Utility Assessments in Synthetic Data

Lucy Mosquera & Xi Fang

April 27th, 2022

Agenda

Introduction to Synthetic Data — 1 — General description of what synthetic data is

Introduction to Utility Assessments — 2 — An overview of state-of-the-art ways to measure utility in synthetic data

Publication Results — 3 — Simulation results to assess the relationship between generic and workload aware utility assessments

Replica Analytics
AN AETION COMPANY

Synthetic Data

Real Data

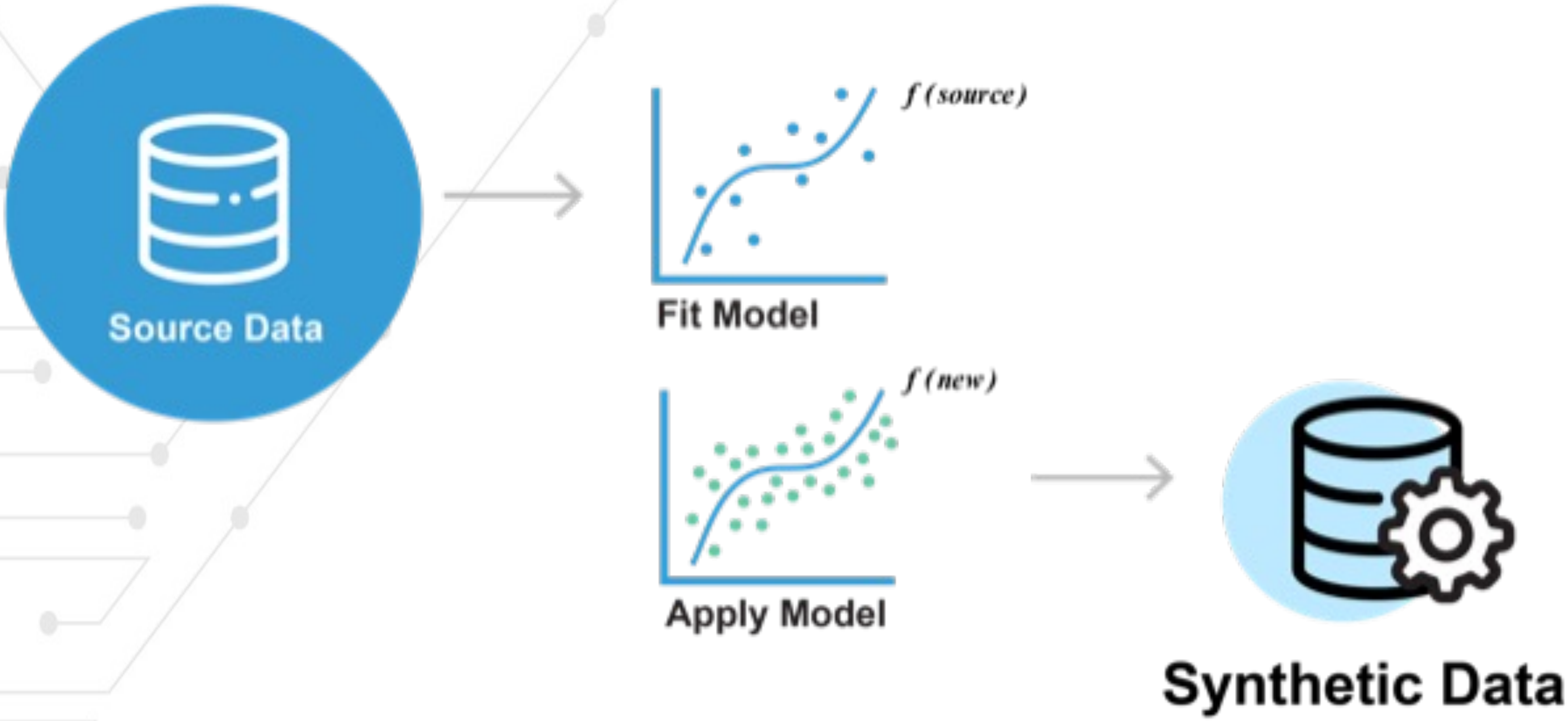
COU1A	AGECAT	AGELE70	WHITE	MALE	BMI
United States	3	1	0	1	25.44585
United States	3	1	1	0	24.09375
United States	3	1	1	1	33.07829
United States	2	1	1	0	33.64845
United States	3	1	1	0	25.66958
United States	3	1	1	0	25.85938
United States	2	1	1	0	24.7357
United States	5	0	0	0	27.75276
United States	5	0	1	1	28.07632

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

Synthetic Data

Replica Analytics
AN AETION COMPANY

The Synthesis Process



Replica Analytics
AN AETION COMPANY

Utility Assessment

Aim to convey how similar synthetic data is to the real data is has been generated from

Utility assessments can be performed in different ways that convey different information to data users; there currently is not a consistent industry-wide standard

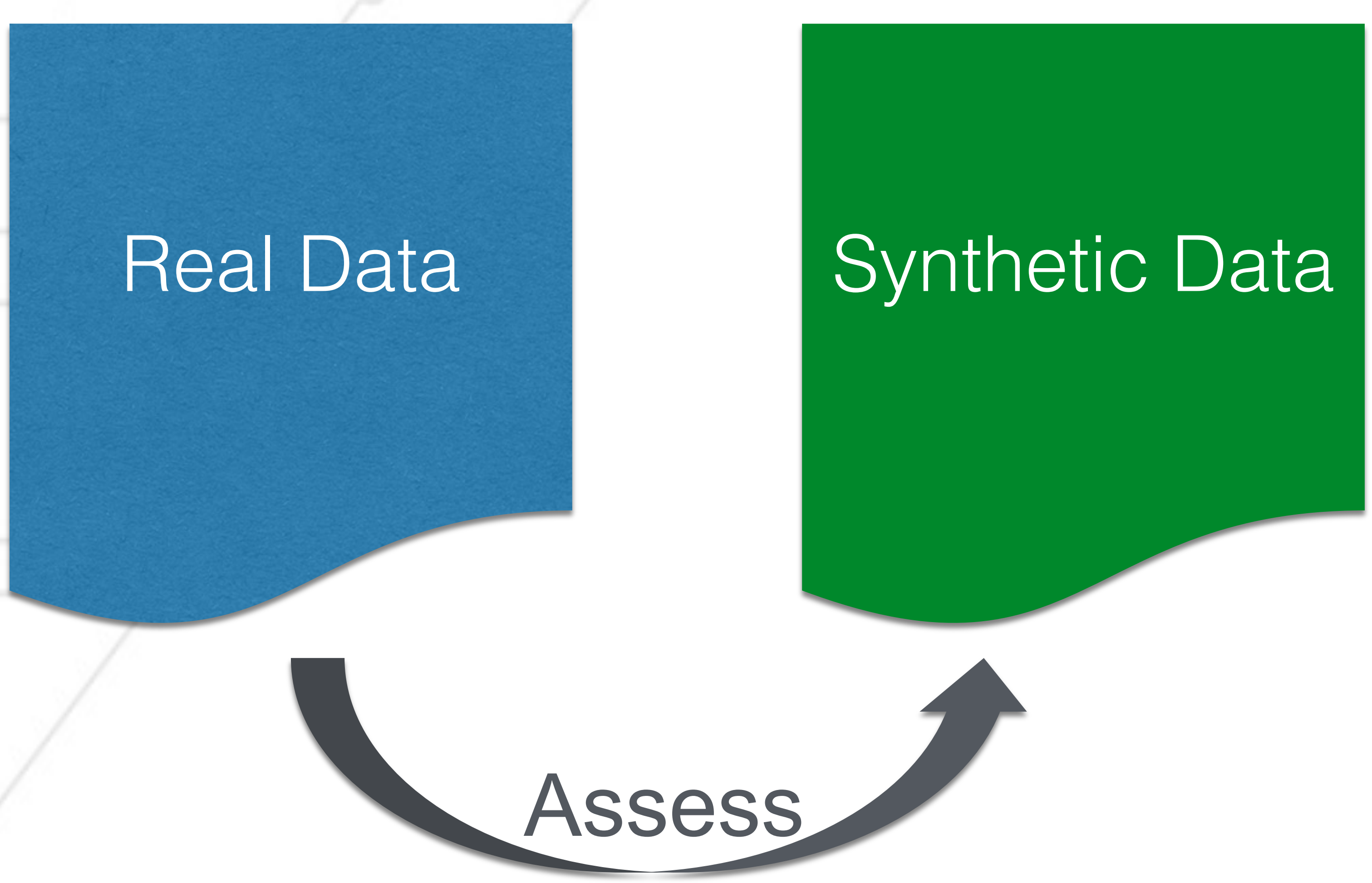
Workload Aware vs Generic

- Workload aware utility assessment illustrate how well synthetic data can be used as a drop-in replacement or proxy for real data for a specific analysis
- Generic or broad utility assessments show how similar synthetic data is to the real data it was generated from without referencing a specific analysis

