On the Evaluation of Synthetic Data utility

Presentation of prior research (2021-2022)

Fida Dankar

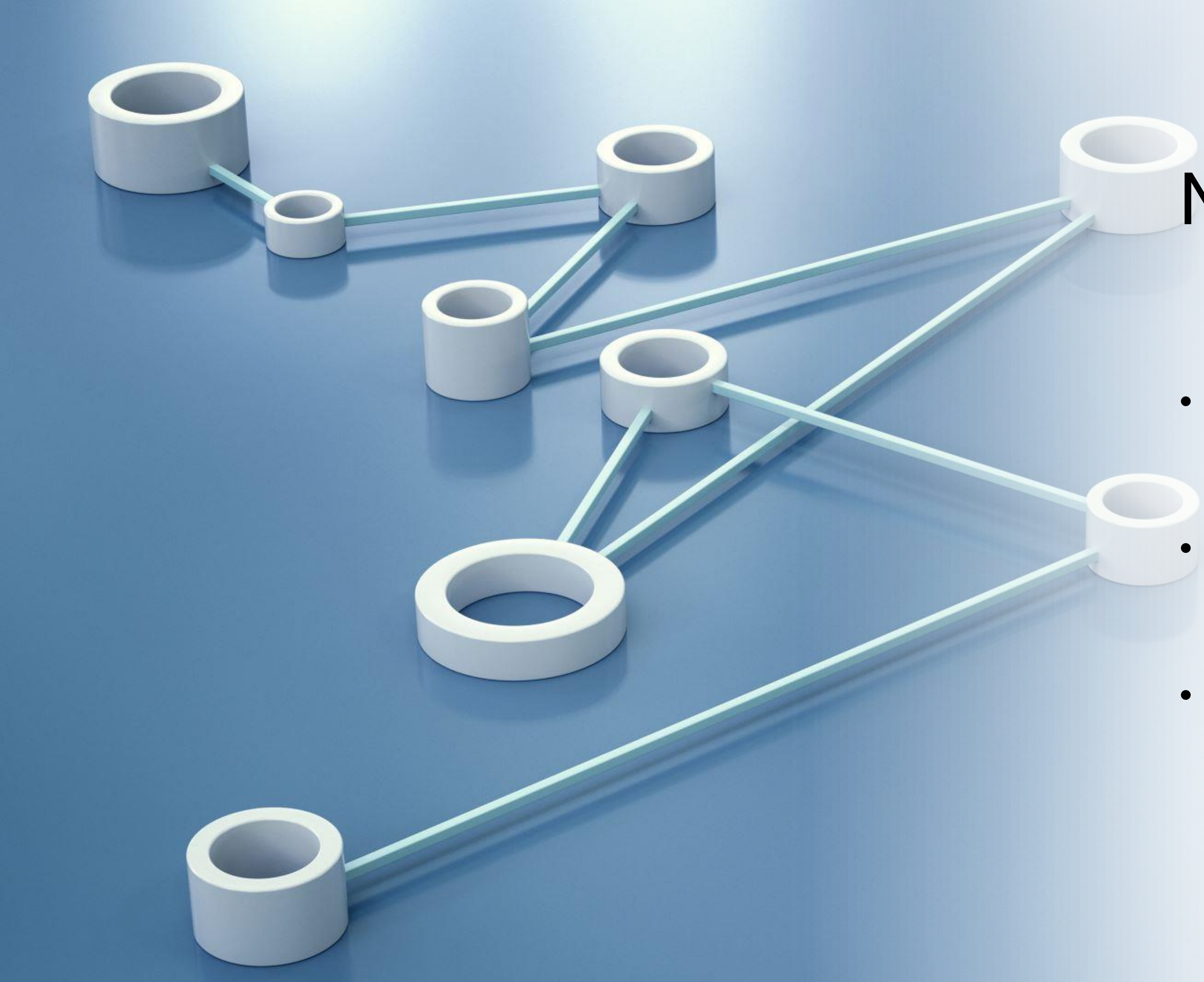
CHEO Research Institute

Agenda

1. PROBLEM DEFINITION

2. A MULTIDIMENSIONAL EVALUATION OF 4 SDG

3. A PCA-BASED MEASURE OF UTILITY



Motivation

- Accessing personal data is often challenging and time-consuming
- An increasingly popular way to overcome these issues is fully synthetic data.
- However, empirical evidence of their utility has not been fully explored.

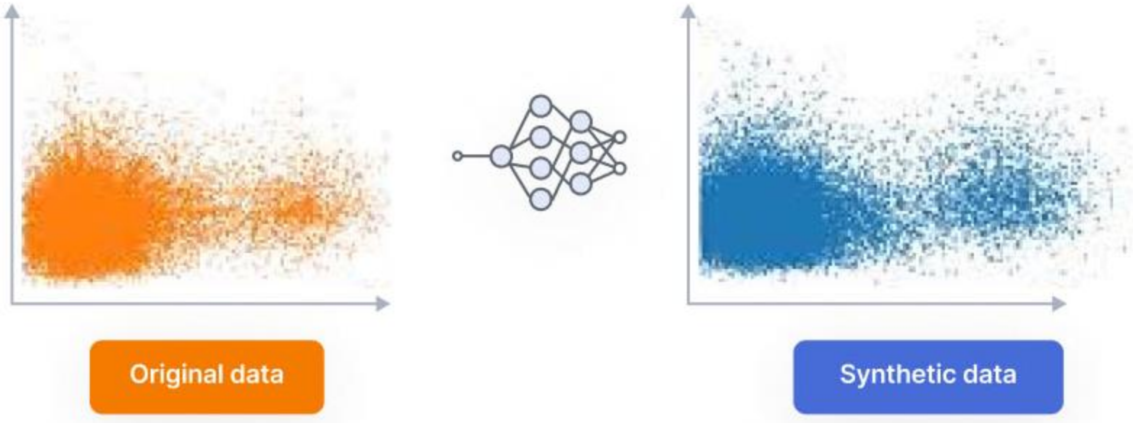
https://www.frontiersin.org/files/Articles/679939/fdata-04-679939-HTML-r1/image_m/fdata-04-679939-g001.jpg

