalization
LabTest
LastObs
MorphineRX
OxycodoneRX


Hellinger Distance
Event Attribute
sojourn_time_cat
dspn_amt_cat
dspn_day_cat
dxcode1_ED_cat
cacs_riw_cat
LOS_cat
dxcode1_hosp_cat
riw_cat
u_cause_cat
test_name
test_results_cat


Real Data
Synthetic Data
Percent of Individuals
Number of Observations Per Individual

Hierarchical datasets require a different approach





Introductory Book on Data Synthesis

Published in 2020

Replica Analytics


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

Use Case: Analyzing Longitudinal Hospital Discharge Data

Replica Analytics

Roles

Claire (Researcher)

Claire is a researcher who is interested in assessing high-cost hospitalizations with lengths of stay greater than 5 days.

Claire puts in a request for access to data to the data provider

Alice

Alice represents the data provider and is authorized to access personal health information. She has a computing background and works in the IT department supporting the data scientists and researchers.

She receives data requests from users for research purposes.

Replica Analytics

Use Case

Alice can provide synthetic data for Clare for research purposes as:

1) A specified cohort of key features
2) Raw longitudinal data

We will illustrate both these use cases

Replica Analytics

Case 1: Synthesis of a Cohort

Data request with analytic cohort definition:
- Hospitalizations with LOS >5
- Highest cost stay per person

Build analytic cohort from raw data

Synthesize the data

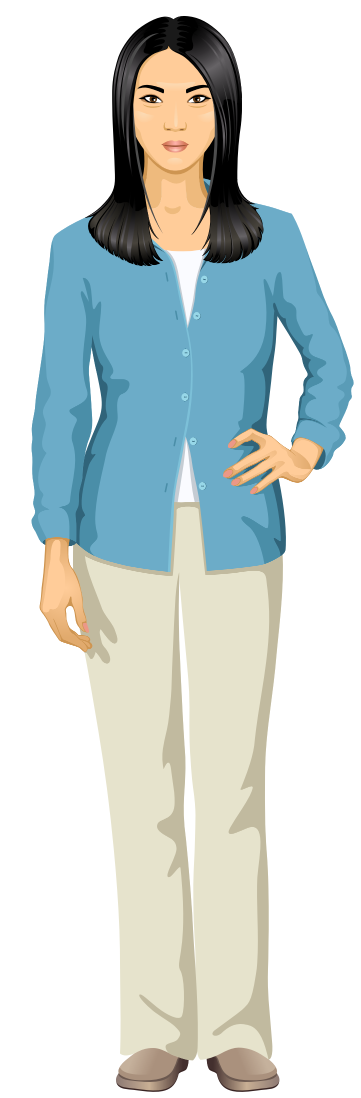
Send synthetic data to the data user

Conduct analysis

Replica Analytics

Case 2: Synthesis of Raw Longitudinal Data

Data request



Synthesize the entire longitudinal dataset

Send synthetic data to the data user

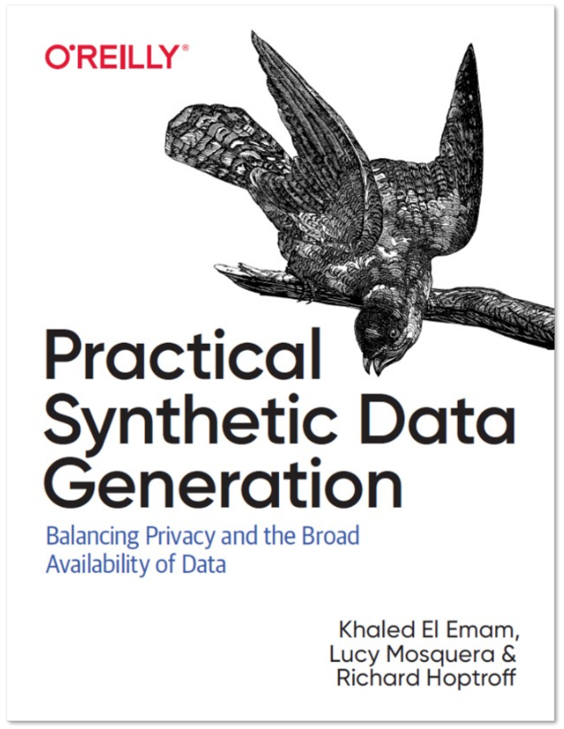
Build analytic cohort:
- Hospitalizations with LOS >5
- Highest cost stay per person

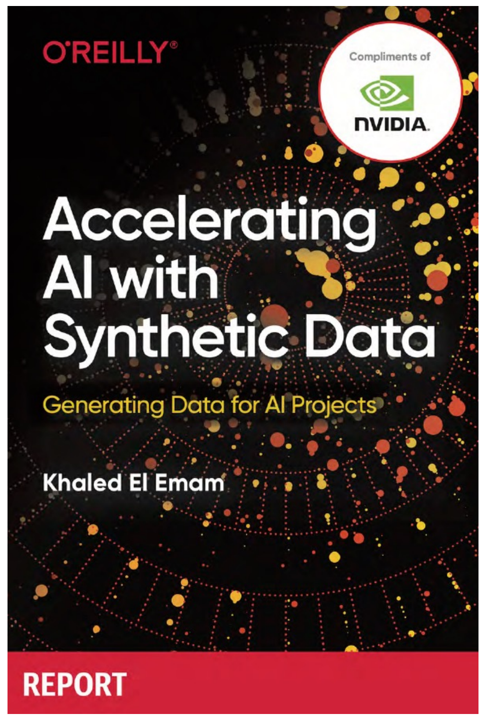
Conduct analysis

cc: an untrained eye - https://www.flickr.com/photos/26312642@N00

To Learn More

- Join our mailing list: https://bit.ly/3gRVAIi
- Follow us on Linkedin: https://bit.ly/2XS3KHF
- Listen to our comprehensive on-line tutorials on data synthesis: https://bit.ly/2TXI0Jy
- Read our introductory report and book on the topic





Replica Analytics