Data Deidentification Research and Resources from the NIST Collaborative Research Cycle

Gary Howarth, NIST
Christine Task, Knexus Research
with content from
Karan Bhagat, Knexus Research
Dhruv Kapur, University of Michigan

NATIONAL INSTITUTE OF STANDARDS AND TECHNOLOGY
U.S. DEPARTMENT OF COMMERCE

Disclaimer

Portions of this talk are presented by a guest speaker. The contents of this presentation do not necessarily reflect the views or policies of the National Institute of Standards and Technology or the U.S. Government.

Certain commercial entities, equipment, or materials may be identified in this document in order to describe an experimental procedure or concept adequately. Such identification is not intended to imply recommendation or endorsement by the National Institute of Standards and Technology, nor is it intended to imply that the entities, materials, or equipment are necessarily the best available for the purpose.

Please note, unless mentioned in reference to a NIST Publication, all information and data presented is preliminary/in-progress and subject to change.

National Institute of Standards and Technology

Measurements essential to commerce, trade, and innovation

Federal role established in the U.S. Constitution

Why does NIST research privacy?

NIST seeks to facilitate organizations and individuals deriving benefits from data while simultaneously encouraging effective management of risks to individuals’ privacy.

NIST seeks to be the world's leader in creating critical measurement solutions and promoting equitable standards.

NIST is a trusted leader in metrology and provides independent, and transparent technical guidance for the benefit of all.

This talk focuses on a NIST metrology program for data **deidnetification techniques**.

Deidentification includes any processing to microdata that produces microdata in the same schema and is *intended* to be resistant to individual reidentification: SDC, synthetic data, differential privacy.

https://pages.nist.gov/privacy_collaborative_research_cycle/

This past year we've launched a massive community benchmarking and meta-analysis project, collecting metrics, algorithms and data samples from stakeholders, researchers and statistical agencies around the world— and making them all freely available and easy to use. We'll give you a tour, and you can check the QR code to access it all yourselves.

NIST Collaborative Research Cycle: Far more than four algorithms

https://pages.nist.gov/privacy_collaborative_research_cycle/

Algorithms: A Sample of Four Deidentification Approaches

DP Histogram: Add randomized noise to counts

Differential Private Histogram (ε = 10)

PATECTGAN Differential Private GAN (ε = 10)

CART-model Synthesis (non-DP synthetic)

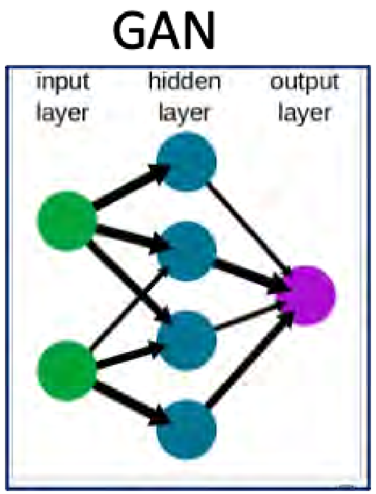
Cell Suppression (k = 6)

Algorithms: A Sample of Four Deidentification Approaches

DP Histogram: Add randomized noise to counts

Differential Private Histogram (ε = 10)

DP GAN: Add randomized noise while training an ML model to reproduce the distribution.



PATECTGAN Differential Private GAN (ε = 10)

CART-model Synthesis (non-DP synthetic)

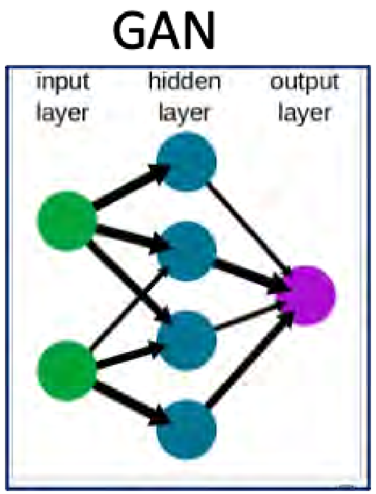
Cell Suppression (k = 6)

Algorithms: A Sample of Four Deidentification Approaches

DP Histogram: Add randomized noise to counts

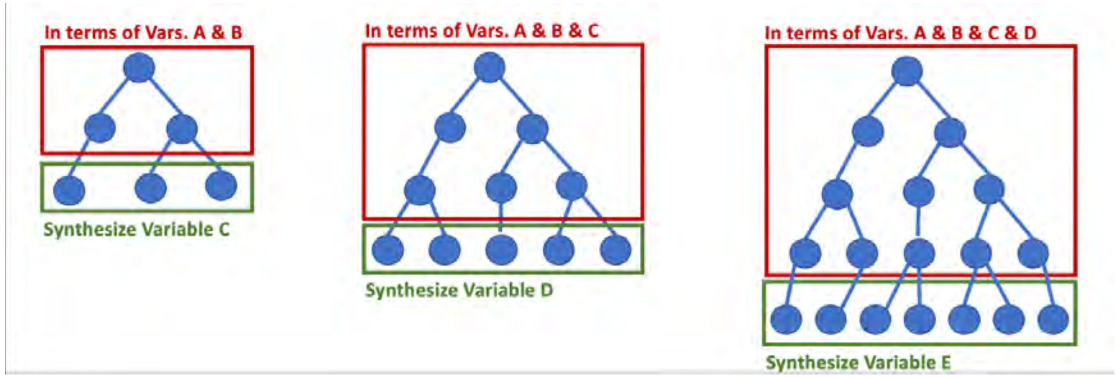
Differential Private Histogram (ε = 10)

DP GAN: Add randomized noise while training an ML model to reproduce the distribution.



PATECTGAN Differential Private GAN (ε = 10)

CART: Use a sequence of decision trees to generate new values for every feature, one at a time.



CART-model Synthesis (non-DP synthetic)

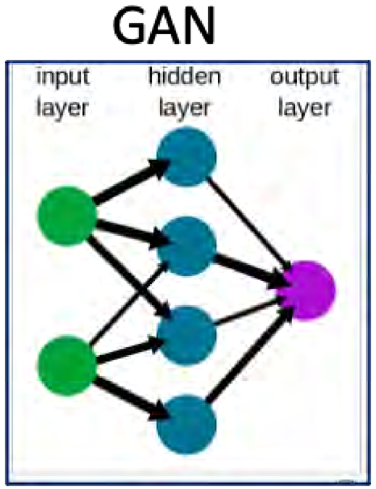
Cell Suppression (k = 6)

Algorithms: A Sample of Four Deidentification Approaches

DP Histogram: Add randomized noise to counts

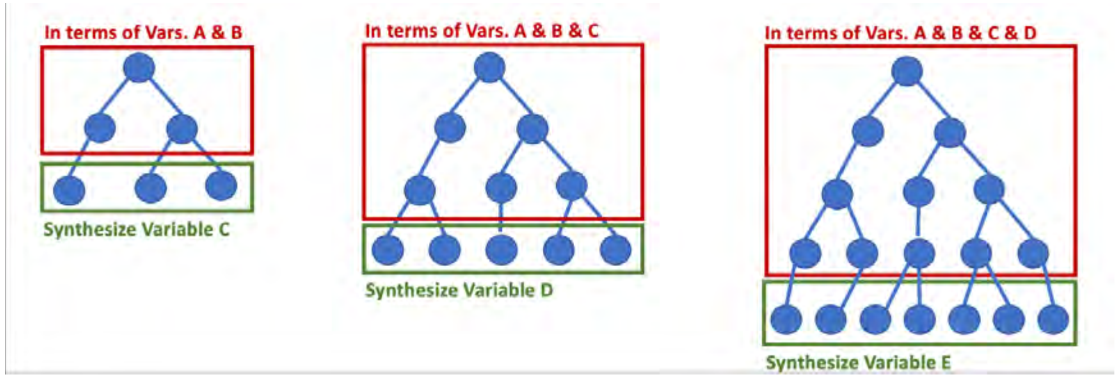
Differential Private Histogram (ε = 10)

DP GAN: Add randomized noise while training an ML model to reproduce the distribution.



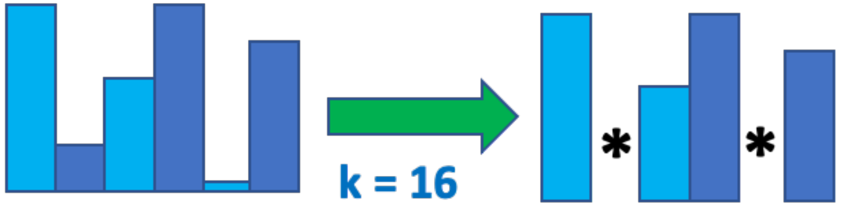
PATECTGAN Differential Private GAN (ε = 10)

CART: Use a sequence of decision trees to generate new values for every feature, one at a time.



CART-model Synthesis (non-DP synthetic)

Cell Suppression: Redact small counts

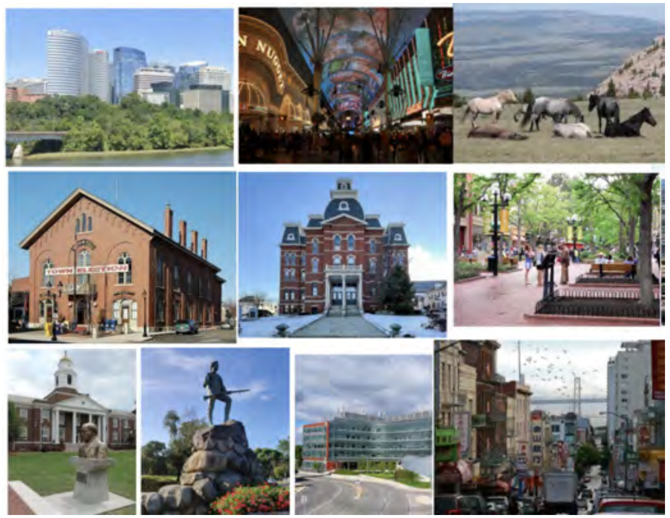


Cell Suppression (k = 6)

Data: Diverse Community Excerpts Benchmark Data

Excerpts of 2019 American Community Survey Data
Tractable schema size for research: 22 Data Features + Weights
Curated to focus on geographies (PUMA) with challenging distributions

Feature Name	Feature Description
PUMA	Public use microdata area code
AGEP	Person's age
SEX	Person's gender
MSP	Marital Status
HISP	Hispanic origin
RAC1P	Person's Race
NOC	Number of own children in household (unweighted)
NPF	Number of persons in family (unweighted)
HOUSING_TYPE	Housing unit or group quarters
OWN_RENT	Housing unit rented or owned
DENSITY	Population density among residents of each PUMA
INDP	Industry codes
INDP_CAT	Industry categories
EDU	Educational attainment
PINCP	Person's total income in dollars
PINCP_DECI	Person's total income in 10-percentile bins
POVPIP	Income-to-poverty ratio (ex: 250 = 2.5 x poverty line)
DVET	Veteran service connected disability rating (percentage)
DREM	Cognitive difficulty
DPHY	Ambulatory (walking) difficulty
DEYE	Vision difficulty
DEAR	Hearing difficulty