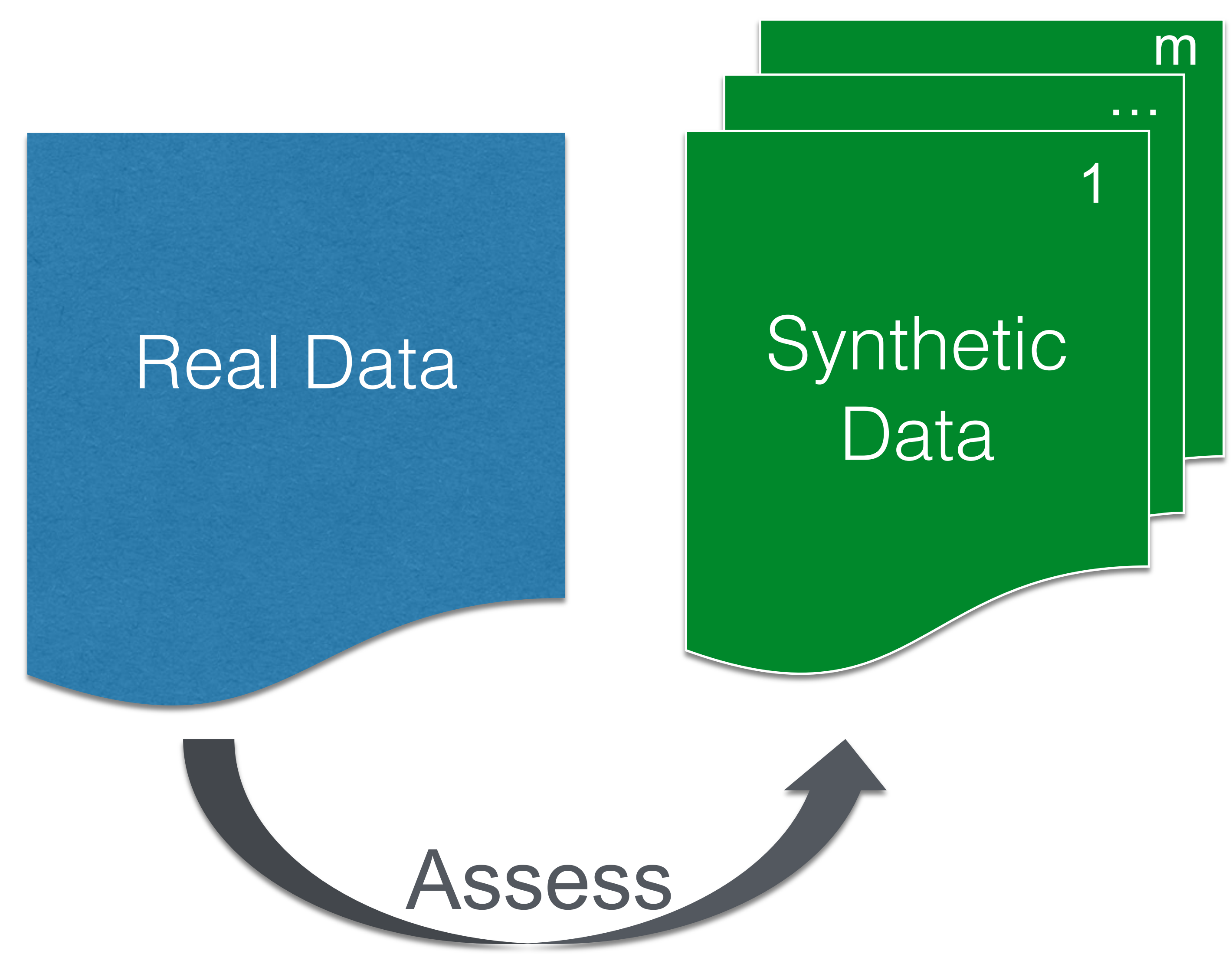
Replica Analytics
AN AETION COMPANY

Dataset vs Generative Model

Specific to a given dataset

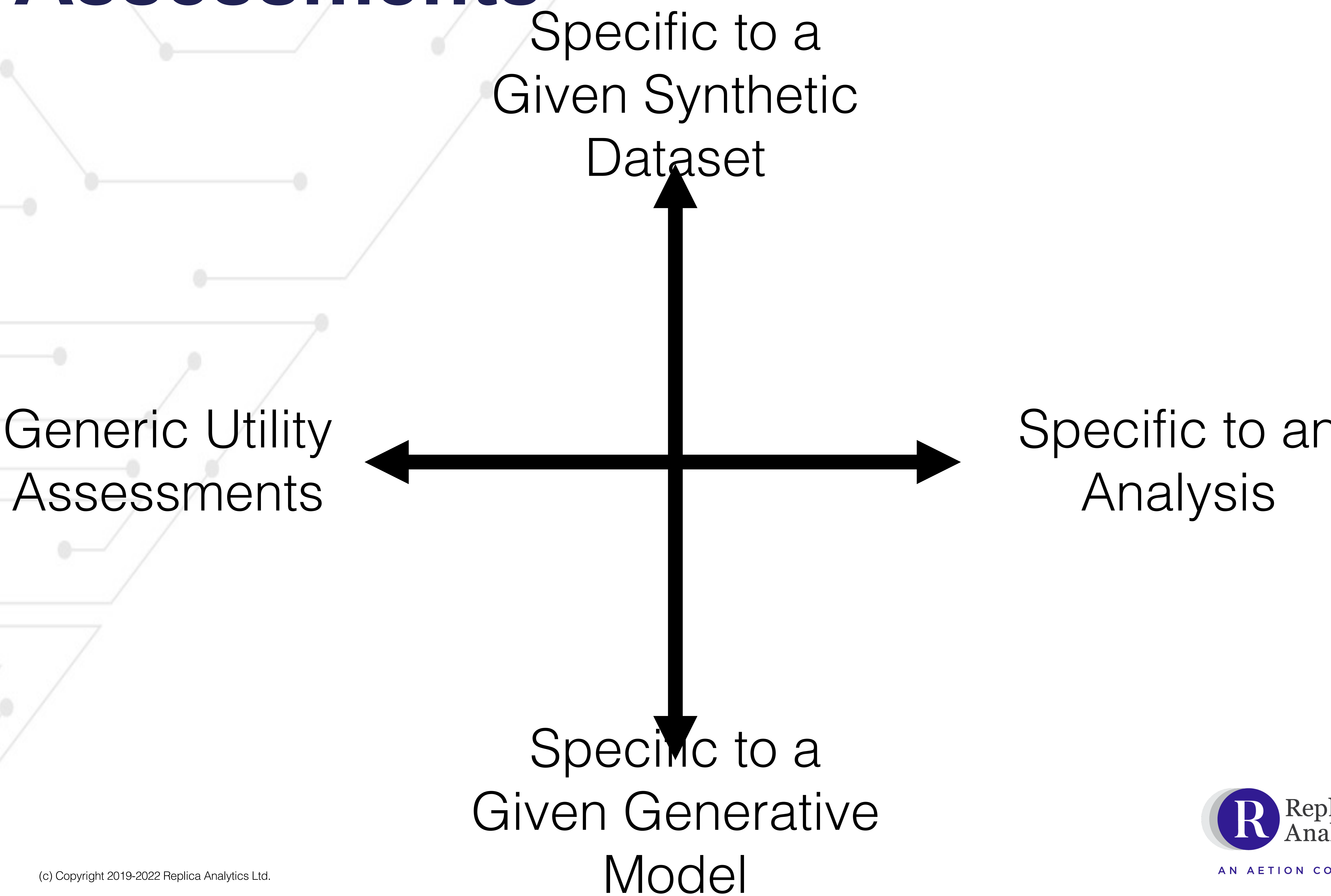


Representative of a generative model



Replica Analytics
AN AETION COMPANY

Continuum of Utility Assessments

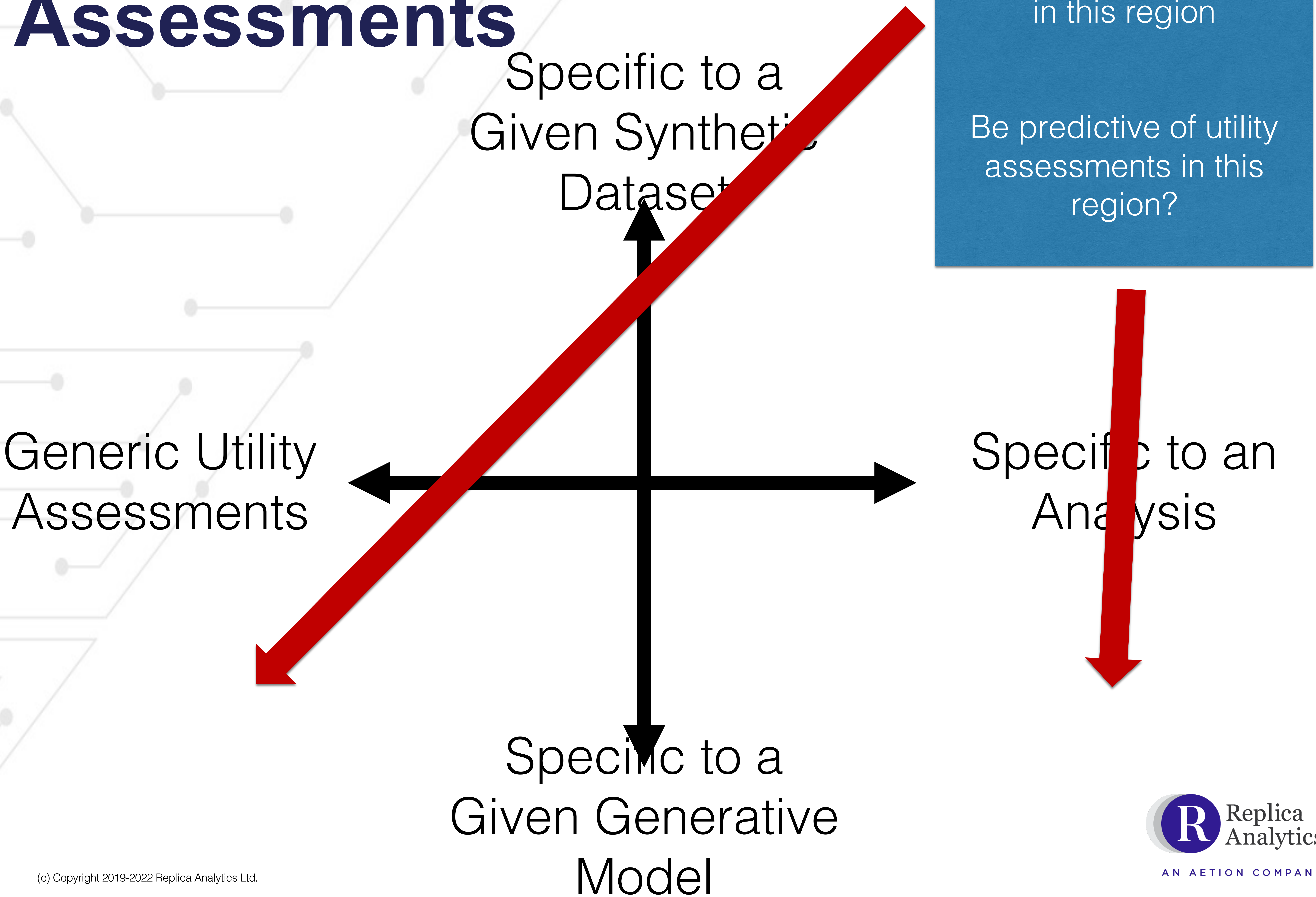


Replica Analytics
AN AETION COMPANY

Continuum of Utility Assessments

Can utility assessments in this region

Be predictive of utility assessments in this region?



Replica Analytics
AN AETION COMPANY

Other Utility Assessment Strategies

- Focus on marginal distributions of specific variables or joint distributions of sets of variables (e.g., Hellinger distance or bivariate correlation)
- Use subject area experts to attempt to select the synthetic records from a mixed dataset

Our Work

JMIR MEDICAL INFORMATICS El Emam et al

Original Paper

Utility Metrics for Evaluating Synthetic Health Data Generation Methods: Validation Study

Khaled El Emam[1,2,3], BEng, PhD; Lucy Mosquera[2,3], BA, MSc; Xi Fang[3], BA, MSc; Alaa El-Hussuna[4], MSc, MD

[1]School of Epidemiology and Public Health, University of Ottawa, Ottawa, ON, Canada
[2]Children's Hospital of Eastern Ontario Research Institute, Ottawa, ON, Canada
[3]Replica Analytics Ltd, Ottawa, ON, Canada
[4]Open Source Research Collaboration, Aarlberg, Denmark

Goal: evaluate how well common utility metrics can rank SDG methods according to performance on a logistic regression prediction models

Assessment Strategy

- Using 30 different health datasets, which utility assessment can be used to reliably rank 3 different synthetic data generation (SDG) methods in terms of their performance on a specific logistic regression analysis
- For each dataset, each SDG method generated 20 copies and the generic utility performance was averaged across all copies