What is synthetic data

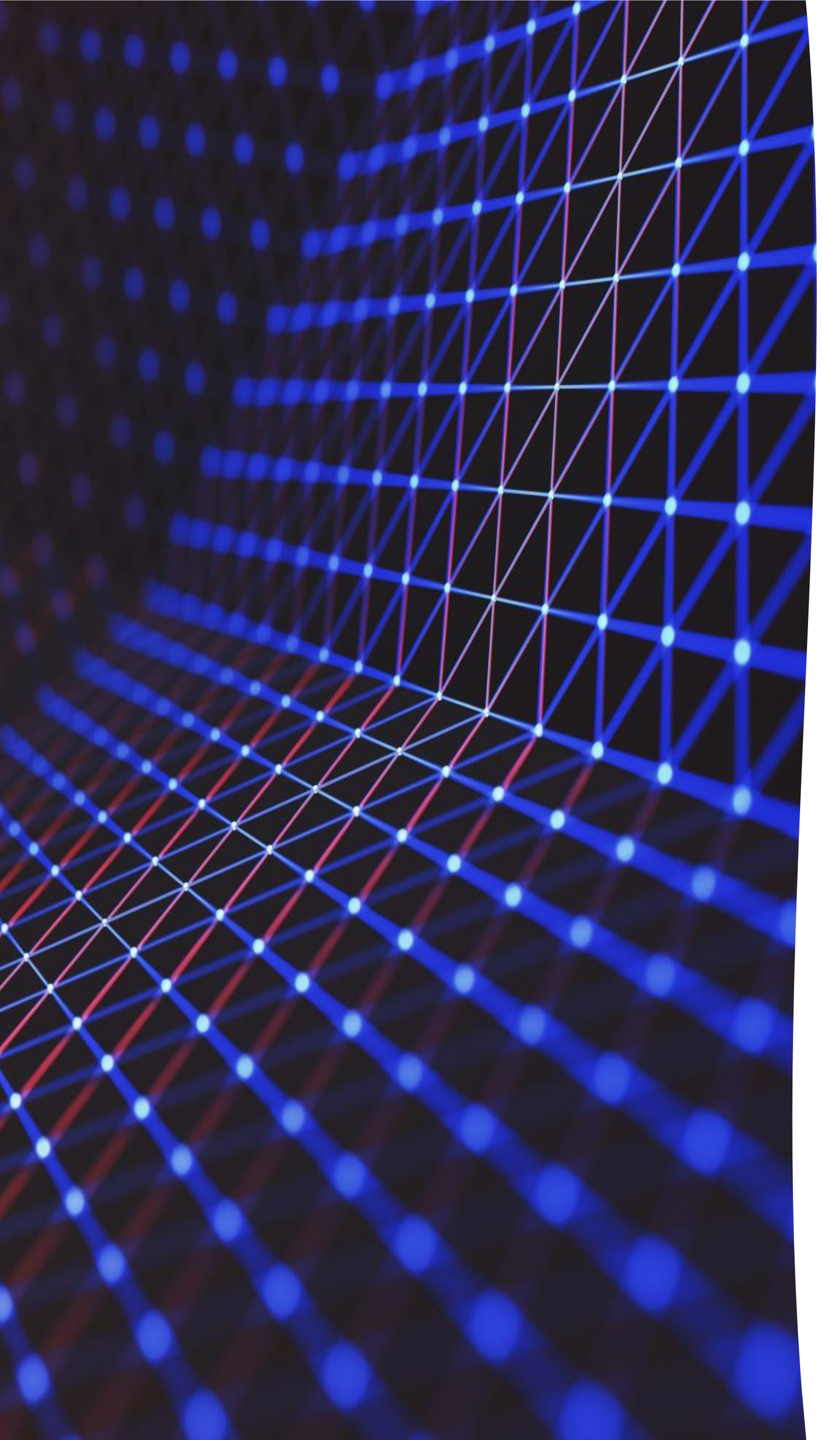
- Privacy protection
- Data availability

Synthetic data generation

New data is created to mirror the statistical properties of original data



https://assets-global.website-files.com/5d7b77b063a9066d83e1209c/63b413dd05d48ab553b6250b_62dfe2bfc453641055c22ecc_HERO%2520-%2520Yellow.jpeg



Synthetic data generation

Machine learning based methods:

Decision trees (CARTs)

Generative adversarial networks (GANs)

Variational autoencoders (VAEs)

Statistical methods:

Copulas

Bayesian networks

Multivariate distributions

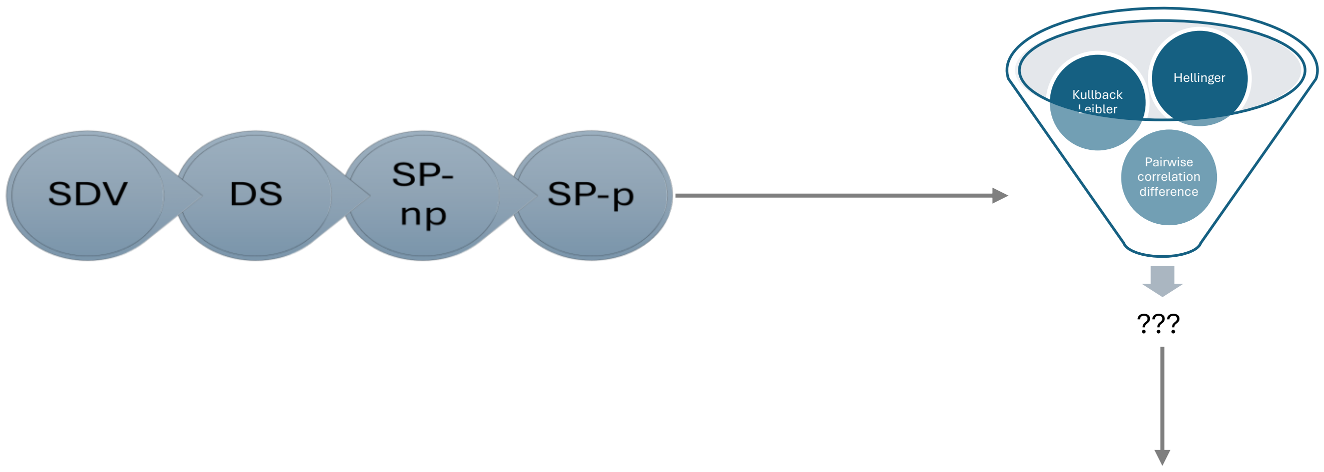
Key areas of investigation

- Evaluating SD utility
- Designing/enhancing synthetic data generation mechanisms
- Privacy preserving techniques
- Bias mitigation
- Regulatory and ethical consideration

Utility

“usefulness of the data for statistical analyses and validity of these analyses”

Problem Defintion



Is it possible to categorize utility metrics and identify a set that addresses different facets of utility?

Is there an SDG that consistently produces better utility across the defined categories?

Can we define a new measure that unifies all these dimensions?

Is this new measure a better indicator of performance?

Part 2

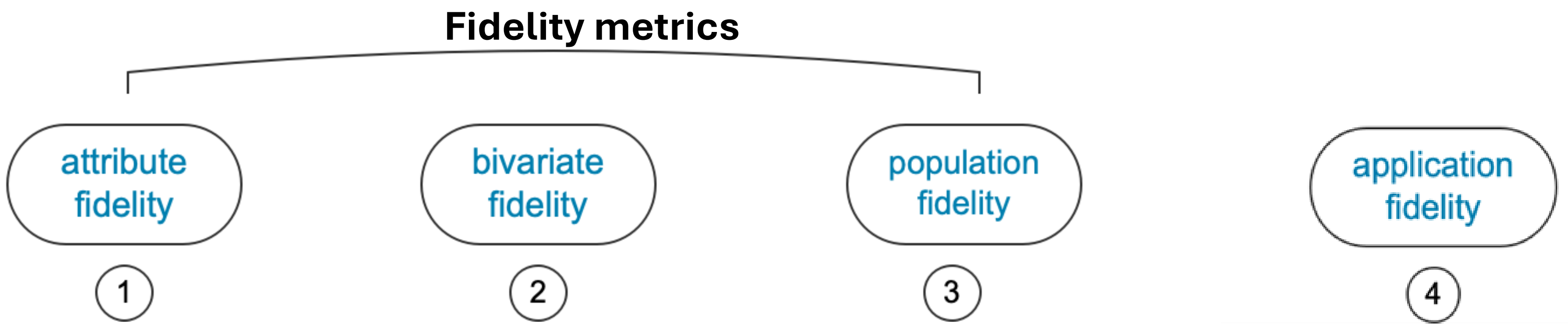
1. Utility measures categorization

Utility measures categorization

- We examined several broad utility metrics used in the generation of synthetic health data.
- Performance across several ML algorithms
- The fidelity metrics used different levels of comparison for assessing the utility :