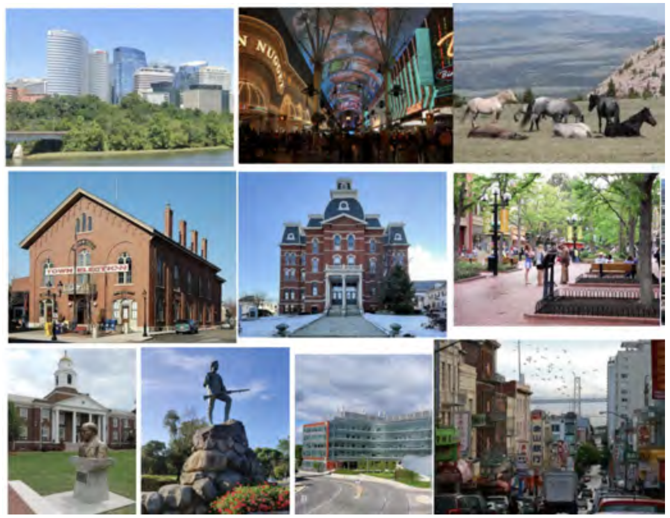
https://github.com/usnistgov/SDNist/tree/main/nist%20diverse%20communities%20data%20excerpts

Data: Diverse Community Excerpts Benchmark Data

Excerpts of 2019 American Community Survey Data
Tractable schema size for research: 22 Data Features + Weights
Curated to focus on geographies (PUMA) with challenging distributions
Recommended Feature Subsets provided for small schema approaches

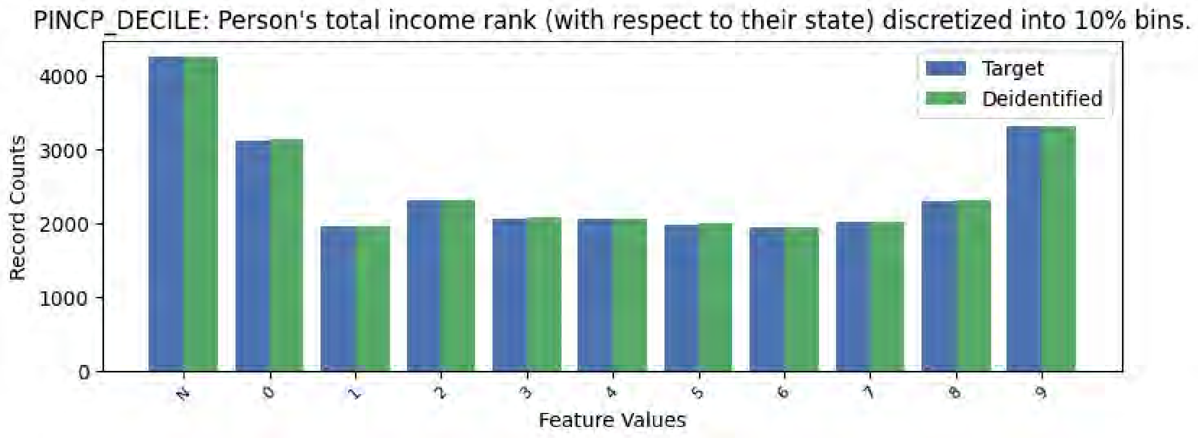
Feature Name	Feature Description
PUMA	Public use microdata area code
AGEP	Person's age
SEX	Person's gender
MSP	Marital Status
HISP	Hispanic origin
RAC1P	Person's Race
NOC	Number of own children in household (unweighted)
NPF	Number of persons in family (unweighted)
HOUSING_TYPE	Housing unit or group quarters
OWN_RENT	Housing unit rented or owned
DENSITY	Population density among residents of each PUMA
INDP	Industry codes
INDP_CAT	Industry categories
EDU	Educational attainment
PINCP	Person's total income in dollars
PINCP_DECILE	Person's total income in 10-percentile bins
POVPIP	Income-to-poverty ratio (ex: 250 = 2.5 x poverty line)
DVET	Veteran service connected disability rating (percentage)
DREM	Cognitive difficulty
DPHY	Ambulatory (walking) difficulty
DEYE	Vision difficulty
DEAR	Hearing difficulty

The next few slides will use the "Demographic-focused" Feature Subset

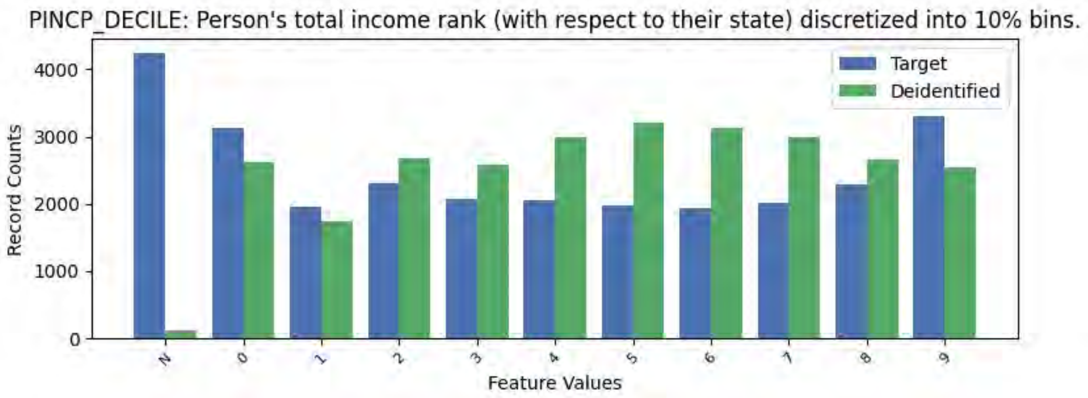


https://github.com/usnistgov/SDNist/tree/main/nist%20diverse%20communities%20data%20excerpts

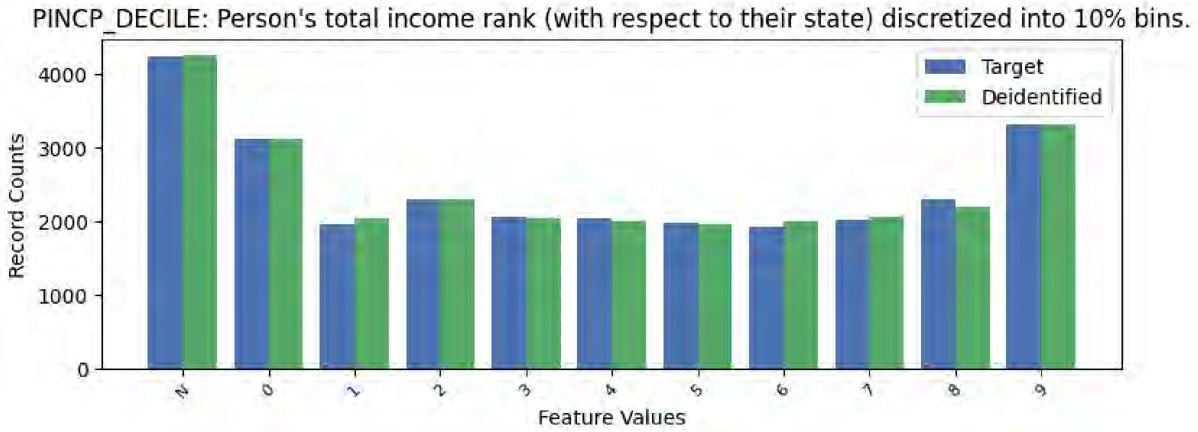
Metrics: Univariate



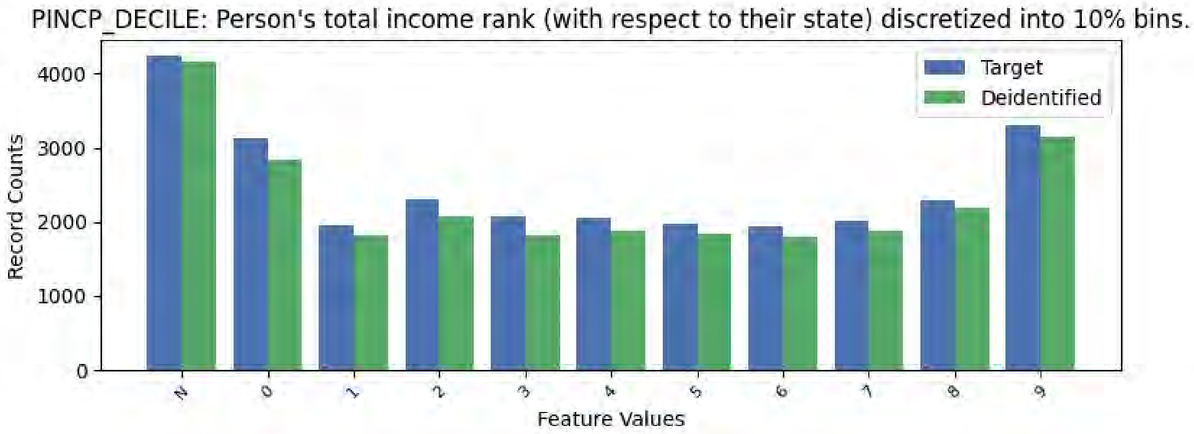
Differential Private Histogram (ε = 10)



PATECTGAN Differential Private GAN (ε = 10)

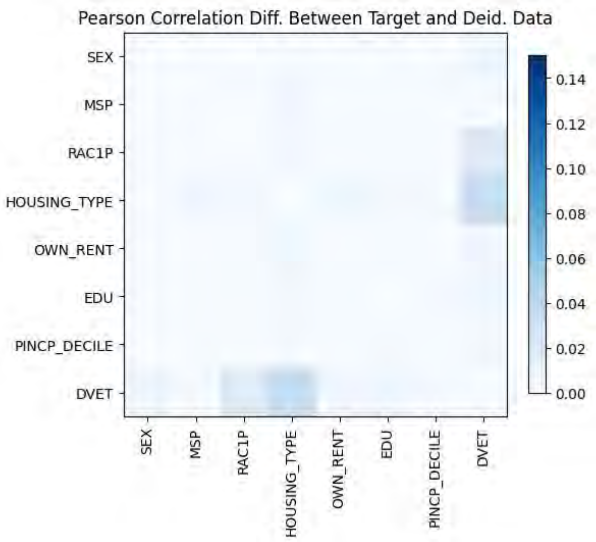


CART-model Synthesis (non-DP synthetic)