Replica Analytics
AN AETION COMPANY

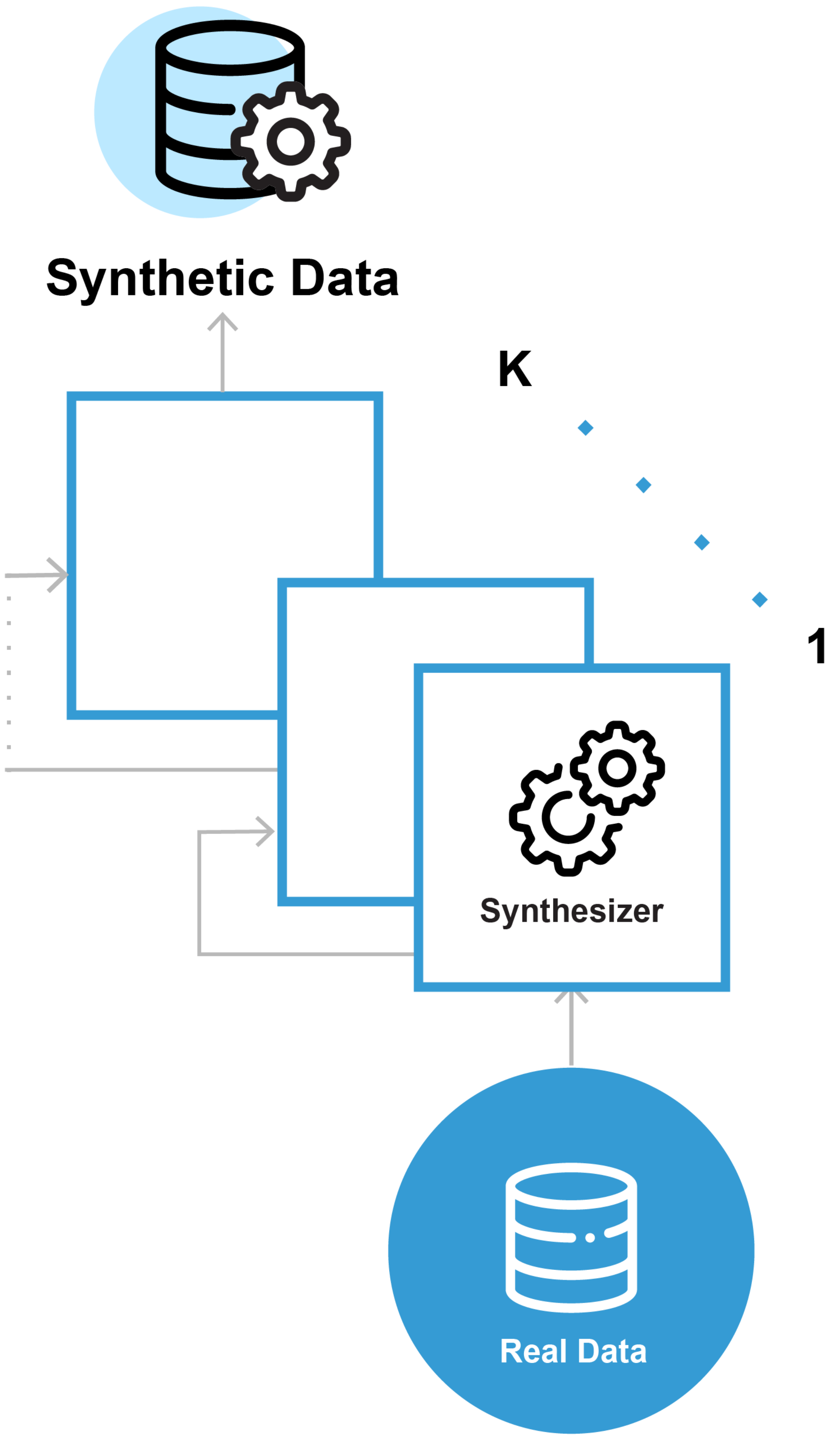
Synthetic Data Generating Methods

Three SDG methods that have very different approaches:

- Bayesian network
- Generative adversarial network
- Sequential tree synthesis

Sequential Tree Synthesis

Flexible synthesis model where variables are synthesized in a sequence.

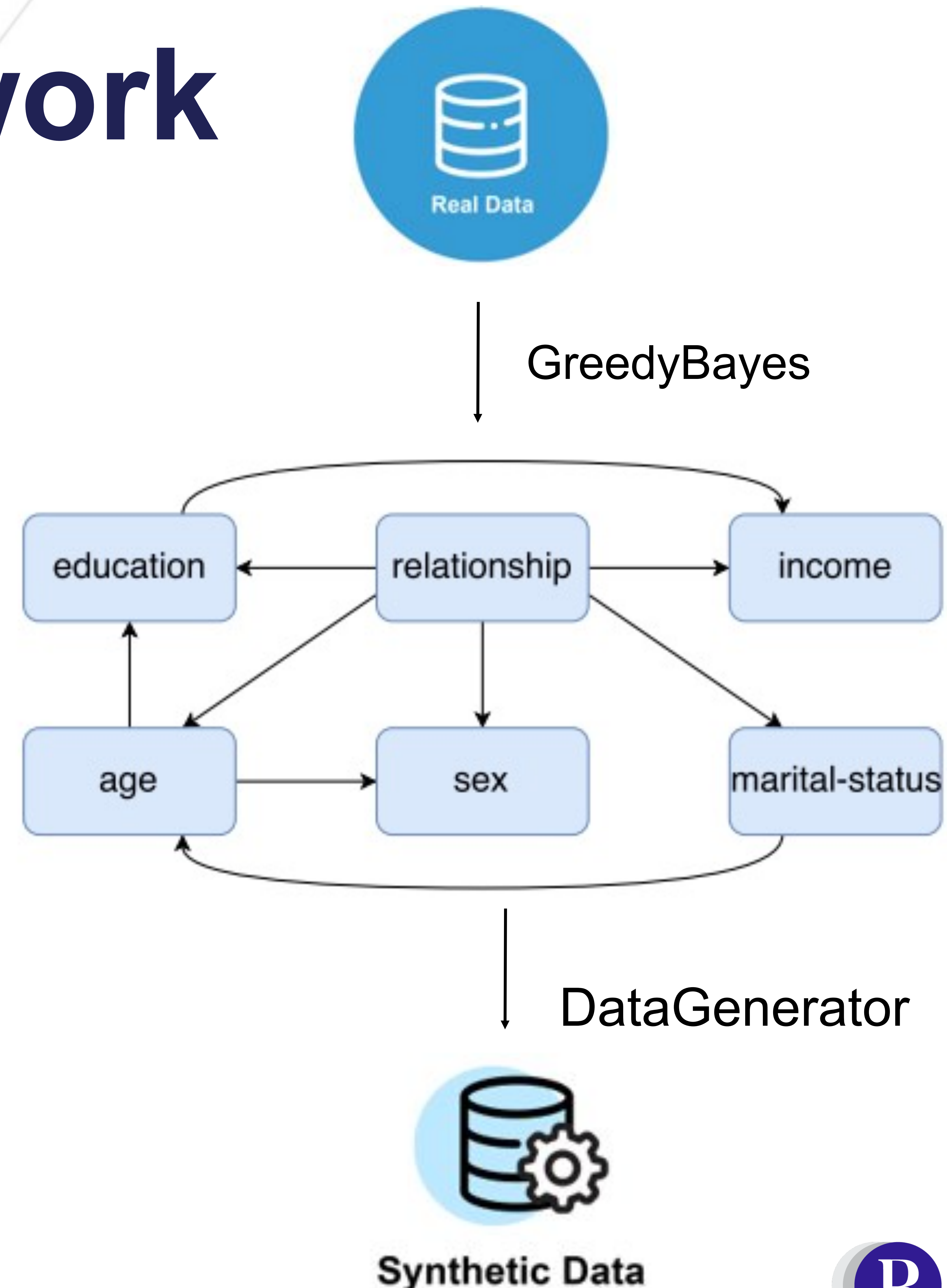


Replica Analytics software

Replica Analytics
AN AETION COMPANY

Bayesian Network

- Build Bayesian Network ([directed acyclic graphs](#))
- Input: a set of values from the node's parent variables
- Output: the probability of the variable represented by the node