1. Basic structural similarity between attributes
2. correlation between pairs of attributes, or
3. Similarity on the entire distribution

Utility measures categorization



Hellinger

PCD

Propensity
Log-cluster

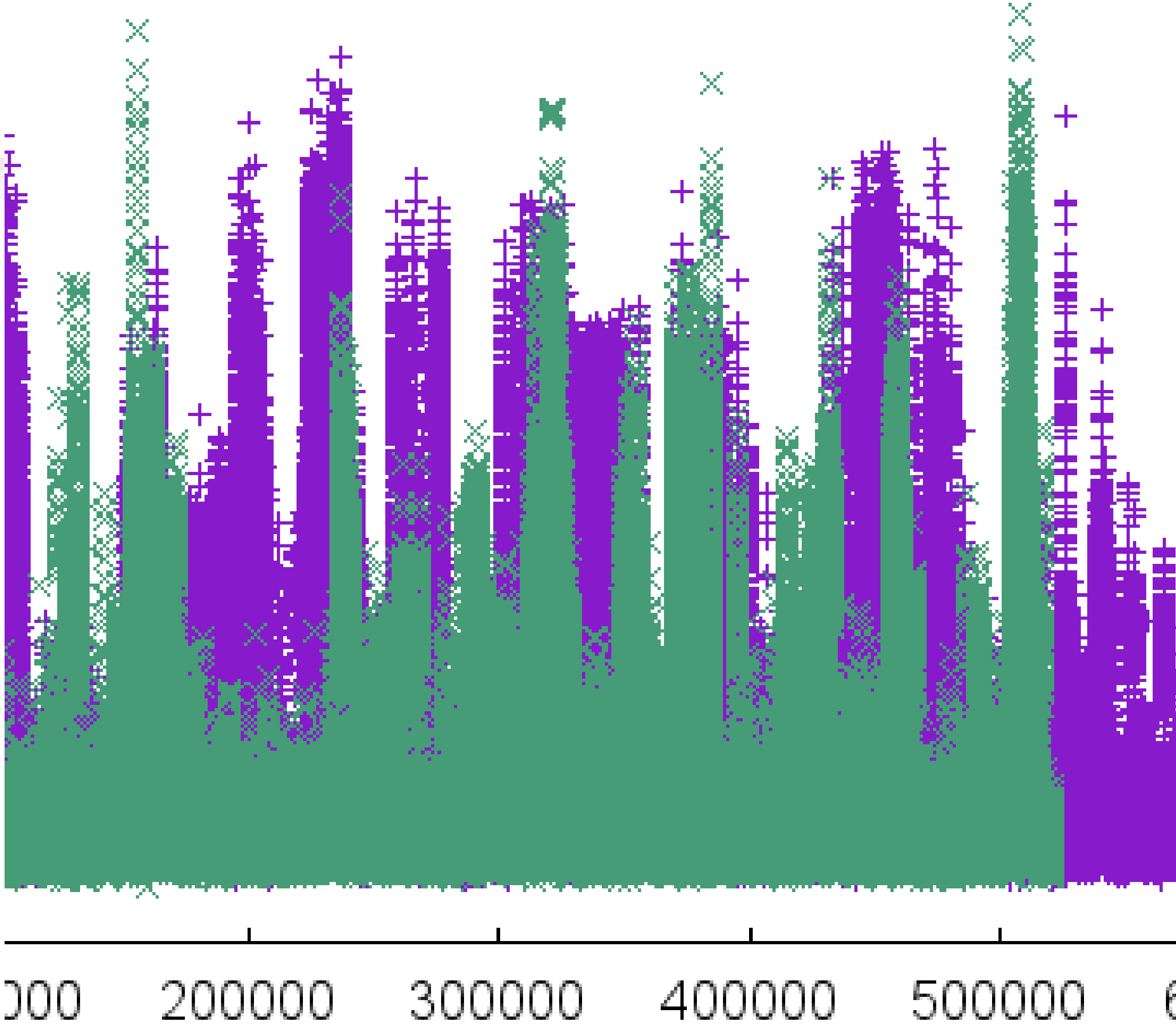
Classification
algorithms

Metrics- Hellinger

- Popular univariate utility measure.
- For each column:

$$H(v_o, v_s) = \frac{1}{\sqrt{2}}\sqrt{\sum_i (\sqrt{p_i} - \sqrt{q_i})^2}$$

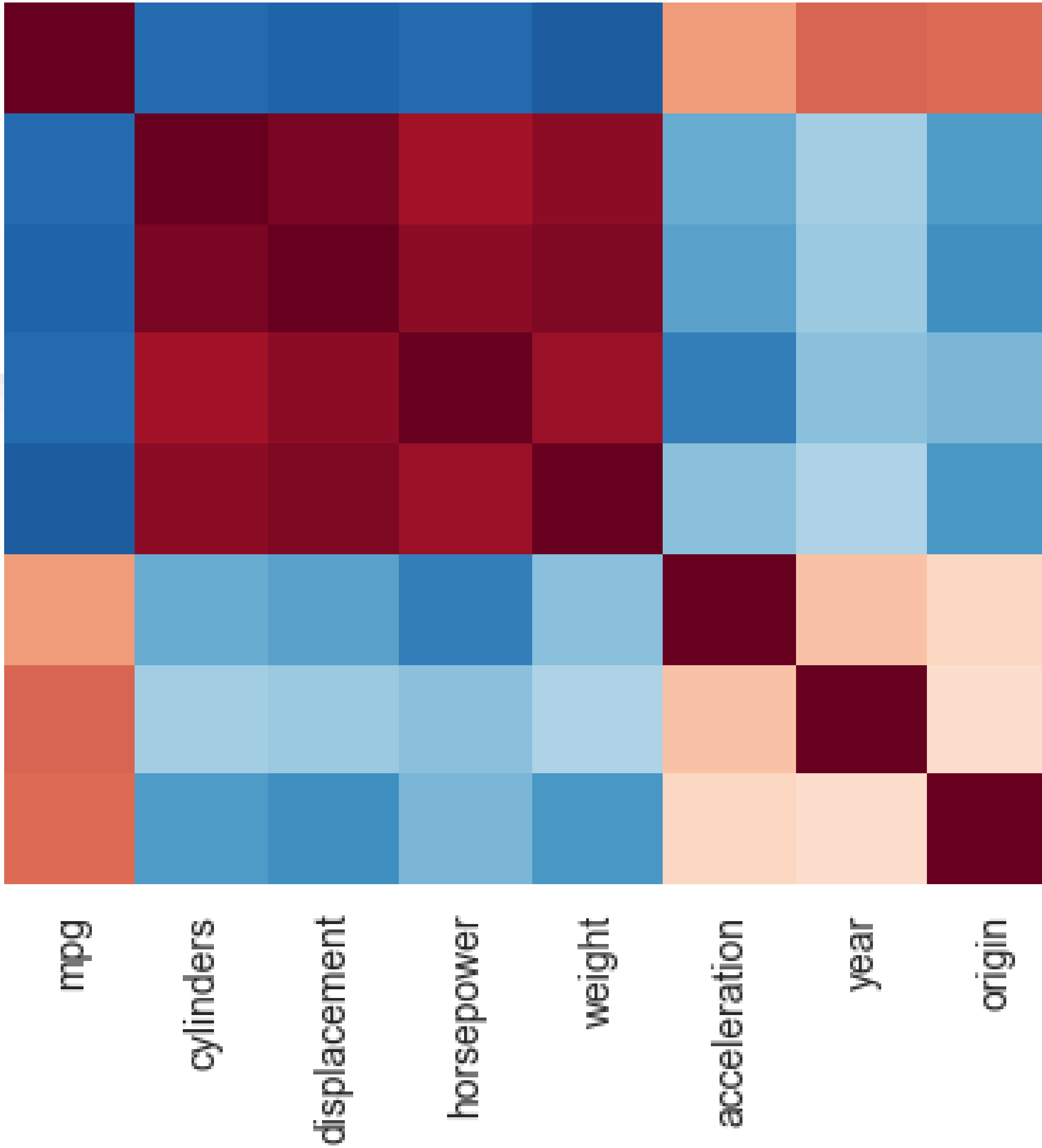
- Then compute the mean Hellinger distance across all variables.
- Shown to be consistent and easy to interpret