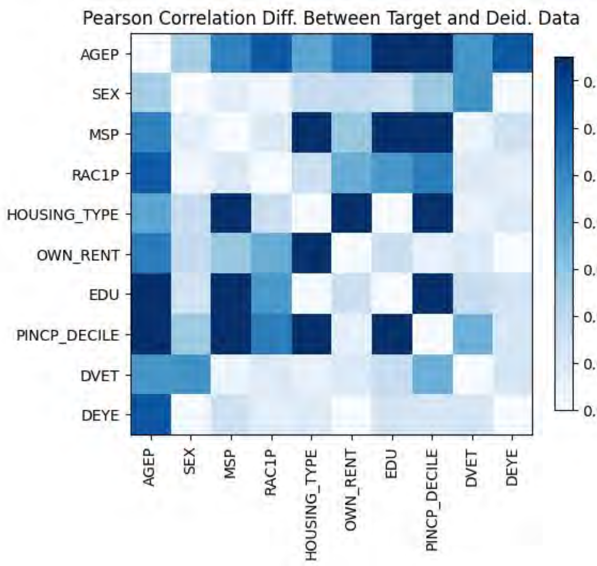


Cell Suppression (k = 6)

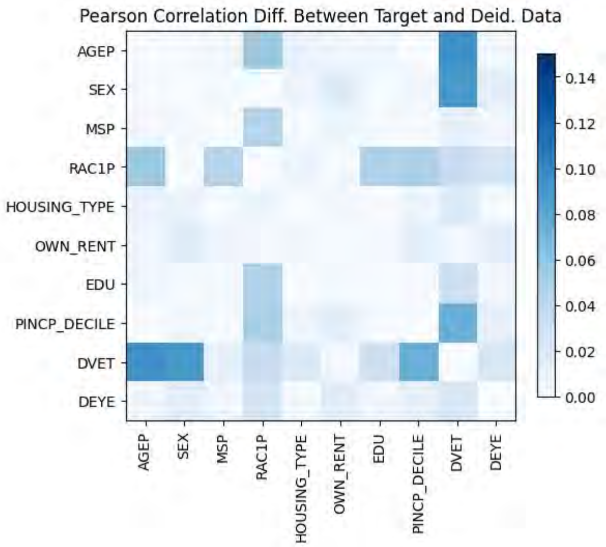
Metrics: Pairwise Correlations



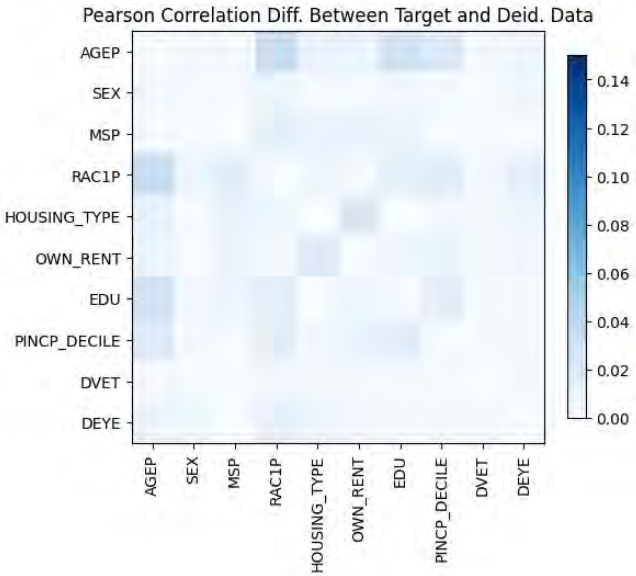
Differential Private Histogram (ε = 10)



PATECTGAN Differential Private GAN (ε = 10)



CART-model Synthesis (non-DP synthetic)



Cell Suppression (κ = 6)

Pairwise Correlations: A key goal of deidentified data is to preserve the feature correlations from the target data, so that analyses performed on the deidentified data provide meaningful insight about the target population. Which correlations are the deidentified data preserving, and which are being altered?

The Pearson Correlation difference was a popular utility metric during the HLG-MOS Synthetic Data Test Drive. Note that darker highlighting indicates pairs of features whose correlations were not well preserved by the deidentified data.

Metrics: K-Marginal Similarity

998
Equivalent to a **90%** uniform random subsample of the input target data.

Differential Private Histogram (ε = 10)

805
Less than a **1%** uniform random subsample of the input target data.

PATECTGAN Differential Private GAN (ε = 10)

984
Equivalent to a **40%** uniform random subsample of the input target data.

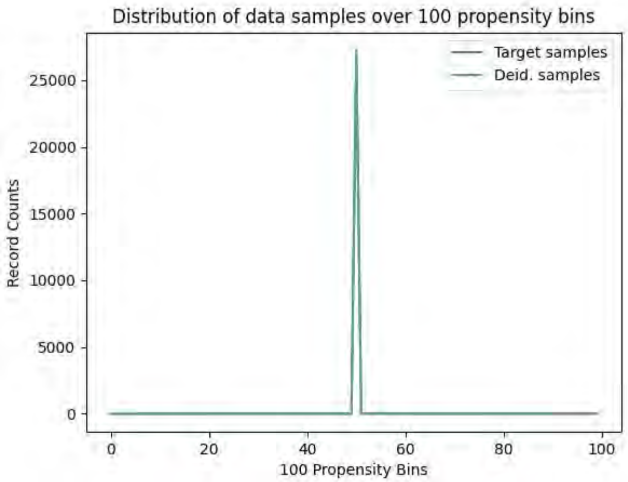
CART-model Synthesis (non-DP synthetic)

977
Equivalent to a **20%** uniform random subsample of the input target data.

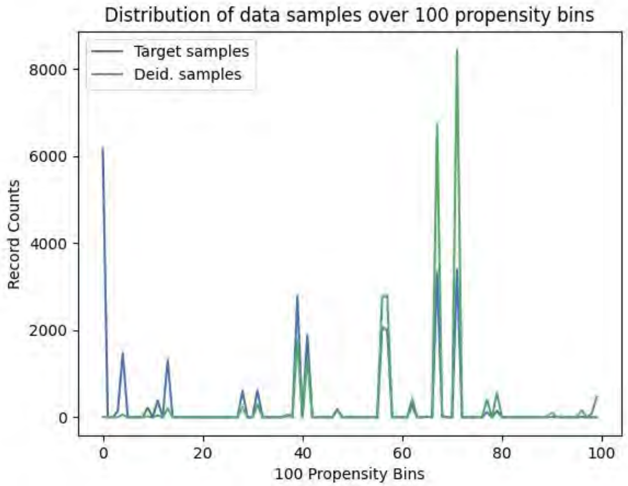
Cell Suppression (k = 6)

K-marginal Similarity: checks how far the shape of the deidentified data distribution has shifted away from the target data distribution, using many 3-dimensional snapshots of the data, averaging the density differences across all snapshots. It was developed as an efficient scoring mechanism for the NIST Temporal Data Challenges, and can be applied to measure the distance between any two data distributions. A score of 0 means zero overlap, while a score of 1000 means the two distributions match identically. More information can be found here.

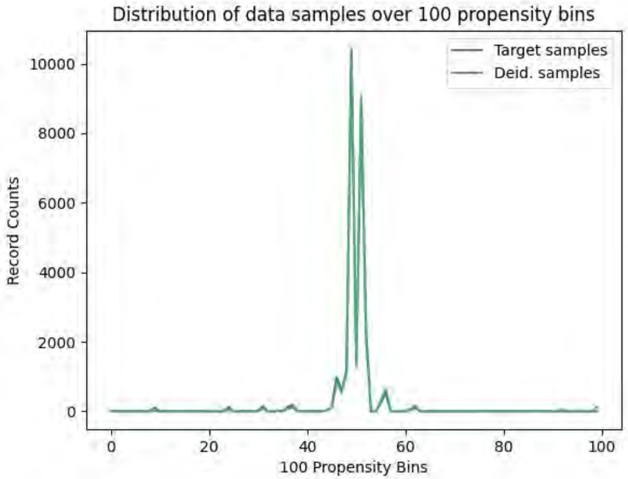
Metrics: Propensity



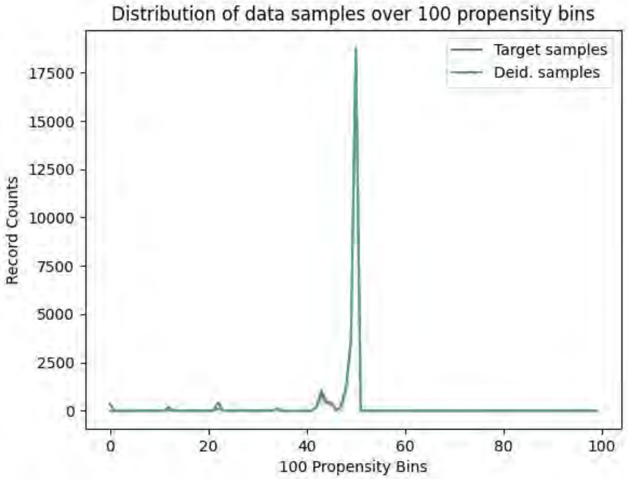
Differential Private Histogram (ε = 10)



PATECTGAN Differential Private GAN (ε = 10)



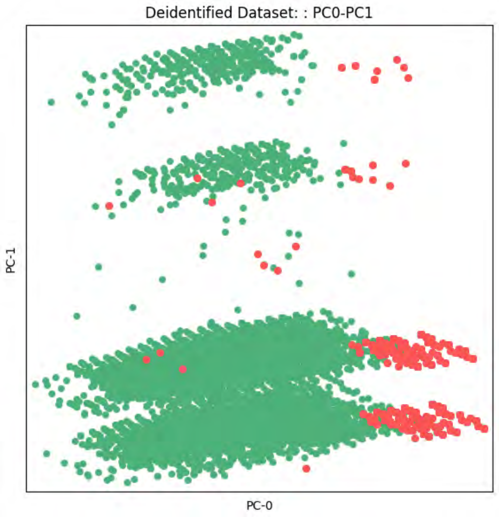
CART-model Synthesis (non-DP synthetic)



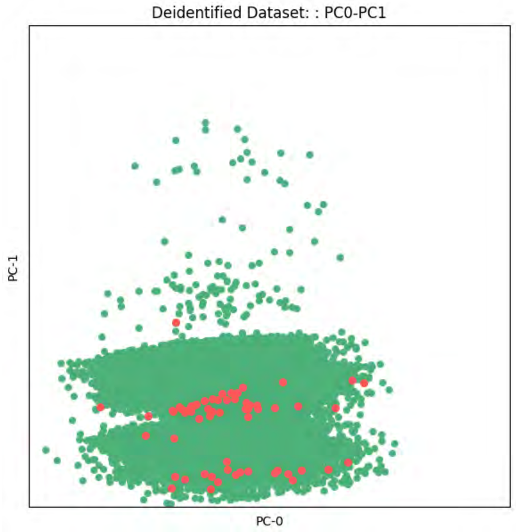
Cell Suppression (k = 6)