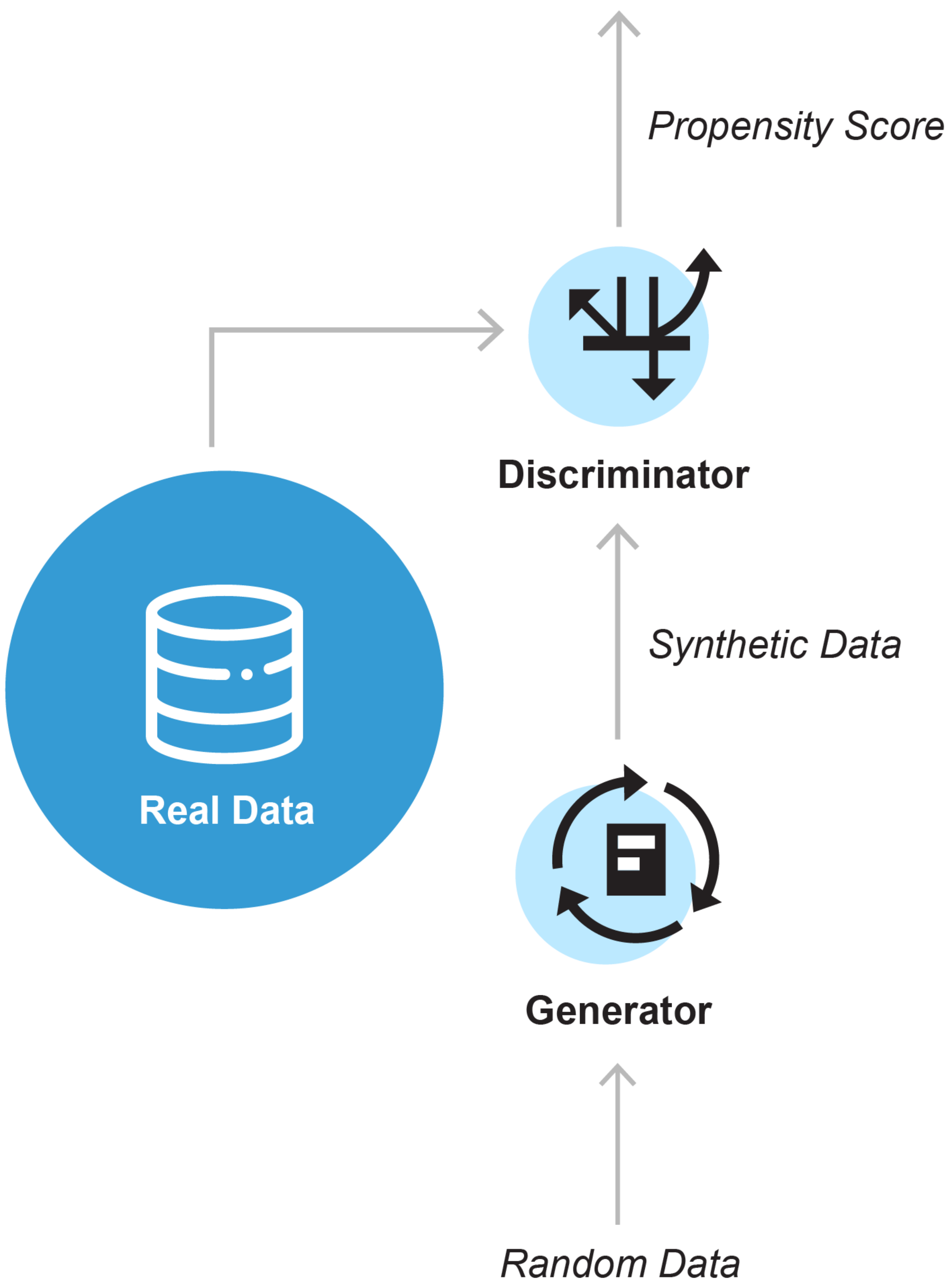


Replica Analytics
AN AETION COMPANY

Generative Adversarial Network

Synthesis model that iteratively trains two neural networks in an arms race to:

1. Generate more representative synthetic data
2. Discriminate real records from synthetic more accurately



Replica Analytics
AN AETION COMPANY

Workload Aware Utility Assessment

- Logistic regression where model performance is assessed using AUROC and AUPRC
- Consider AUROC & AUPRC differences as the workload aware utility metrics
- Designed to represent a typical analysis for health data

Generic Utility Assessments

Focused on assessments that look at the synthetic dataset as whole:

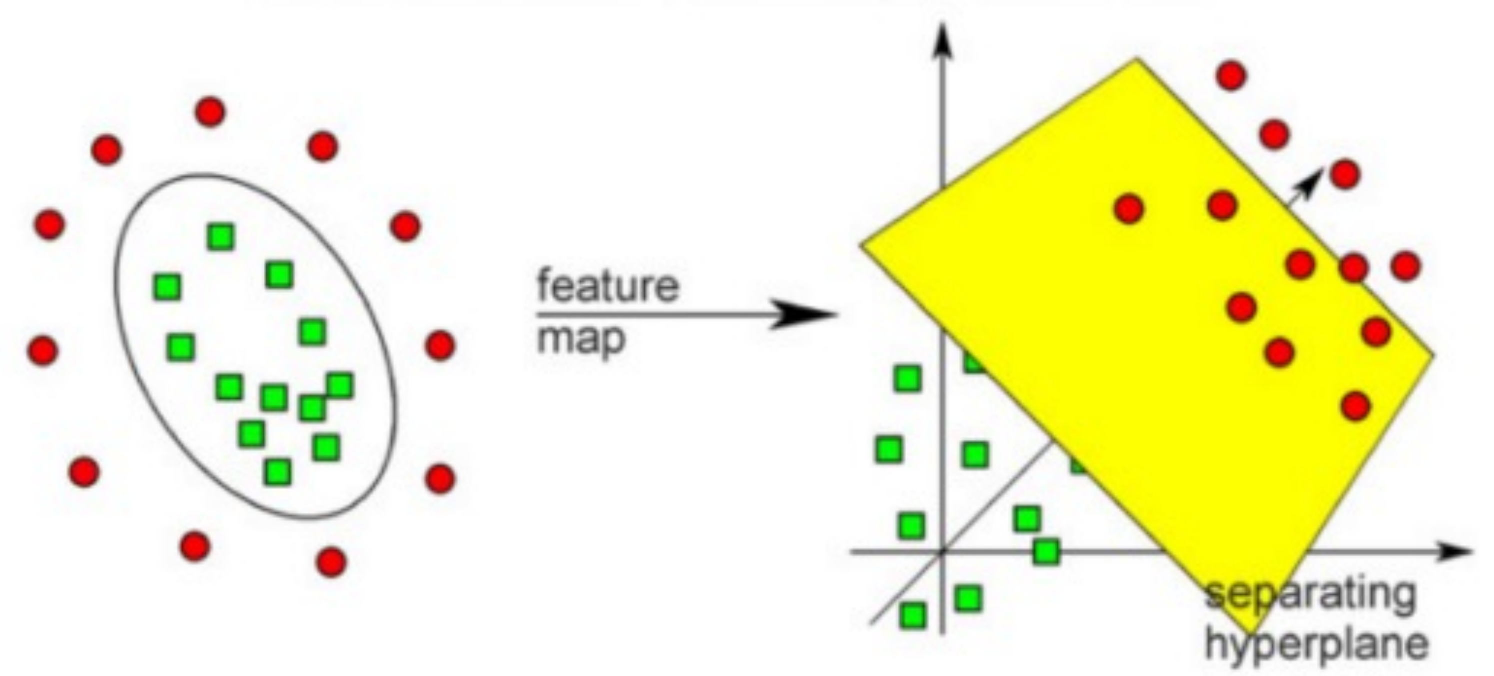
- Maximum Mean Discrepancy
- Multivariate Hellinger Distance
- Wasserstein Distance
- Cluster Analysis Measure
- Distinguishability Metrics (pMSE)