Metrics- PCD

- Pairwise correlation difference

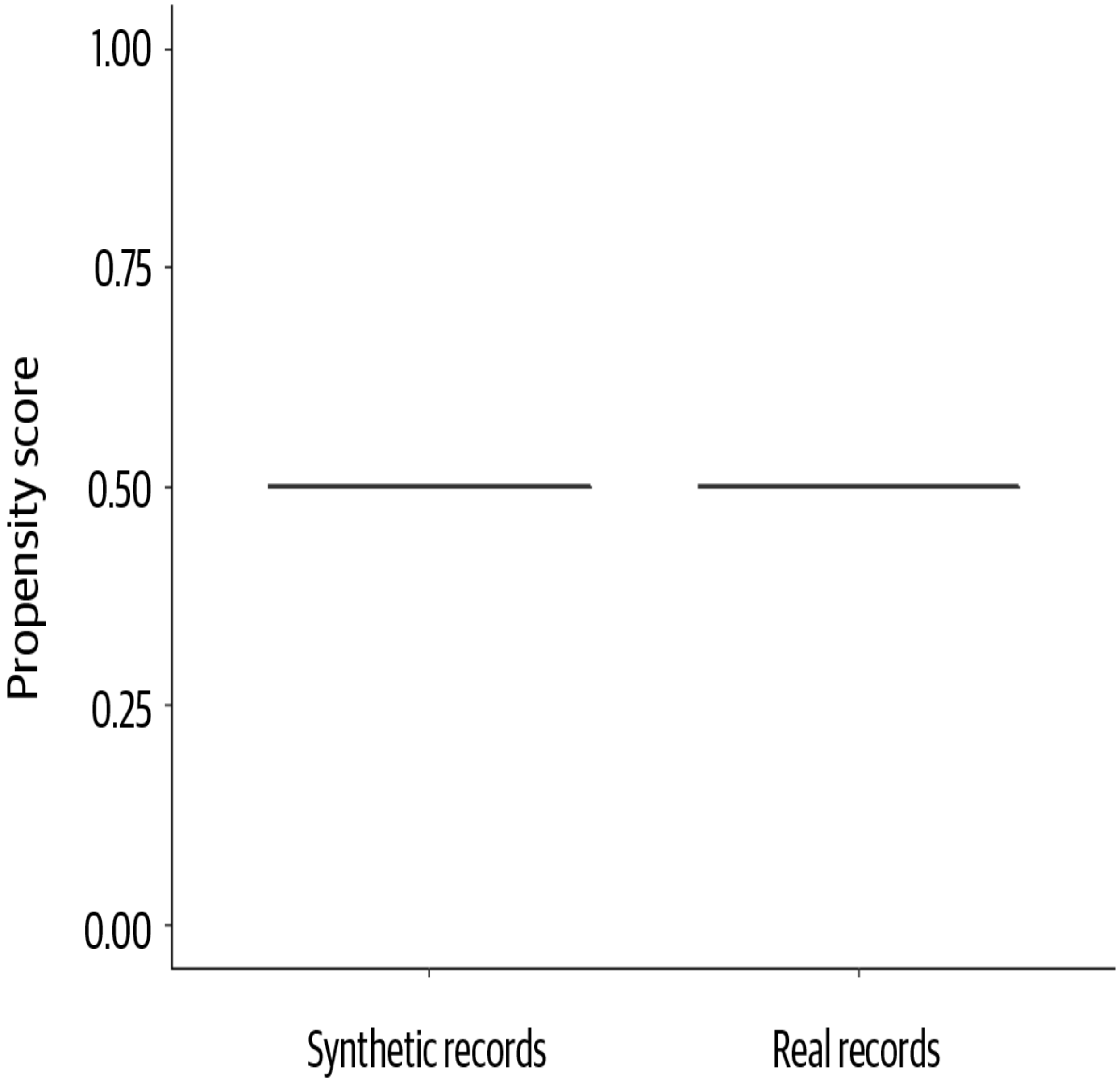
$$PCD(R,S) = ||Corr(R) - Corr(S)||_F$$



Metrics- Propensity

- Most popular broad metric
- The original and synthetic datasets are joined in one group with a binary indicator assigned to each record depending on whether the record is real or synthesized
- A binary classification model is constructed to discriminate between real and synthetic records.