Propensity Metrics:
Can a decision tree classifier tell the difference between the target data and the deidentified data? If a classifier is trained to distinguish between the two data sets and it performs poorly on the task, then the deidentified data must not be easy to distinguish from the target data. If the green line matches the blue line, then the deidentified data is high quality. Propensity based metrics have been developed by Joshua Snoke and Gillian Raab and Claire Bowen

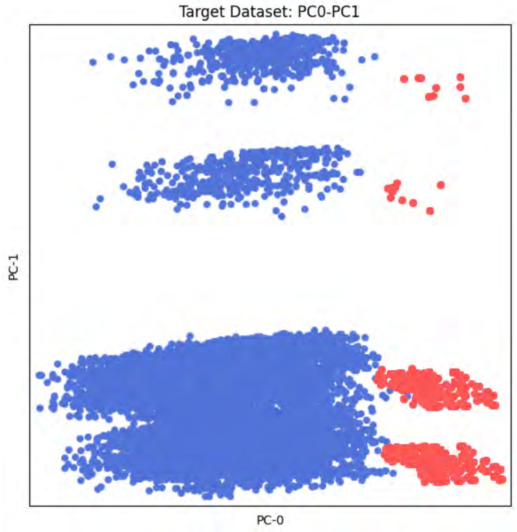
Metrics: Pairwise PCA



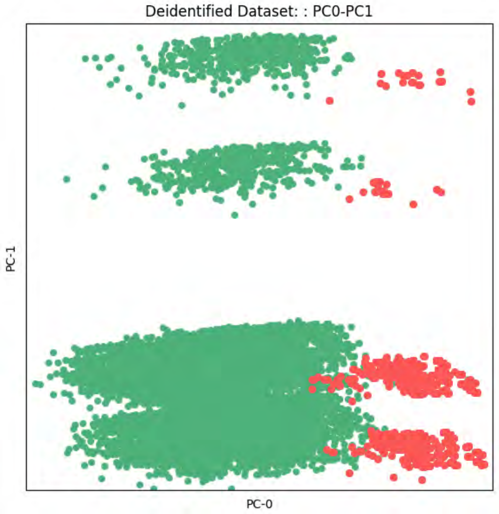
Differential Private Histogram (ε = 10)



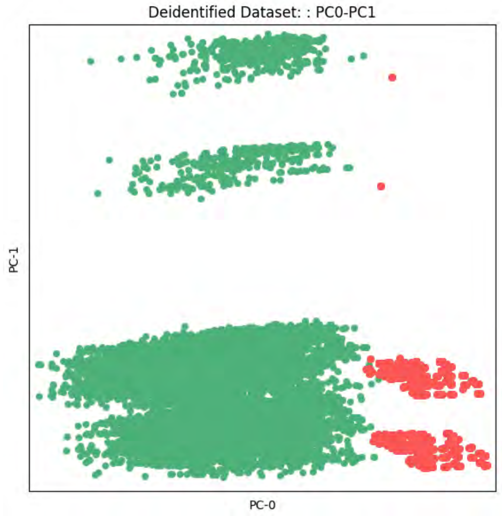
PATECTGAN Differential Private GAN (ε = 10)



PCA Metric visually compares a synthetic data set with the original input data. It plots high dimensional data as a 2D scatterplot using the first two principal component axes; each point represents an individual in the data. Good synthetic data should recreate the shape of the original data with new points (new synthetic individuals). The plot above shows the shape of the original sensitive data; the synthetic data generators are trying to reproduce this distribution. To display more detail, we've used **red points** to highlight records that represent **children** (MSP value = 'N')



CART-model Synthesis (non-DP synthetic)



Cell Suppression (k = 6)

Metrics: Consistency Checks

Inconsistency Group	Number of Records Inconsistent
Age	17
Work	0
Housing	42

Differential Private Histogram (ε = 10)

Inconsistency Group	Number of Records Inconsistent
Age	517
Work	0
Housing	122

PATECTGAN Differential Private GAN (ε = 10)

Inconsistency Group	Number of Records Inconsistent
Age	59
Work	0
Housing	0

CART-model Synthesis (non-DP synthetic)

Inconsistency Group	Number of Records Inconsistent
Age	0
Work	0
Housing	0

Cell Suppression (k = 6)

Age Inconsistencies: These inconsistencies deal with the AGE feature; records with age-based inconsistencies might have children who are married, or infants with high school diplomas

Work Inconsistencies: These inconsistencies deal with the work and finance features — such as high incomes while being in poverty.

Housing Inconsistencies: Records with household inconsistencies might have more children in the house than the total household size, or be residents of group quarters (such as prison inmates) who are listed as owning their residences.

Metrics: Unique Exact Match Rate

Percent of unique Target Data records exactly matched in Deid. Data: **100%**

Differential Private Histogram (ε = 10)

Percent of unique Target Data records exactly matched in Deid. Data: **7.1%**

PATECTGAN Differential Private GAN (ε = 10)

Percent of unique Target Data records exactly matched in Deid. Data: **20.32%**

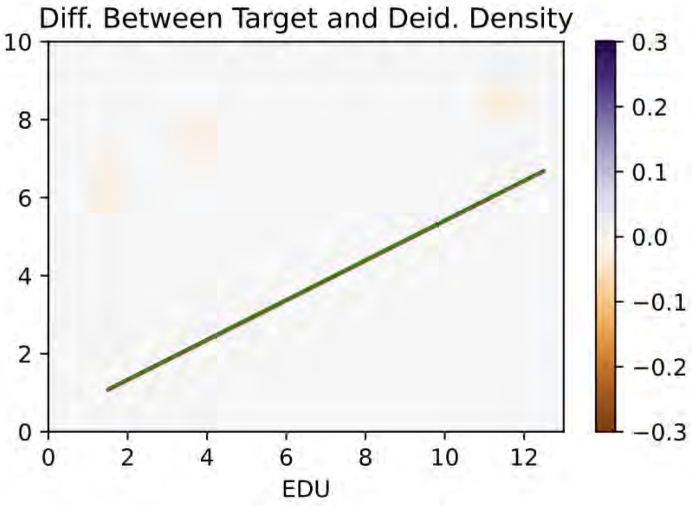
CART-model Synthesis (non-DP synthetic)

Percent of unique Target Data records exactly matched in Deid. Data: **48.5%**

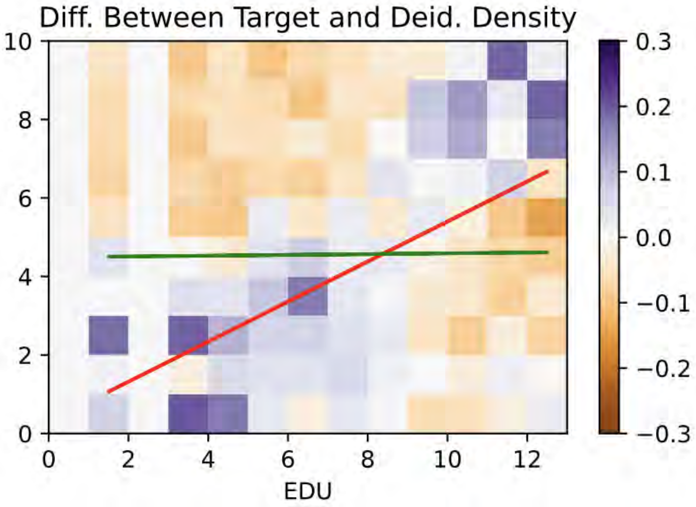
Cell Suppression (k = 6)

Unique Exact Match Rate: This is a count of unique records in the target data that were exactly reproduced in the deidentified data. Because these records were unique outliers in the target data, and they still appear unchanged in the deidentified data, they are potentially vulnerable to reidentification.

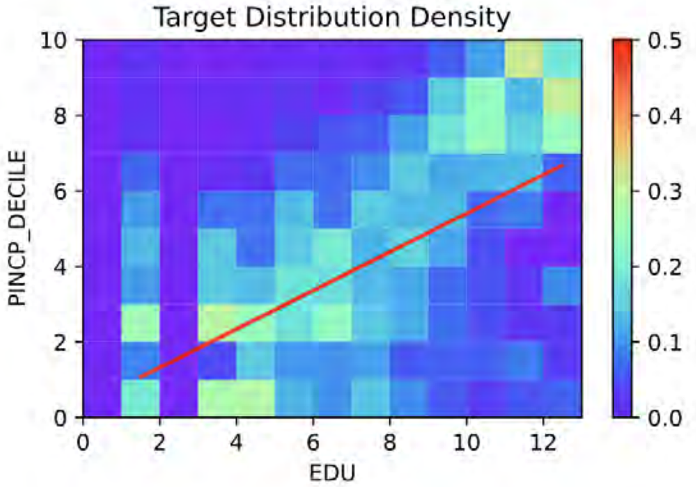
Application Specific Metrics: Relationship between Education and Income



Differential Private Histogram (ε = 10)

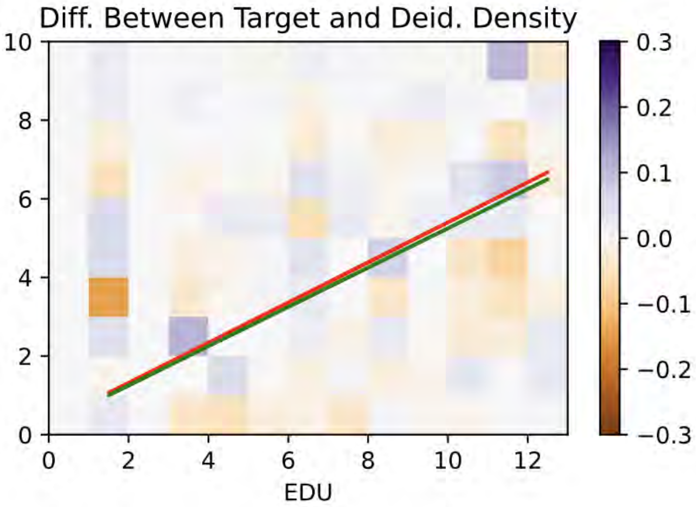


PATECTGAN Differential Private GAN (ε = 10)

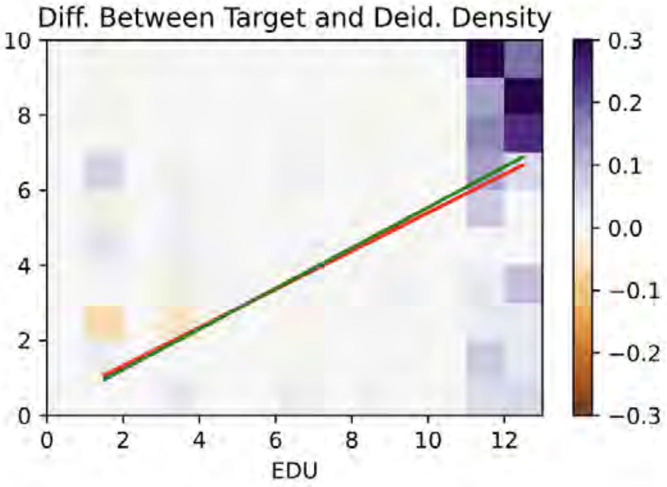


This data-specific metric looks at **linear regression** on adults (AGEP > 15) across two features: Income Decile and Educational Attainment. Higher values of EDU should lead to higher values of PINCP_DECILE, however the relationship is different for different demographic subgroups.