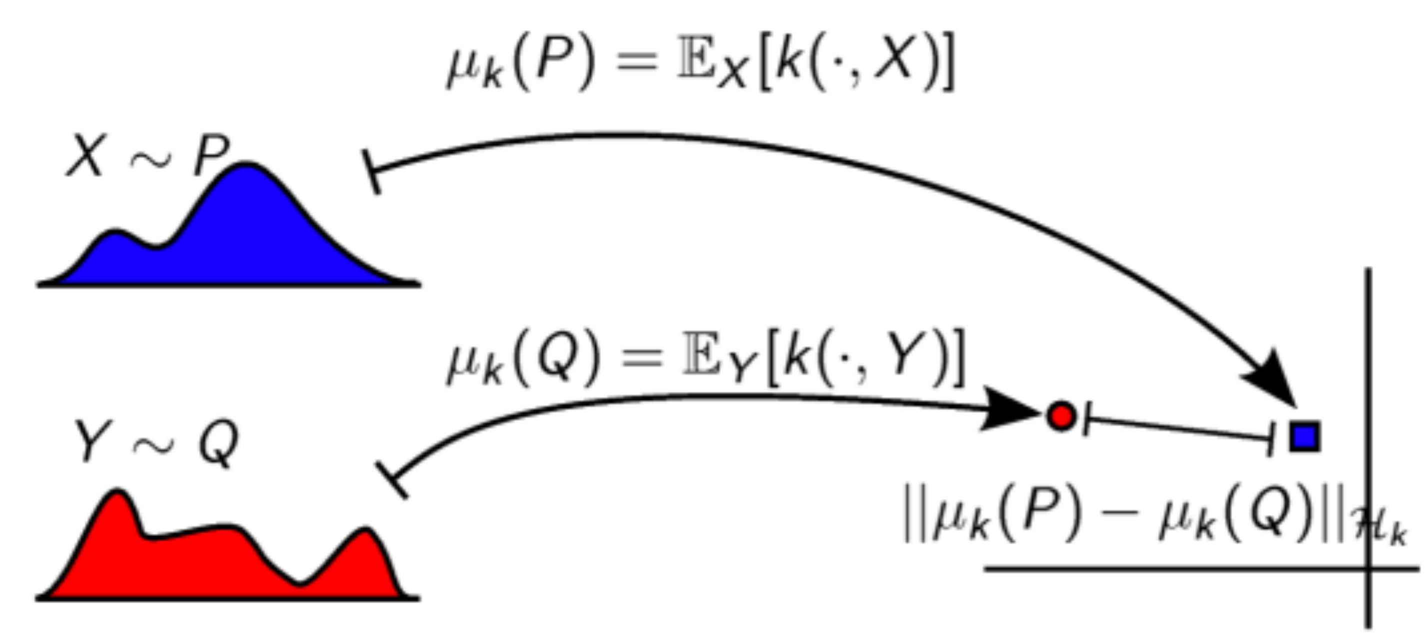
Replica Analytics
AN AETION COMPANY

Maximum Mean Discrepancy

The maximum mean discrepancy(MMD) is proposed by Gretton et al. [1] to test whether the samples are from different distributions.

Empirical Estimate, let **K** be a class of smooth function:





[Cortes & Vapnik, 1995; Schölkopf & Smola, 2001]

$$MMD[\mathrm{k}, X, Y] = sup_{\mathrm{k} \in K}\left(\frac{1}{m}\sum_{i=1}^{m} k(x_i) - \frac{1}{n}\sum_{i=1}^{n} k(y_i)\right)$$

Replica Analytics
AN AETION COMPANY

Multivariate Hellinger Distance

The Hellinger distance is used to quantify the similarity between two probability distributions. It can be derived from the multivariate normal Bhattacharyya distance

- Bhattacharyya distance: the degree of dissimilarity between two probability distributions

$$D_B(p,q) = -\ln\left(BC(p,q)\right) \qquad BC(p,q) = \int \sqrt{p(x)q(x)}dx$$

- Hellinger distance: the degree of similarity between two probability distributions

$$H^2(p,q) = \frac{1}{2}\int(\sqrt{p(x)} - \sqrt{q(x)})^2 dx = 1 - \int \sqrt{p(x)q(x)}dx$$

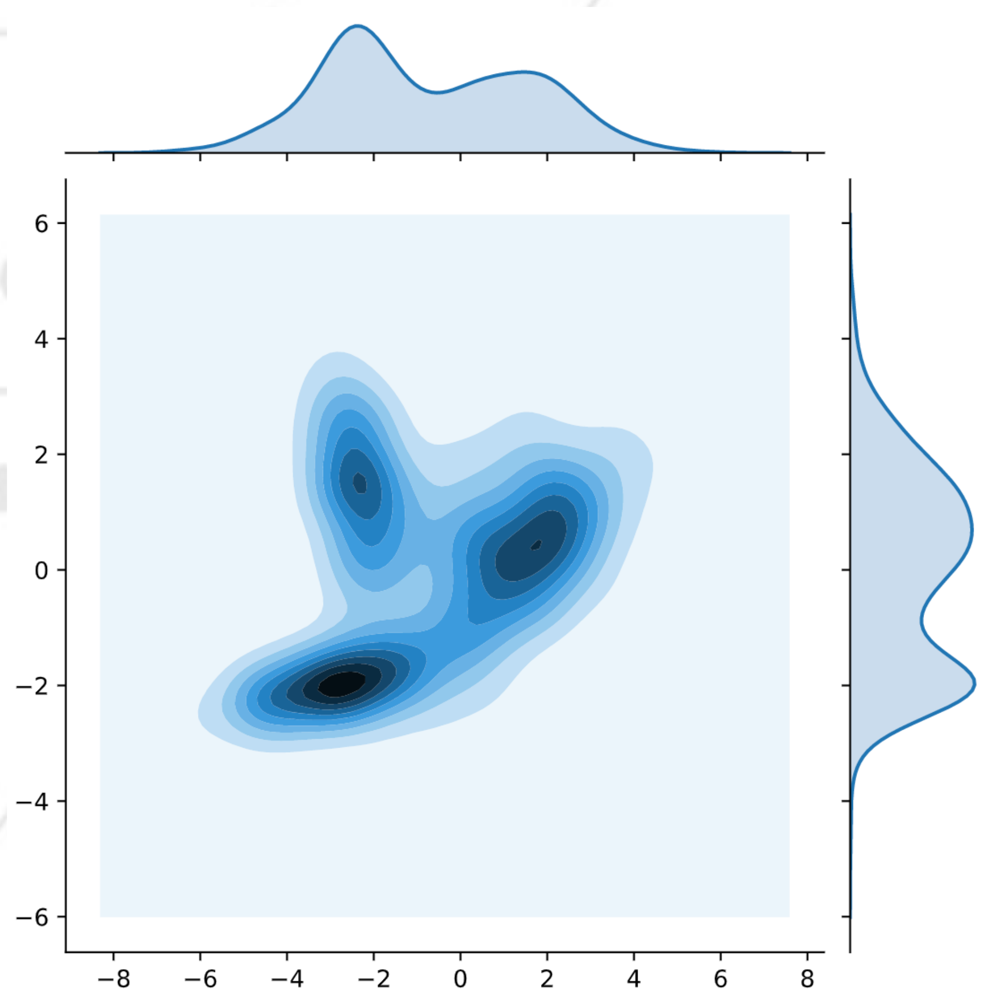
bound between 0 and 1, more interpretable

Replica Analytics

AN AETION COMPANY

Wasserstein Distance

For a distribution of mass p(x) on a space X, we wish to transport the mass in such a way that it is transformed into the distribution q(x) on the same space.