$$pMSE = \frac{1}{N}\sum_{i}(\hat{p}_i - 0.5)^2$$

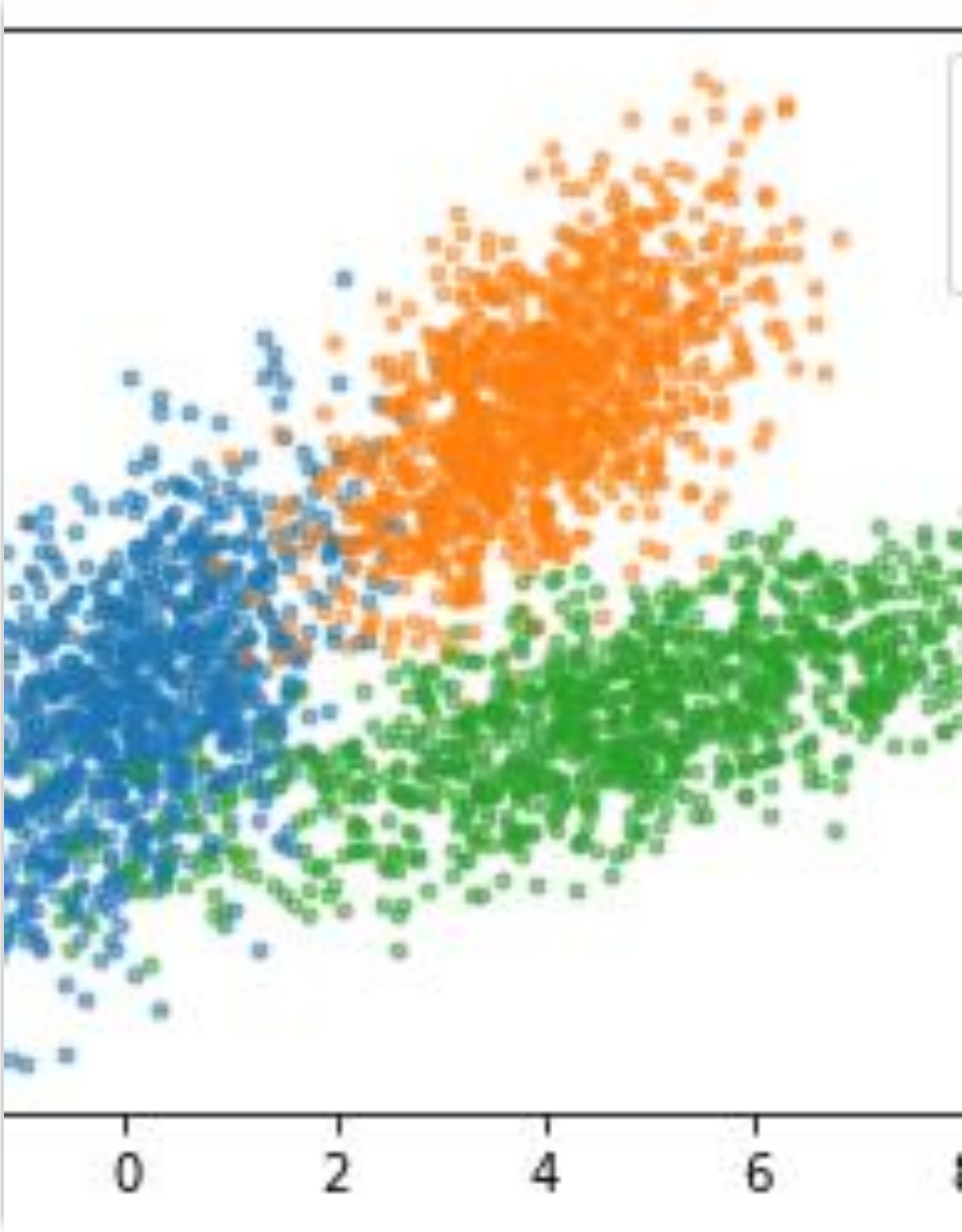


https://www.oreilly.com/library/view/practical-synthetic-data/9781492072737/ch04.html

Metrics- log cluster

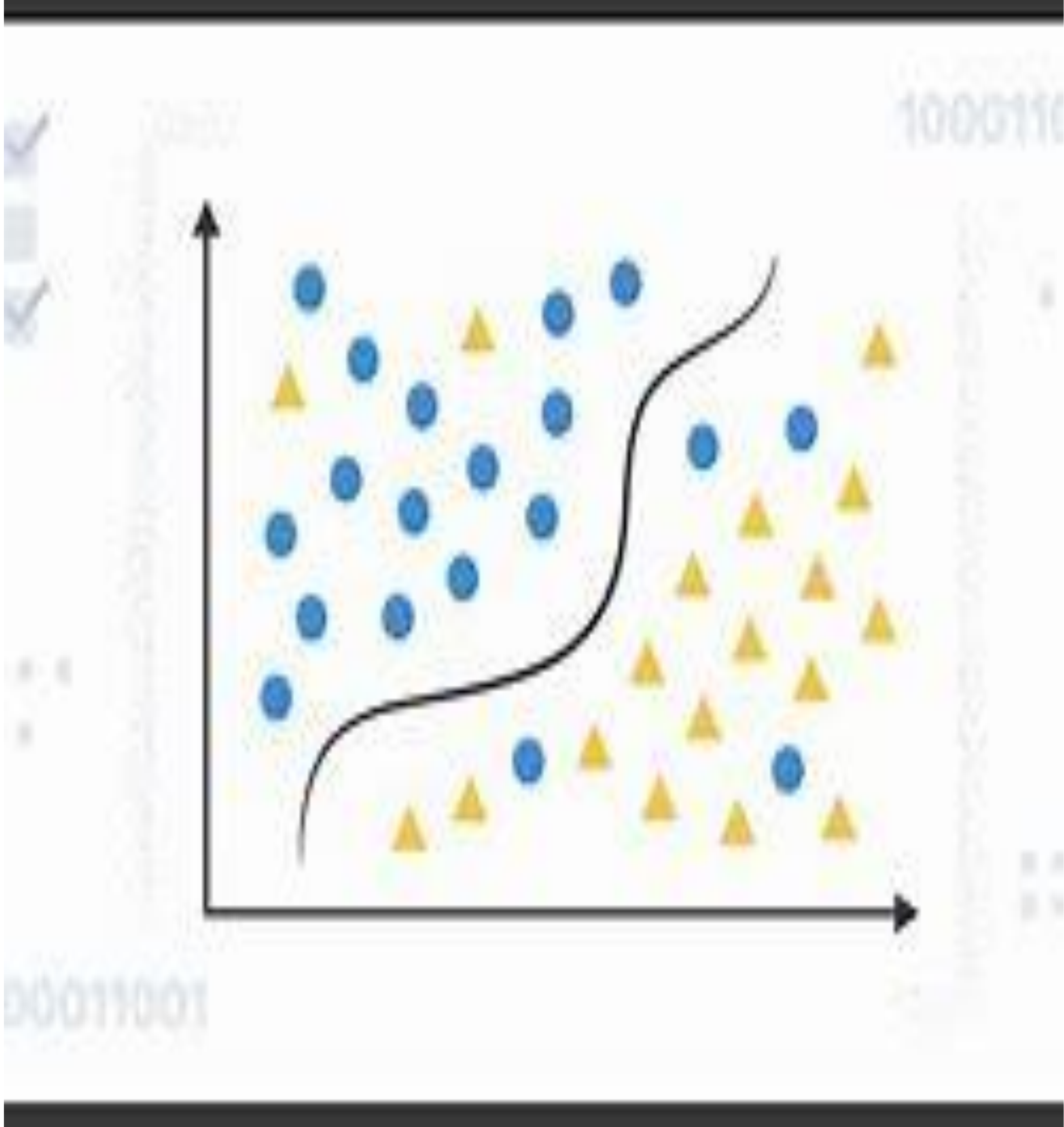
- Popular broad metric.
- Measures the similarity of the underlying dependency structure between the original and synthesized datasets
- The real and synthetic datasets are merged and clustering algorithms are applied on the data to partition the observations into $c$lusters,
- The proportion of real vs synthetic data is assessed within each cluster.

$$U_c(R,S) = \log\left(\frac{1}{G}\sum_{j=1}^{G}\left[c_j - 1/2\right]^2\right)$$



Application Fidelity

- Logistic regression, SVM, RF and DT models are trained on the real and synthetic datasets and tested on the real data.
- Accuracy and F1



Part 2

2. Analysis

Experimental design

We use 4 SDGs for our evaluations:

DataSynthesizer (DS): Bayesian network -based data synthesis technique

Synthetic Data Vault (SDV): Copula-Based data synthesis technique

Synthpop parametric (SP-P): sequential synthesizing of attributes using linear and logistic regression

Synthpop non-parametric (SP-NP): sequential synthesizing of attributes using Classification and regression trees