Here we show how well the deidentified data preserves the distribution of black women, using a deviation heatmap: Purple indicates the deidentified data contains too few individuals in that area, brown indicates too many. The original target distribution is shown above in blue.



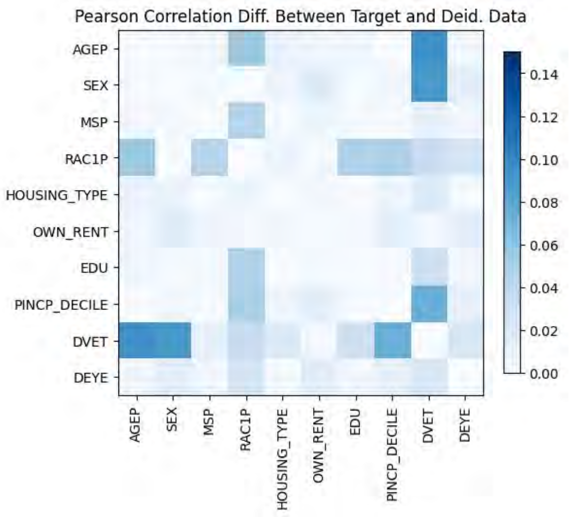
CART-model Synthesis (non-DP synthetic)



Cell Suppression (k = 6)

What Happens on the Full Feature Set?: Pairwise Correlations

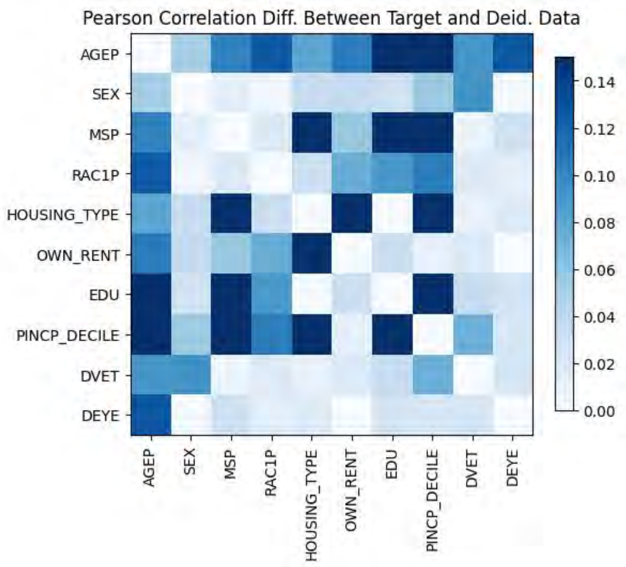
10 Feature Subset



21 Features

CART-model Synthesis (non-DP synthetic)

10 Feature Subset



21 Features

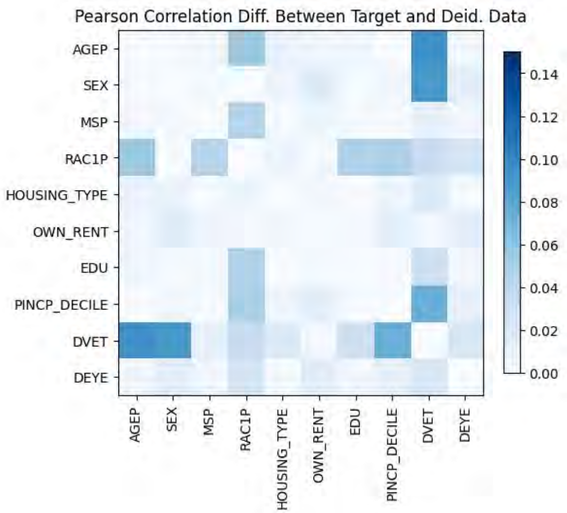
PATECTGAN Differential Private GAN (ε = 10)

Pairwise Correlations: A key goal of deidentified data is to preserve the feature correlations from the target data, so that analyses performed on the deidentified data provide meaningful insight about the target population. Which correlations are the deidentified data preserving, and which are being altered?

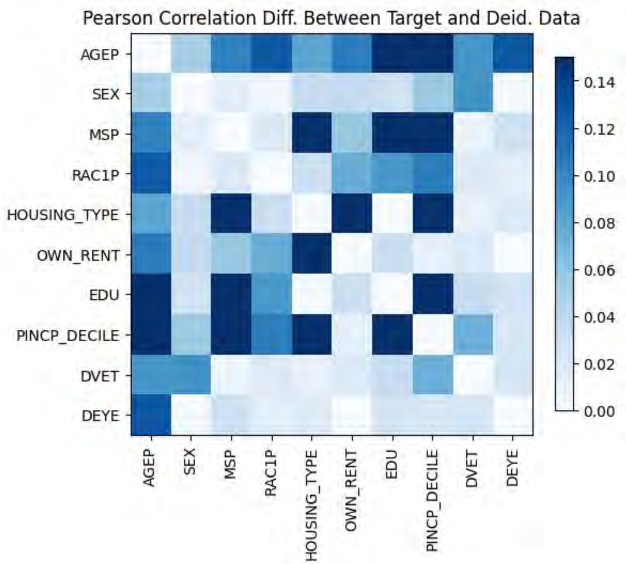
The Pearson Correlation difference was a popular utility metric during the HLG-MOS Synthetic Data Test Drive. Note that darker highlighting indicates pairs of features whose correlations were not well preserved by the deidentified data.

What Happens on the Full Feature Set?: Pairwise Correlations

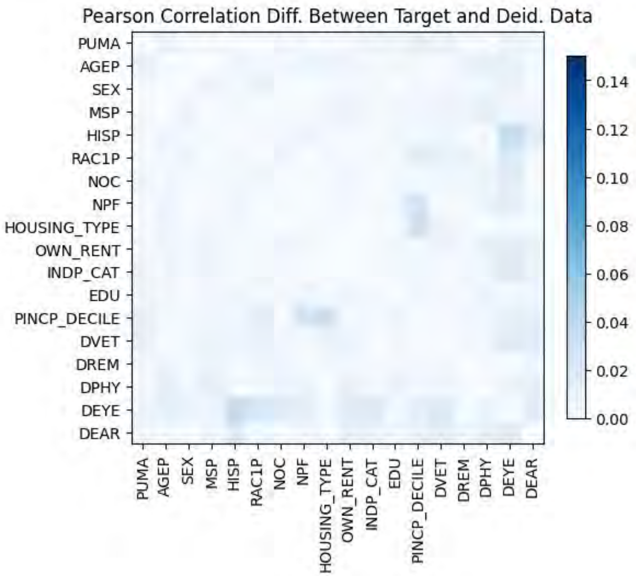
10 Feature Subset



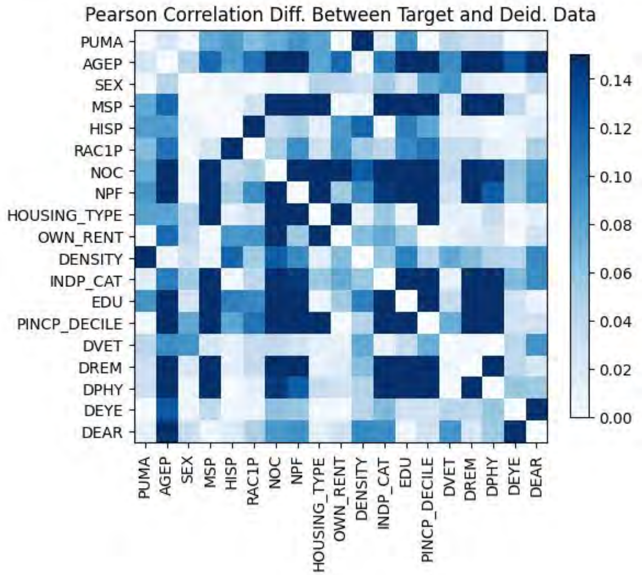
10 Feature Subset



21 Features



21 Features



CART-model Synthesis (non-DP synthetic)

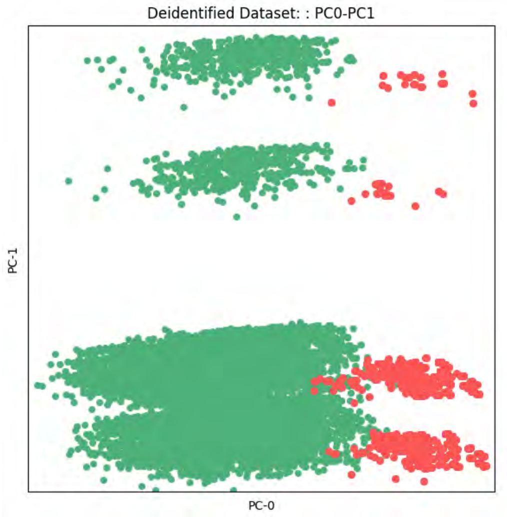
PATECTGAN Differential Private GAN (ε = 10)

Pairwise Correlations: A key goal of deidentified data is to preserve the feature correlations from the target data, so that analyses performed on the deidentified data provide meaningful insight about the target population. Which correlations are the deidentified data preserving, and which are being altered?