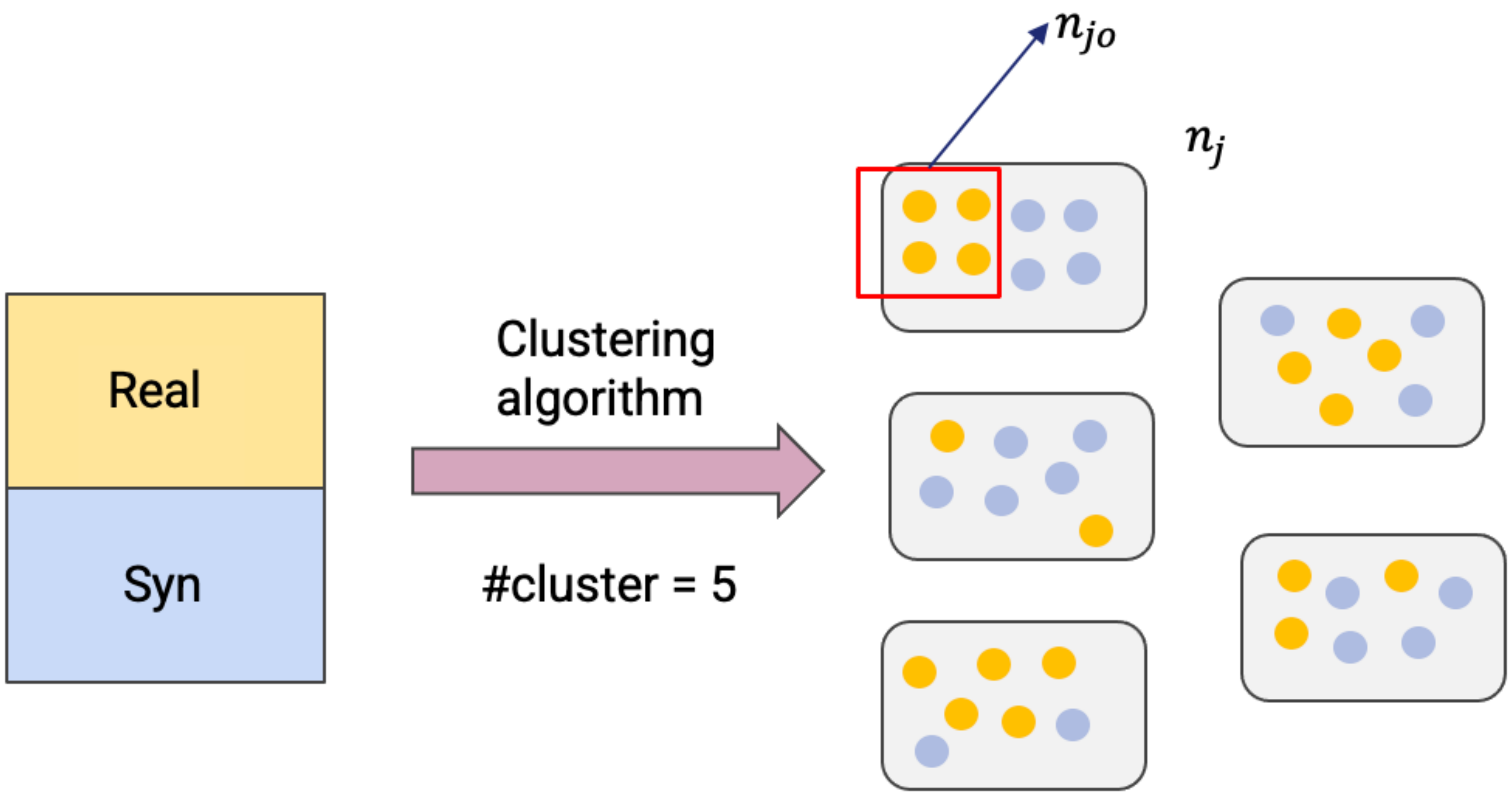


$$C = inf_{\gamma \in \Gamma(p,q)} \int c(x,y) d\gamma(x,y)$$

Given a cost function, the optimal transport plan is the plan with the minimal cost out of all possible transport plans. If the cost of a move is simply the distance between the two points, then the optimal cost is identical to the definition of the ***W1 distance***.

Replica Analytics
AN AETION COMPANY

Cluster Analysis Measure

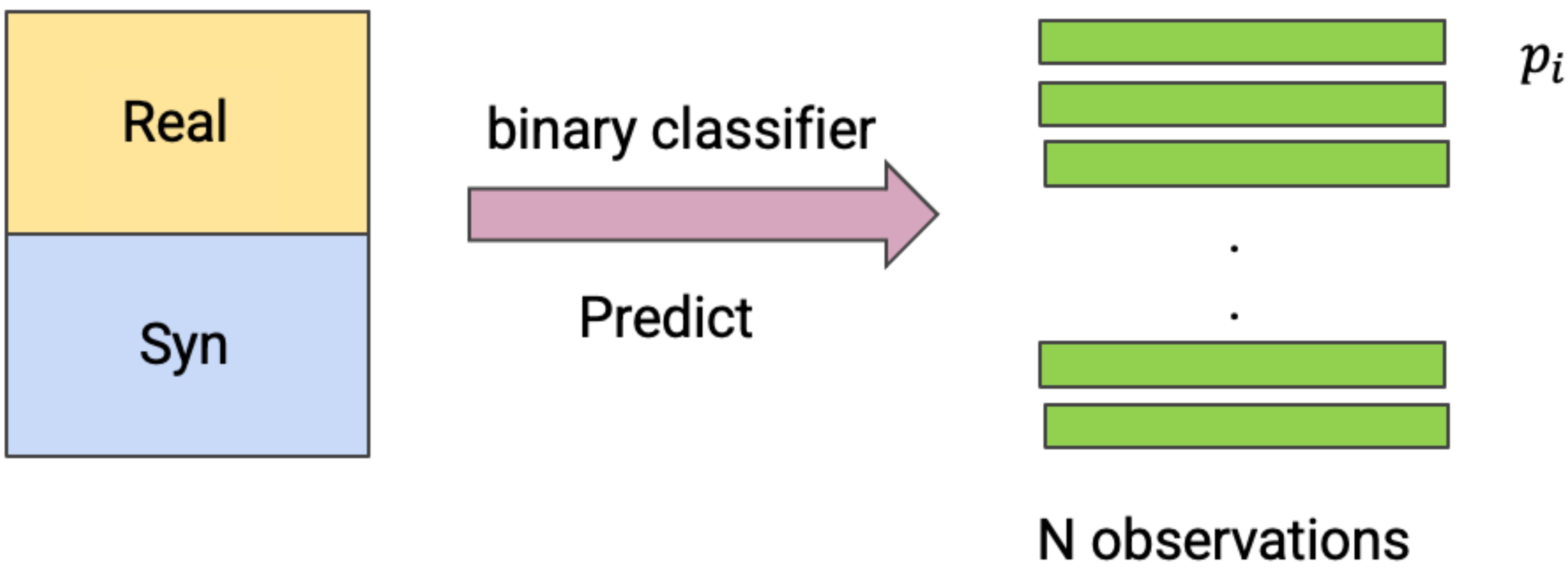


$$U_c = \frac{1}{G}\sum_{j=1}^{G} w_j \left[\frac{n_{jo}}{n_j} - c\right]^2 \qquad c = \frac{N_O}{N_O + N_M}$$

Where nj denotes number of observations in the jth cluster

Replica Analytics
AN AETION COMPANY

Distinguishability Metric



$$propensityMSE = \frac{1}{N}\sum_i (p_i - 0.5)^2$$

Page Test

Adult Data	Multivariate Hellinger Distance	Group	AUROC DIFF	AUPRC Diff
RA	0.2	L	0.1	0.2
CTGAN	0.5	M	0.3	0.3
BN	0.7	H	0.5	0.4

$$H0_{AUROC}\text{: median(AUROC\_Diff}_L) = \text{median(AUROC\_Diff}_M) = \text{median(AUROC\_Diff}_H)$$

$$H1_{AUROC}\text{: median(AUROC\_Diff}_L) > \text{median(AUROC\_Diff}_M) > \text{median(AUROC\_Diff}_H)$$