Key questions

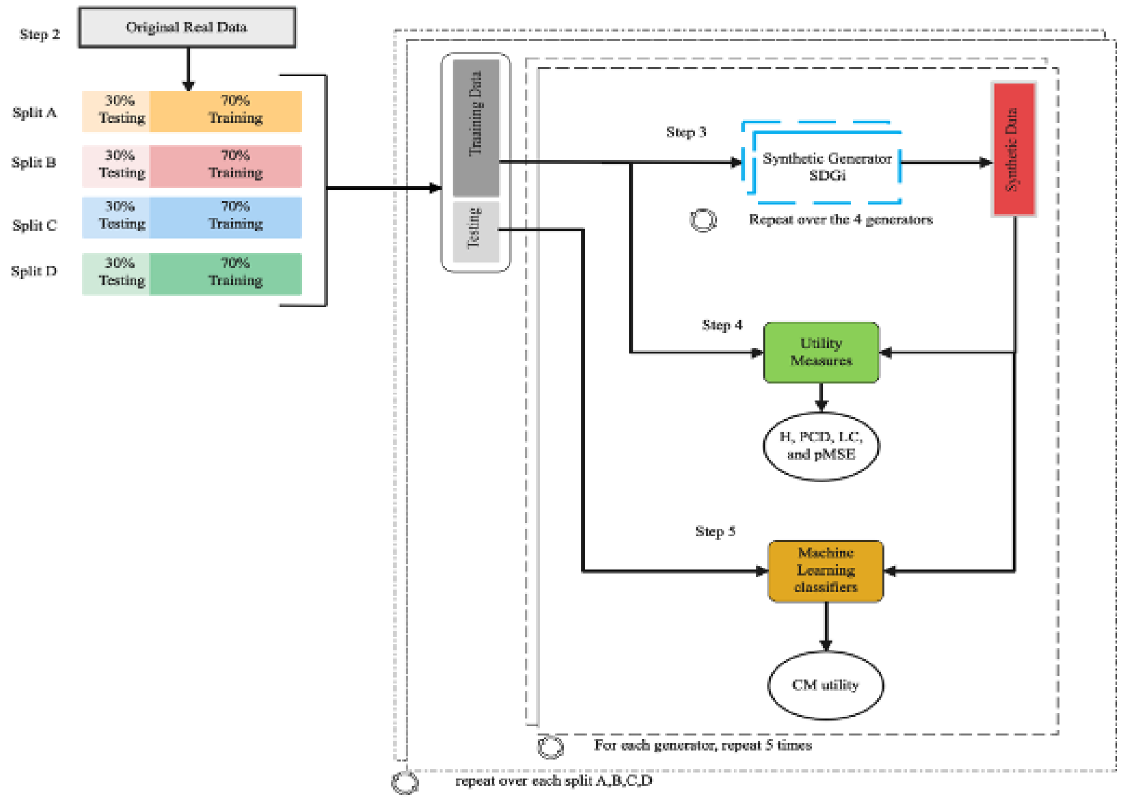
19 datasets from University of California Irvine repository, OpenML platform, Datasphere, Cerner clinical database and Kaggle community platform.

Results were used to address the following:

1. Considering all metrics, Is there a winning SDG?
2. Do metrics agree on a winning generator?
3. Are metrics correlated?

Experimental design

- 4 random splits are created for each dataset, and data synthesis methods are repeated 5 times for each SDG (5*4*4=80 SD per dataset)
- utility metrics are calculated for each of the synthetic datasets generated.
- Logistic regression, SVM, RF and DT models are trained on the real and synthetic datasets and tested on the real data.



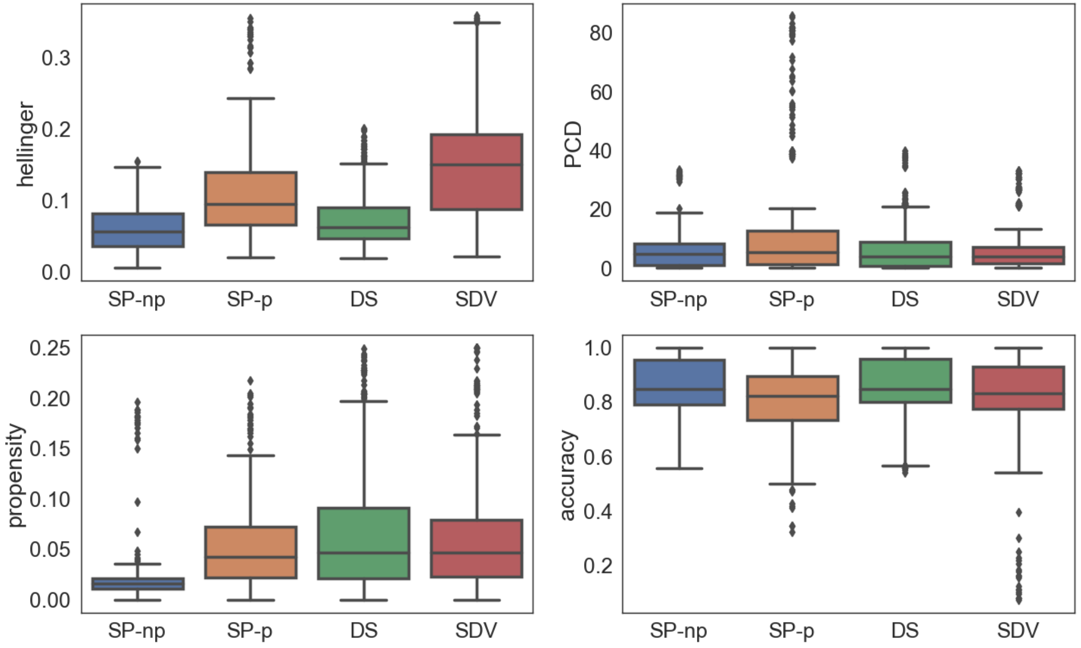
Guidelines[1]

Prior study on evaluating the effect of various synthetic data generation and usage settings on the utility of the generated synthetic data and its derived models.

- there is **no benefit from pre-processing** real data prior to synthesizing it (imputing missing values, encoding categorical values as integers encoding categorical values as integers, and standardizing numeric features)
- **tuning the ML when using synthetic datasets** does not enhance the performance of the generated models (choosing the best hyperparameters of the model and selecting the best set of predictors)

SDG performance

- Considering all metrics, Is there a winning SDG?



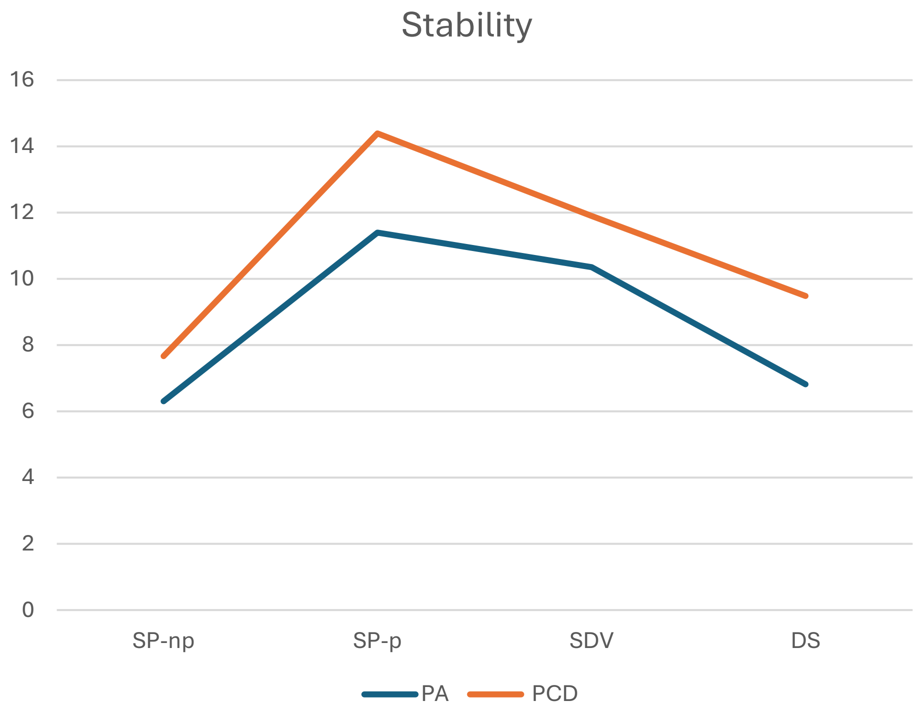
Performance of the different synthetic data generators on each metric and on classification accuracy across all datasets.