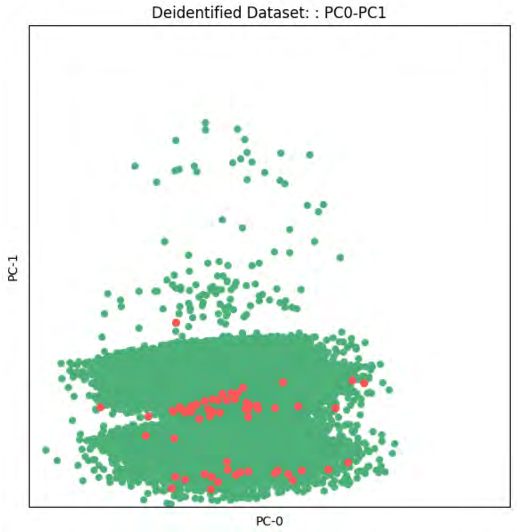
The Pearson Correlation difference was a popular utility metric during the HLG-MOS Synthetic Data Test Drive. Note that darker highlighting indicates pairs of features whose correlations were not well preserved by the deidentified data.

What Happens on the Full 24 Feature Set?: Pairwise PCA

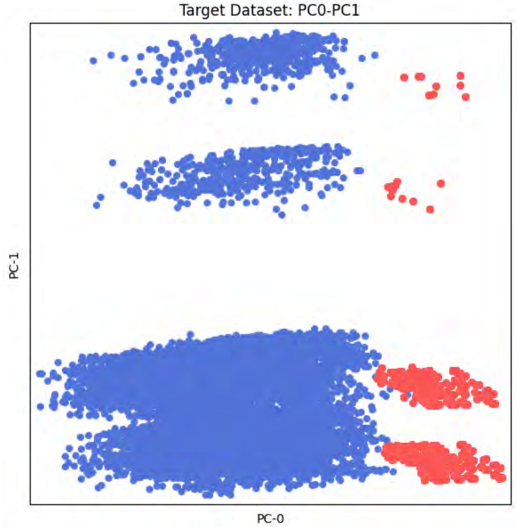
10 Feature Subset



10 Feature Subset



10 Feature Subset Target



21 Features

21 Features

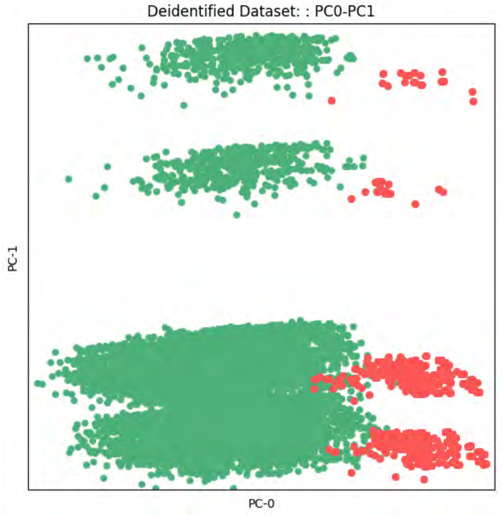
24 Feature Target

CART-model Synthesis (non-DP synthetic)

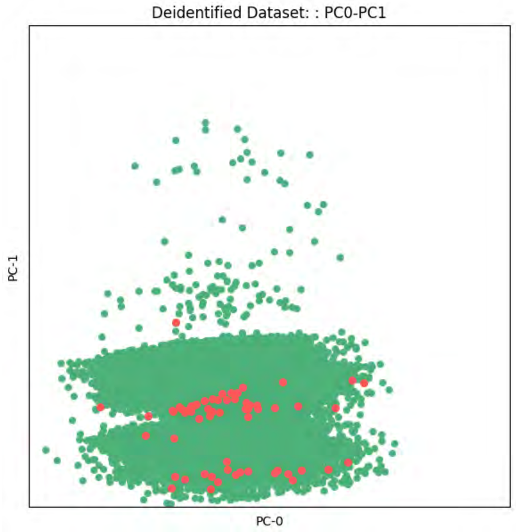
PATECTGAN Differential Private GAN (ε = 10)

What Happens on the Full 24 Feature Set?: Pairwise PCA

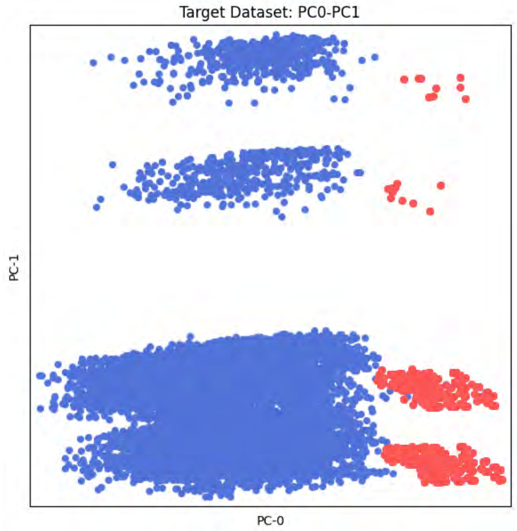
10 Feature Subset



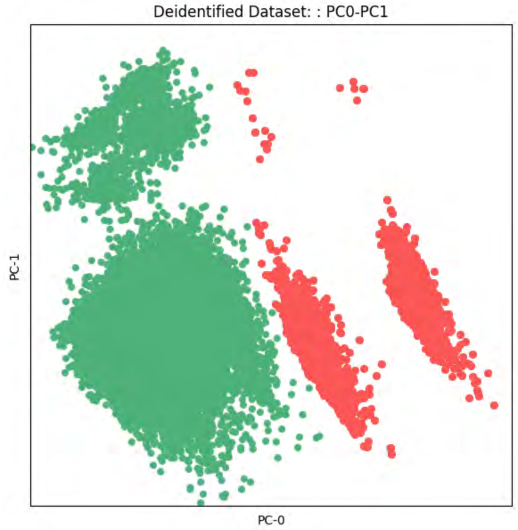
10 Feature Subset



10 Feature Subset Target

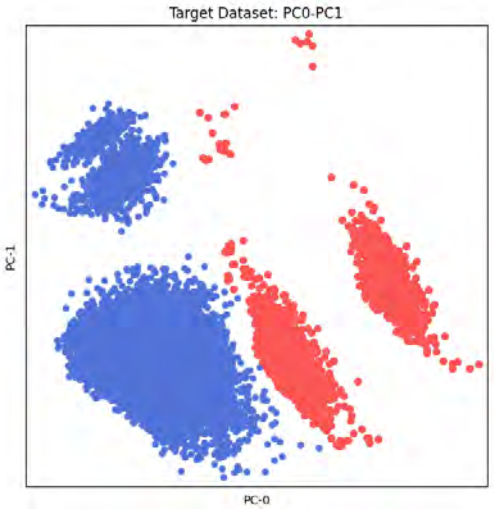


21 Features



21 Features

24 Feature Target



CART-model Synthesis (non-DP synthetic)

PATECTGAN Differential Private GAN (ε = 10)

What Happens on the Full 24 Feature Set?: UEM

10 Feature Subset

Percent of unique Target Data records exactly matched in Deid. Data: **20.32%**

21 Features

Percent of unique Target Data records exactly matched in Deid. Data: **2.4%**

CART-model Synthesis (non-DP synthetic)

10 Feature Subset

Percent of unique Target Data records exactly matched in Deid. Data: **7.1%**

21 Features

Percent of unique Target Data records exactly matched in Deid. Data: **0%**

PATECTGAN Differential Private GAN (ε = 10)

Unique Exact Match Rate: This is a count of unique records in the target data that were exactly reproduced in the deidentified data. Because these records were unique outliers in the target data, and they still appear unchanged in the deidentified data, they are potentially vulnerable to reidentification.

NIST Collaborative Research Cycle: Far more than four algorithms

https://pages.nist.gov/privacy_collaborative_research_cycle/

Collaborative Research Cycle

The CRC is an ongoing NIST program that provides resources for researching the behavior of deidentification (data privacy) on diverse populations.

Resources include:

- **Techniques Directory**
- Evaluation Reports
- Archive of Deidentified Data Samples

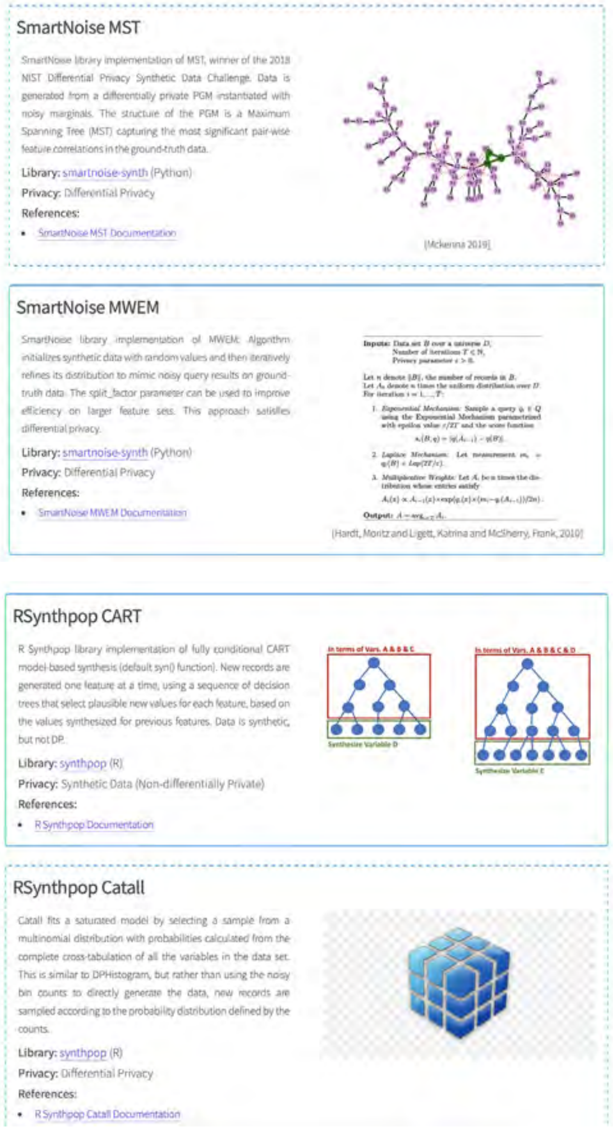
Contents:

Open Source:

- SmartNoise MST
- SmartNoise MWEM
- SmartNoise PACSynth
- SmartNoise PATE-CTGAN
- RSynthpop-CART
- RSynthpop Catall
- RSynthpop IPF
- SDV Copula-GAN
- SDV CTGAN
- SDV TVAE
- SDV Gaussian Copula
- SDV FAST-ML
- Synthcity DPGAN
- Synthcity PATEGAN
- Synthcity adsgan
- Synthcity bayesian_network
- Synthcity privbayes
- Synthcity TVAE
- Sdcmicro PRAM
- Sdcmicro K-anonymity

Commercial Products:

- MostlyAI-SD
- Sarus-SDG



SmartNoise MST

SmartNoise library implementation of MST, winner of the 2018 NIST Differential Privacy Synthetic Data Challenge. Data is generated from a differentially private PGM instantiated with noisy marginals. The structure of the PGM is a Maximum Spanning Tree (MST) capturing the most significant pair-wise feature correlations in the ground-truth data.