Replica Analytics
AN AETION COMPANY

Results

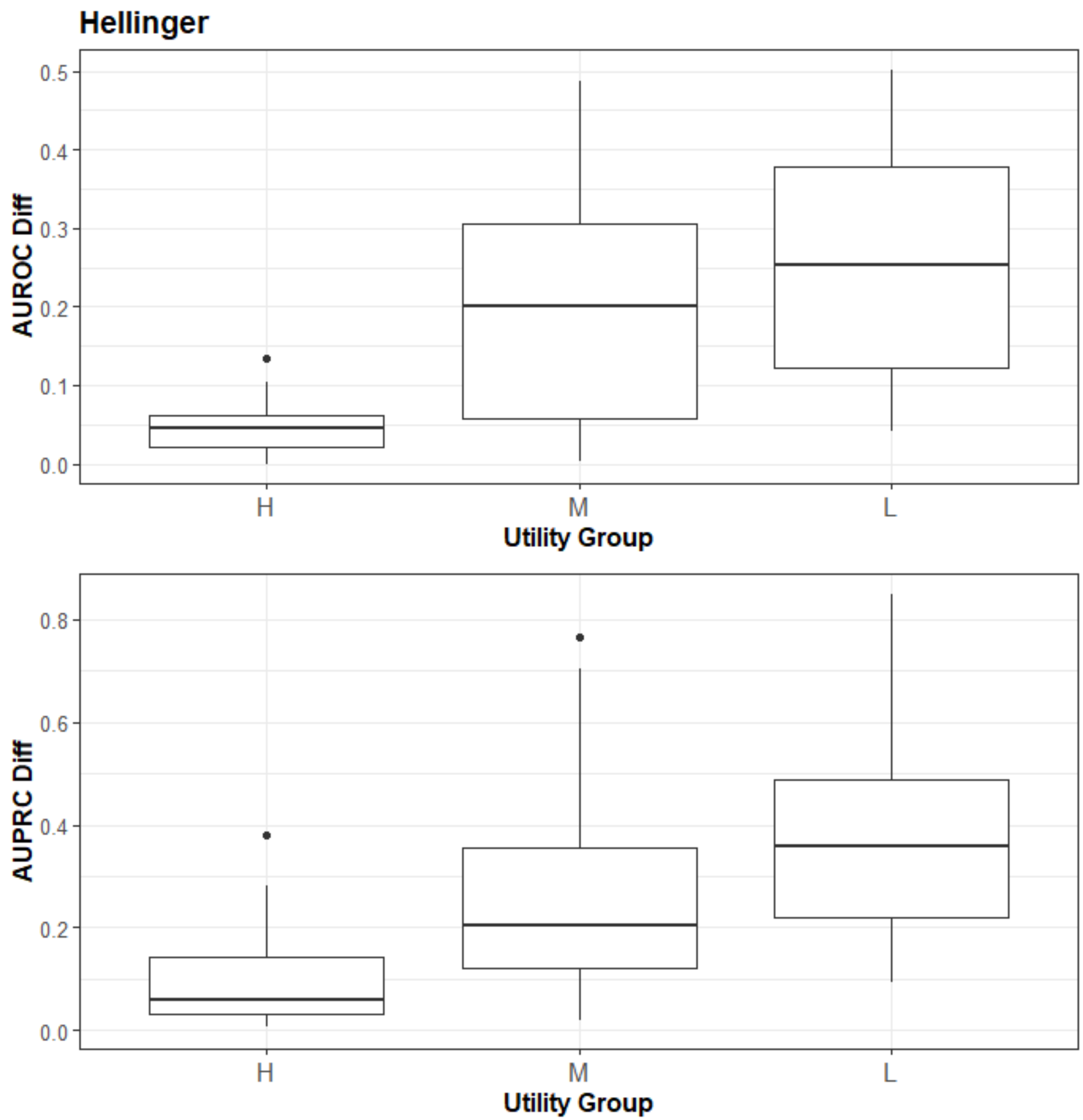
Table 1. Page test results for each of the utility metrics and prediction accuracy

Utility metric	AUROC[a] difference		AUPRC[b] difference	
	L value	*P* value	*L* value	*P* value
Maximum mean discrepancy	384	.00104[c]	392	<.001[c]
Hellinger distance[d]	$\boxed{398}$	<.001[c]	409	<.001[c]
Wasserstein distance	392	<.001[c]	403	<.001[c]
Cluster analysis	396,	<.001[c]	405	<.001[c]
Propensity mean squared error	390	<.001[c]	394	<.001[c]

- The test statistic (L) indicates the strength of the ordering of data. The Hellinger distance had the highest *L* value, suggesting that it has an advantage in ordering the SDG methods

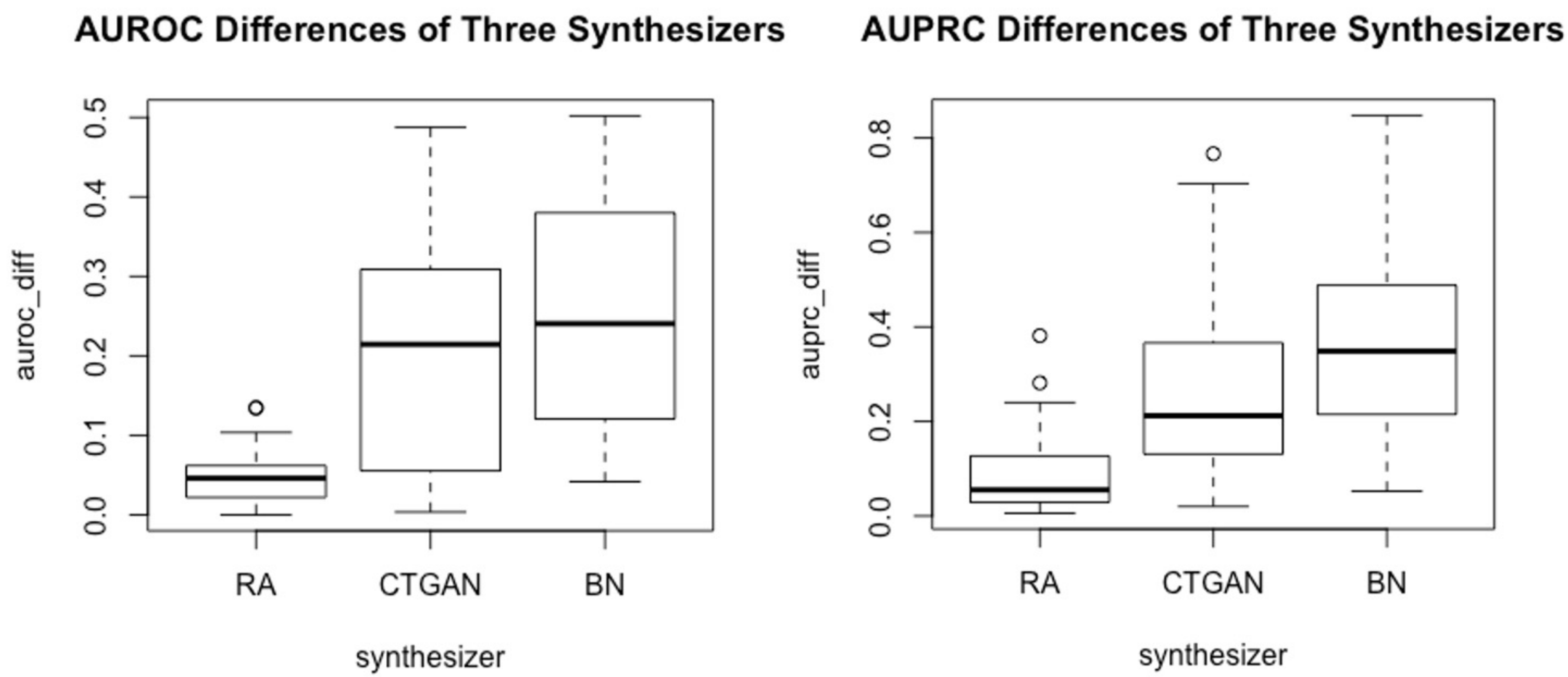
Replica Analytics
AN AETION COMPANY

Results: Multivariate Hellinger Distance



Replica Analytics
AN AETION COMPANY

Results



Comparing SDG methods, the RA sequential tree method produced data with the smallest differences in regression model performance

Replica Analytics
AN AETION COMPANY

Conclusions

- Many generic utility metrics can be predictive of analysis specific utility
- Multivariate Hellinger distance was the most predictive generic utility assessment considered by a small margin
- Sequential tree synthesis led to synthetic data with the smallest differences in predictive ability on average

Shows that generic utility metrics can be used to select the best generative model for a given analysis

Replica Analytics
AN AETION COMPANY

Limitations

- Use case was ranking SDG methods
- Workload aware assessment was logistic regression
- Did not assess privacy implications

Replica Analytics
AN AETION COMPANY

Future Work

- Extend these results from selecting a SDG method to hyperparameter tuning within a method (e.g., tree depth in sequential trees)
- Assess combined metric that includes utility and privacy

Replica Analytics
AN AETION COMPANY

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

Discussion Questions

- How much better is hellinger distance compared to the other generic assessments?

Discussion Questions

- What happens if you try to combine these metrics and create an aggregate?

Discussion Questions

- What if you vary the number of copies of synthetic datasets generated for each SDG?

Replica Analytics
AN AETION COMPANY

Discussion Questions

- What if the synthesized datasets were a different size than the original datasets?