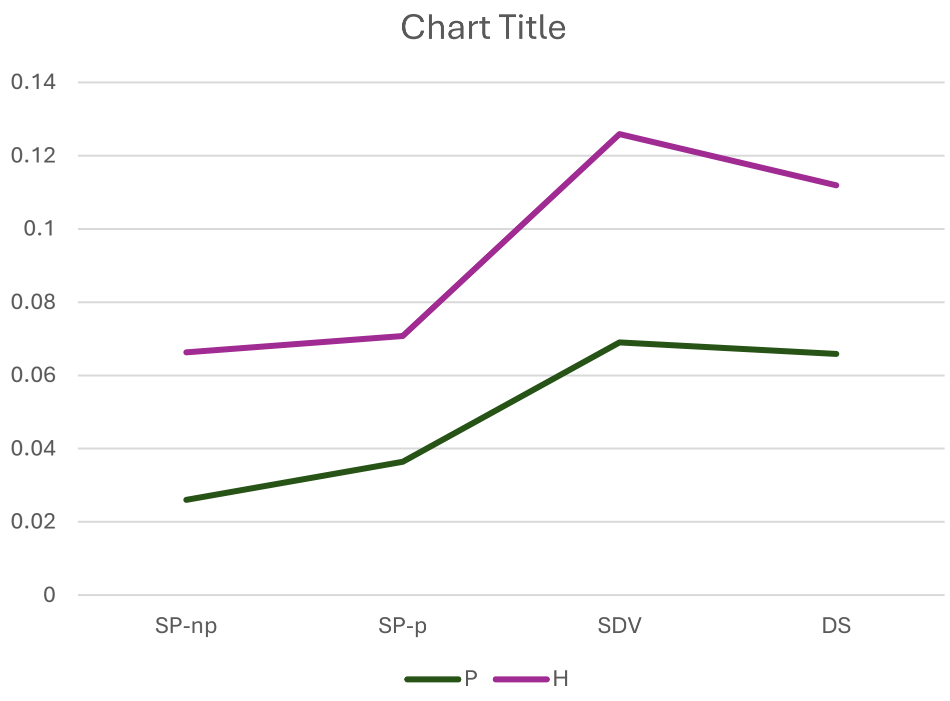
SDG performance

- Average performance

SDG	Hellinger	Average PA loss	PCD	Propensity
SP-np	0.0617296	3.5	6.2960989	0.0233651
SP-p	0.1171702	9.5	14.130144	0.0557168
SDV	0.1539435	9.9	7.4813360	0.0647602
DS	0.0829068	4.5	10.193042	0.0724779
winning	SP-np	SP-np	SP-np	SP-np

SDG performance





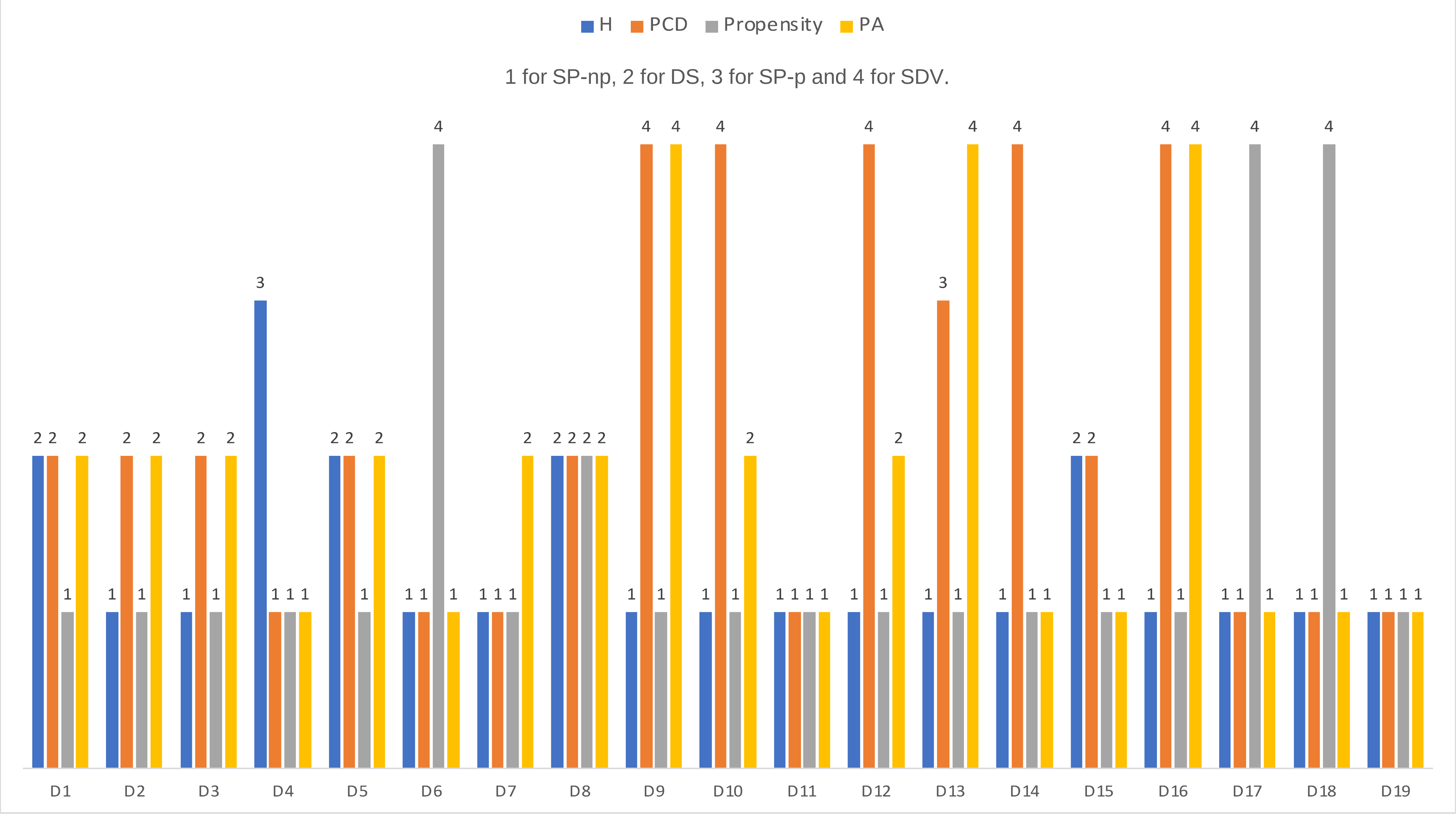
Agreement

Do metrics agree on a winning generator?