Library: smartnoise-synth (Python)

Privacy: Differential Privacy

References:

- SmartNoise MST Documentation

[Mckenna 2019]

SmartNoise MWEM

SmartNoise library implementation of MWEM: Algorithm initializes synthetic data with random values and then iteratively refines its distribution to mimic noisy query results on ground-truth data. The split_factor parameter can be used to improve efficiency on larger feature sets. This approach satisfies differential privacy.

Library: smartnoise-synth (Python)

Privacy: Differential Privacy

References:

- SmartNoise MWEM Documentation

[Hardt, Moritz and Ligett, Katrina and McSherry, Frank, 2010]

RSynthpop CART

R Synthpop library implementation of fully conditional CART model-based synthesis (default syn() function). New records are generated one feature at a time, using a sequence of decision trees that select plausible new values for each feature, based on the values synthesized for previous features. Data is synthetic, but not DP.

Library: synthpop (R)

Privacy: Synthetic Data (Non-differentially Private)

References:

- R Synthpop Documentation

RSynthpop Catall

Catall fits a saturated model by selecting a sample from a multinomial distribution with probabilities calculated from the complete cross-tabulation of all the variables in the data set. This is similar to DPHistogram, but rather than using the noisy bin counts to directly generate the data, new records are sampled according to the probability distribution defined by the counts.

Library: synthpop (R)

Privacy: Differential Privacy

References:

- R Synthpop Catall Documentation

Collaborative Research Cycle

NIST

The CRC is an ongoing NIST program that provides resources for researching the behavior of deidentification (data privacy) on diverse populations.

Resources include:

- Techniques Directory
- **Evaluation Reports**
- Archive of Deidentified Data Samples

Data Evaluation Report

Report created on: May 19, 2023 18:14:41

Created with SDNIST v2.2.1

Data Description

Deidentified (Deid.) Data:

Label Name	Label Value
Algorithm Name	pacsynth
Library	smartnoise-synth
Feature Set	family-focused
Target Dataset	national2019
Epsilon	10
Variant Label	preprocessor-epsilon: 3
Privacy	Differential Privacy

Property	Value
Filename	pac_synth_e_10_family_focused_na2019
Records	4579
Features	11

Target Data:

Property	Value
Filename	national2019
Records	27253
Features	24

Collaborative Research Cycle

The CRC is an ongoing NIST program that provides resources for researching the behavior of deidentification (data privacy) on diverse populations.

Resources include:

- Techniques Directory
- Evaluation Reports
- **Archive of Deidentified Data Samples**

The NIST Collaborative Research Cycle (CRC) Research Acceleration Bundle v1.1

- Direct download link for deidentified data and reports (537 MB)
- Direct download link for the metareports (484 MB)

Introduction

Welcome!

This repository contains deidentified data submitted to the CRC and their evaluation results as generated by SDNist v2.3.0. The CRC homepage provides more detailed information about the program, its goals, and how to participate.

In short, the CRC seeks to equip the research community with resources to explore, evaluate, and discuss deidentification approaches. The original data for this project are the NIST Diverse Communities Excerpts, curated data drawn from the American Community Survey.

There are three ground truth partitions, corresponding to three geographic regions (Boston-area (ma), Dallas-Fort Worth Area (tx), and a national sample (national). Submissions may include any or all of these partitions.

The original data contains 24 features. We also have a list of recommended reduced-size feature sets which can be found in the Excerpts Readme. Deidentified data may include any combination of feature set, though we have encouraged participants to use one of the recommended combinations to facilitate comparison of techniques.

What do we have here?

This repository contains the results of the first round of submissions. Additional submissions will be added with the next drop (expected in July 2023). The repository contains the navigable structure for the entire bundle. You can find all of the compressed data in Releases or you can use the links at the top of this readme.

The `crc-data-and-metrics-bundle` file contains:

- All of the deidentified data submissions and their evaluation metric results in the current release of our archive,
- An index.csv file that tracks all submission metadata, algorithm properties and definitions,
- A comprehensive set of tutorial jupyter notebooks and utilities that teach users how to programmatically explore the archive using the index file, and

Collaborative Research Cycle

The CRC is an ongoing NIST program that provides resources for researching the behavior of deidentification (data privacy) on diverse populations.

Resources include:

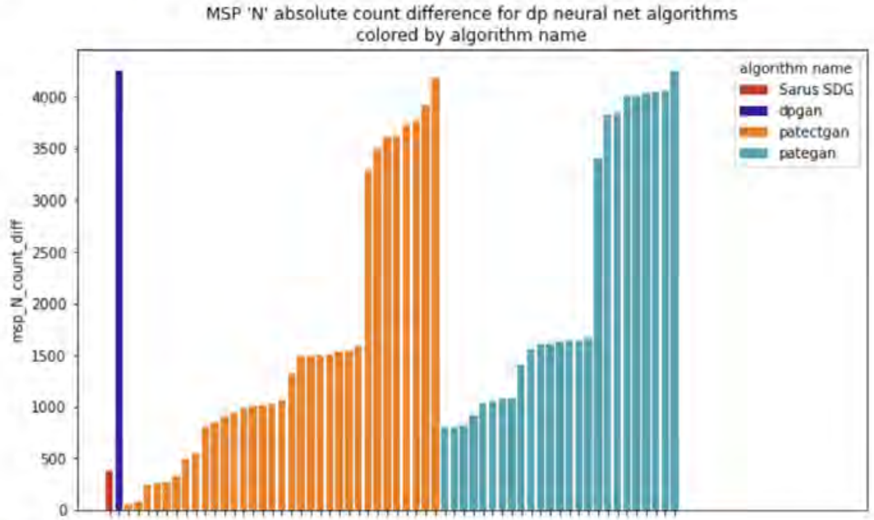
- Techniques Directory
- Evaluation Reports
- **Archive of Deidentified Data Samples**

Library	Algorithm	Team	#Entries	#Feature sets	Avg. Feat. Space Size	ε	Utility: SsE	Privacy Leak: UEM
rsynthpop	ipf_NonDP	Rsynthpop-categorical	1	1	3.405e+08		50.0	15.82
rsynthpop	catall_NonDP	Rsynthpop-categorical	1	1	2.270e+08		50.0	63.37
subsample_40pcnt	subsample_40pcnt	CRC	15	5	4.363e+25		40.67	39.93
rsynthpop	cart	CRC	12	4	3.457e+20		40.0	16.14
sdcmicro	pram	CRC	12	3	9.747e+10		38.33	56.27
MostlyAI SD	MostlyAI SD	MOSTLY AI	6	1	1.891e+26		30.0	0.01
rsynthpop	catall	Rsynthpop-categorical	6	1	2.270e+08	1, 10, 100	22.33	47.24
rsynthpop	cart	CBS-NL	3	1	2.270e+08		21.67	28.6
tumult	DPHist	CRC	5	2	5.732e+07	1, 2, 4, 10	18.8	92.14
smartnoise-synth	mst	CRC	36	5	3.781e+25	1, 5, 10	14.03	6.8
Genetic SD	Genetic SD	DataEvolution	19	2	9.454e+25	1, 10	11.84	0.11
LostInTheNoise	MWEM+PGM	LostInTheNoise	1	1	5.178e+26	1	10.0	0.0
synthcity	bayesian_network	CRC	12	4	5.672e+25		7.17	17.86
subsample_5pcnt	subsample_5pcnt	CRC	4	4	1.295e+26		5.0	4.97
Sarus SDG	Sarus SDG	Sarus	1	1	2.270e+08	10	5.0	13.99
sdv	ctgan	CBS-NL	6	1	1.891e+26		4.33	0.0

NIST Collaborative Research Cycle: Far more than four algorithms

Meta-analysis notebooks and tools available on the NIST CRC site make it easy to explore the archive

Introduction Tutorial:

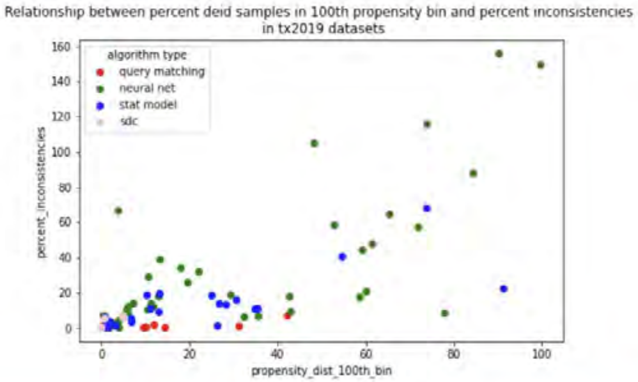


We teach all the basics for performing meta-analysis on the deidentified data archive:

1. Setup notebook.
2. Load deid datasets index file (index.csv).
3. Select specific deid. datasets from the index dataframe.
4. Working with the deidentified data csv files.
5. Working with the target data csv files.
6. Compare target and deid datasets.
7. Use index.csv to highlight plots by algorithm properties.
8. Access SDNIST evaluation reports.
9. Show relationship between two evaluation metrics.
10. Identify specific data samples of interest.
11. Show images from SDNist evaluation reports.
12. Get evaluation metrics for specific samples of interest.

Example Notebook 2: Imposter plot (propensity scores and inconsistencies).

We show how to collect the counts of individuals in the 100% confidence propensity bin (obviously synthetic records, 'imposters') across all deidentified data submissions. We then provide a scatterplot comparing imposter count with inconsistencies count. This notebook demonstrates accessing metrics in report.json files and metric result .csv files, and using the CRC plotting utility to make highlighted scatter plots.