	Hellinger	PCD	Propensity	PA
Hellinger		0.368421	0.508772	0.298246
PCD	0.368421		0.017544	0.578947
Propensity	0.508772	0.017544		0.087719
PA	0.298246	0.578947	0.087719	

Kappa score measuring the agreement of different metrics on the winning SDGs

Agreement



H
PCD
Propensity
PA
1 for SP-np, 2 for DS, 3 for SP-p and 4 for SDV.
D1
D2
D3
D4
D5
D6
D7
D8
D9
D10
D11
D12
D13
D14
D15
D16
D17
D18
D19

Correlation

Can one metric be used as an indicator/predictor for all utility dimensions?

	Hellinger	PCD	Propensity	PA
Hellinger	1	0.535184	0.268217	-0.2636
PCD	0.535184	1	0.257282	-0.2684
Propensity	0.268217	0.257282	1	-0.33437
PA	-0.2636	-0.2684	-0.33437	1

Correlation matrix

Part 3

1. A multi-dimensional measure of utility

PCA based utility measure

- Unifying measure
  - We used the 4 fidelity metrics introduced previously to define a new utility measure
  - The measure unifies the 4 measures using principal component analysis (PCA)
  - It is evaluated against propensity

PCA based utility measure

- PCA:
  - For each SD, we consider the tuple $\{H, LC, P, PCD\}$ (16 per dataset)
  - PCA is used to reduce dimensionality to 1
  - 10 datasets are used for training and 9 for testing

$$\{H, LC, P, PCD\} \longrightarrow \boxed{\text{PCA}} \longrightarrow \{u_{PCA}\}$$

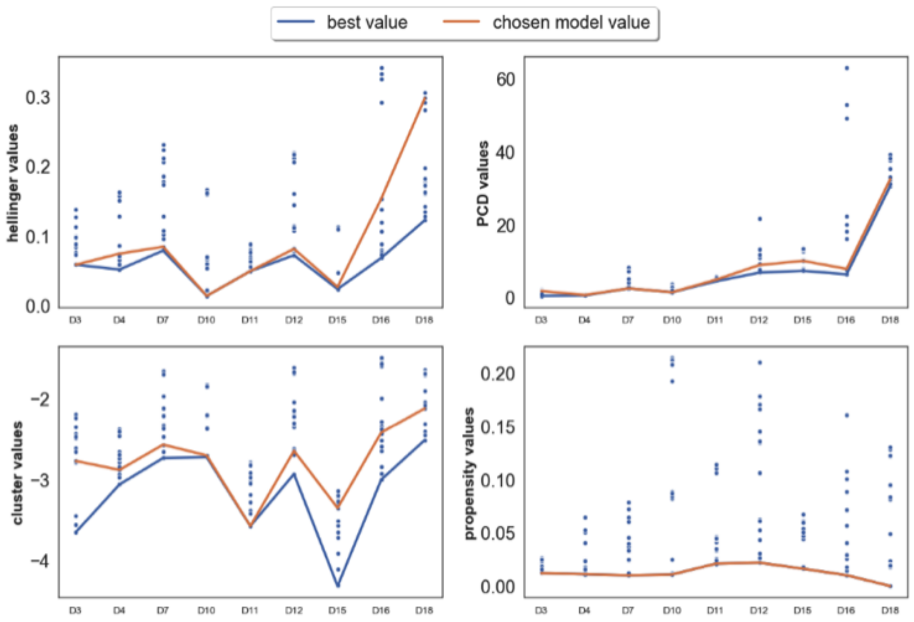
Part 2

2. Experimental evaluation:
    1. *New metric performance in comparison to propensity*
    2. *Correlation with prediction accuracy*

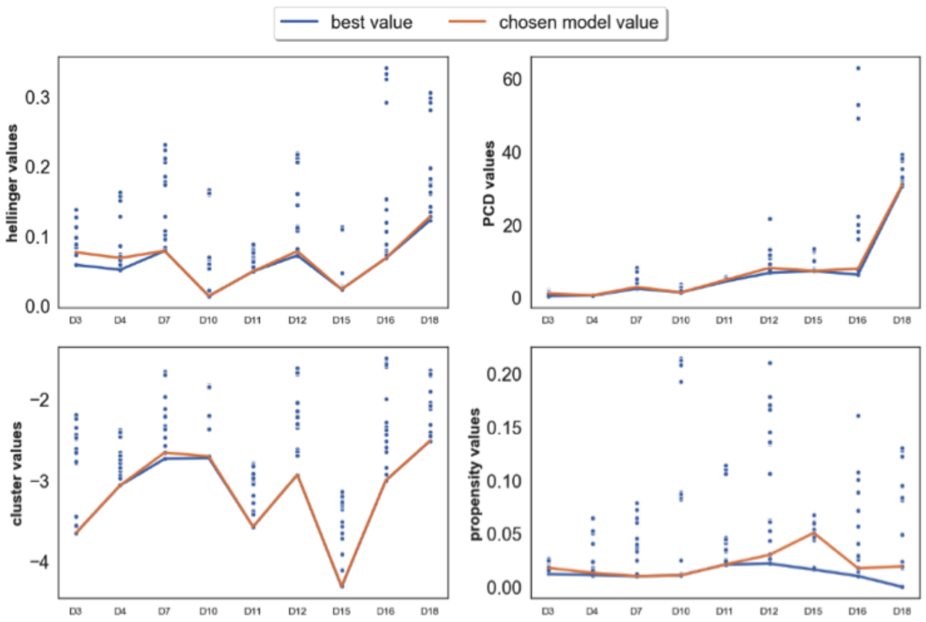
Experimental evaluation

- *Q1: Granular comparison between PCA based metric and propensity score across all utility dimensions*

Propensity



PCA measure



Experimental evaluation

- *Q1: Coarse comparison between PCA based metric and propensity score*

Metrics (metric range)	Average abs diff (p)	Average abs diff (pca)
H (0-1)	0.0335	0.0052
Prop (0-.025)	0.0000	0.0085
LC (-4.7,-1.45)	0.3847	0.0117
PCD (0.06-85.84)	1.1132	0.5587
Average	0.38285	0.146025

Experimental evaluation

- Q2: Correlation with prediction accuracy

	p	pca
DS	0.046	0.537
SDV	-0.308	0.548
SP-np	0.575	0.670
SP-p	0.663	0.708
Overall	0.006	0.525

Limitations

- Further investigations with **more datasets** and machine learning algorithms are needed to validate the above results and to refine the eigenvectors for the PCA based measure.
- Further investigations into the **best broad measures** to include are also needed.
- PCA is most effective when the original variables are highly correlated. We need to explore other **dimensionality reduction** techniques (non-linear)

Readings

1. Fake it till you make it: Guidelines for effective synthetic data generation, FK Dankar, M Ibrahim - Applied Sciences, 2021, Fake It Till You Make It: Guidelines for Effective Synthetic Data Generation (mdpi.com)
2. A Multi-Dimensional Evaluation of Synthetic Data Generators, F. K. Dankar, M. K. Ibrahim and L. Ismail, IEEE Access, vol. 10, A Multi-Dimensional Evaluation of Synthetic Data Generators | IEEE Journals & Magazine | IEEE Xplore.
3. A new PCA-based utility measure for synthetic data evaluation, F. K. Dankar and M. K. Ibrahim, 2022,arXiv, https://arxiv.org/abs/2212.05595

Questions