Example Notebook 3: Race distribution

We show how to check deidentification impact on race distribution by directly counting individuals in the target and deidentified data csv files. We filter the index.csv metadata to print a data frame containing this score alongside relevant algorithm properties.

	library name	algorithm name	algorithm type	target dataset	privacy category	epsilon	feature set name	race_edit_distance_percent
241	synthcity	adsgan	neural net	ma2019	non_dp	NaN	industry-focused	14.38
224	subsample_40pcnt	subsample_40pcnt	sdc	national2019	sdc	NaN	all-features	0.61
67	sdcmicro	kanonymity	sdc	tx2019	sdc	NaN	demographic-focused	24.58
179	smartnoise-synth	pacsynth	query matching	tx2019	dp	5.0	demographic-focused	26.72
72	sdcmicro	kanonymity	sdc	tx2019	sdc	NaN	industry-focused	31.16
85	sdv	copula-gan	neural net	ma2019	non_dp	NaN	custom-features-23	88.43
155	smartnoise-synth	mwem	query matching	national2019	dp	10.0	custom-features-12	1.09
274	synthcity	pategan	neural net	national2019	dp	1.0	all-features	22.96
281	synthcity	pategan	neural net	tx2019	dp	10.0	all-features	5.13
121	smartnoise-synth	mst	stat model	ma2019	dp	1.0	demographic-focused	1.47
295	synthcity	privbayes	stat model	national2019	dp	1.0	demographic-focused	43.08
132	smartnoise-synth	mst	stat model	national2019	dp	10.0	demographic-focused	0.05
320	tumult	DPHist	histogram	national2019	dp	1.0	demographic-focused-except-AGEP-DEYE	113.01
229	subsample_40pcnt	subsample_40pcnt	sdc	tx2019	sdc	NaN	all-features	1.82
306	synthcity	privbayes	stat model	tx2019	dp	1.0	industry-focused	0.86
6	Genetic SD	Genetic SD	query matching	national2019	dp	1.0	all-features	4.45
137	smartnoise-synth	mst	stat model	national2019	dp	10.0	family-focused	0.15

NIST Collaborative Research Cycle: Far more than four algorithms

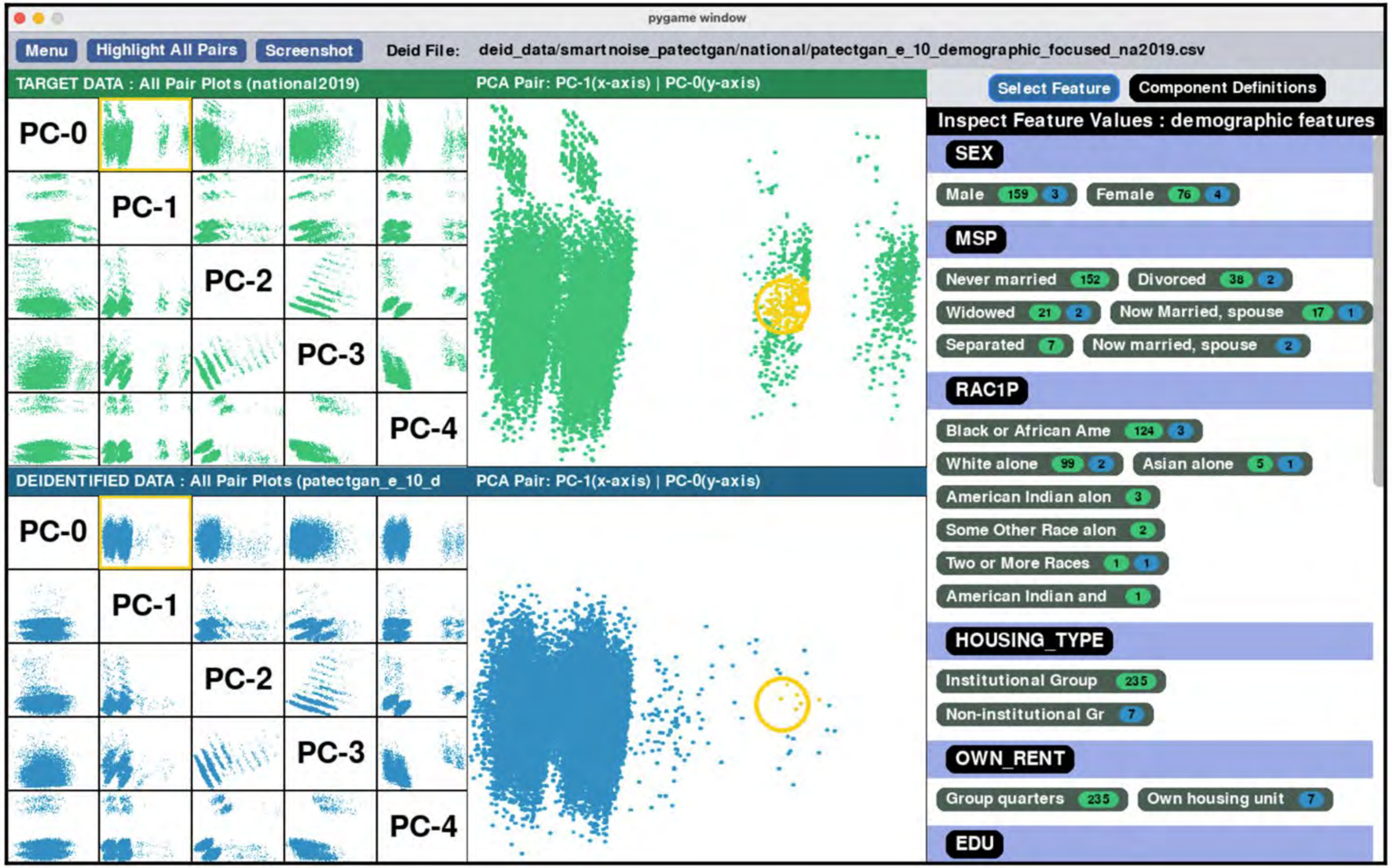
Pair-wise PCA Inspection Tool

Pairwise PCA is a relatively new visualization metric that was introduced by the IPUMS International team during the HLG-MOS Synthetic Data Test Drive.

It lets us look at the high dimensional data distribution using a set of 2D scatterplots along principle component axes. The plots look at the deidentified data and target data from the same angle (ie, using axes from the target data), so we can directly see where their distributions differ from each other.

The pairwise PCA tool lets you interactively explore these plots using a GUI interface.

You can install it by following the directions here: https://github.com/usnistgov/pair-wise_PCA



NIST Collaborative Research Cycle: Far more than four algorithms

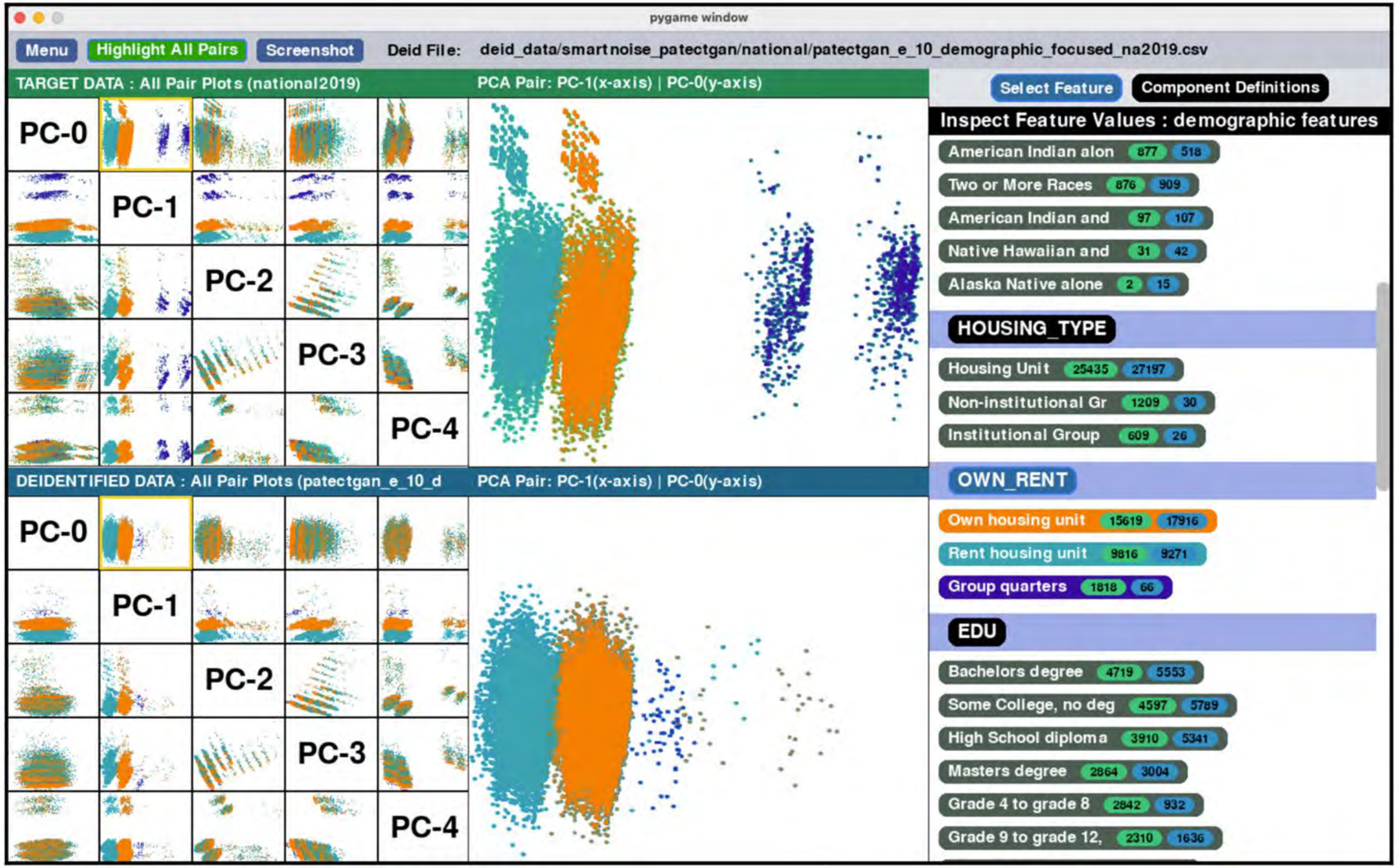
Pair-wise PCA Inspection Tool

Pairwise PCA is a relatively new visualization metric that was introduced by the IPUMS International team during the HLG-MOS Synthetic Data Test Drive.

It lets us look at the high dimensional data distribution using a set of 2D scatterplots along principle component axes. The plots look at the deidentified data and target data from the same angle (ie, using axes from the target data), so we can directly see where their distributions differ from each other.

The pairwise PCA tool lets you interactively explore these plots using a GUI interface.

You can install it by following the directions here: https://github.com/usnistgov/pair-wise_PCA



Collaborative Research Cycle (CRC) Strategy

- 12-month program that collects, disseminates, and analyzes synthetic data
- No prize money, low barrier to participation, emphasizes cooperation

Exploratory Phase (Feb. 2023 - July. 2023)

- NIST releases Diverse Community Excerpt Data
- Participants submit de-identified data and abstract on techniques
- NIST releases machine readable analysis of submissions (*acceleration bundle*)

Explanatory Phase (July. 2023 - Dec. 2023)

- Participants perform meta-analysis on the acceleration bundle.
- NIST hosts a seminar series and conducts outreach
- Participants submit papers on their analyses Nov 17th 2023
- NIST hosts conference Dec. 18 2023

Thank you!
Questions?

gary.howarth@nist.gov
christine.task@knexusresearch.com

https://pages.nist.gov/privacy_collaborative_research_cycle/

google keywords: NIST collaborative research cycle
or NIST CRC