The Collaborative Research Cycle

Christine Task, PhD
Lead Privacy Researcher
Knexus Research Corporation

Karan Bhagat
Research Developer
Knexus Research Corporation

Gary Howarth, PhD
Physical Scientist
NIST

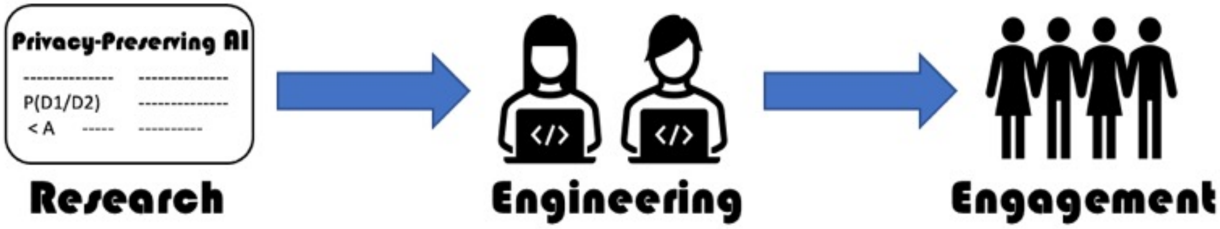
NIST
NATIONAL INSTITUTE OF STANDARDS AND TECHNOLOGY
U.S. DEPARTMENT OF COMMERCE

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.



Research

Engineering

Engagement

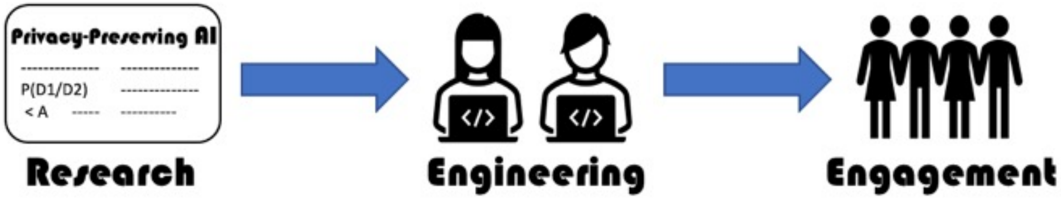
Here’s a basic path from research to production: We start with an abstract problem definition and begin researching potential solutions.

When one approach seems to work very well in evaluations, we begin engineering a production-grade implementation.

When the implementation is complete, we release the tool for real world use.

…this isn’t how anything works

USCB Disclosure Avoidance System: 2017 -- 2018



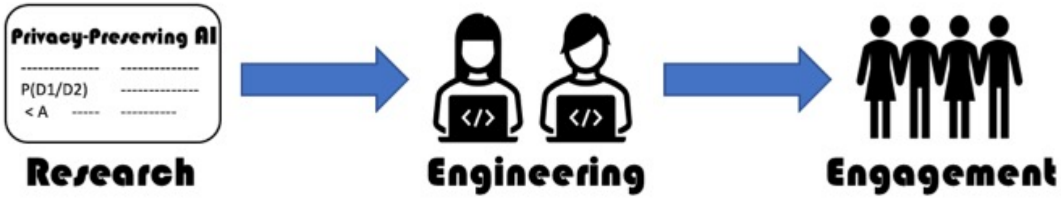
Research

Engineering

Engagement

Disclaimer: Assembled from census.gov, CNSTAT website, google scholar, and vague memory.

USCB Disclosure Avoidance System: 2017 -- 2018



Research

- Understanding hierarchical methods for differentially private histograms. [Qardaji, Wahbeh and Yang, Weining and Li, Ninghu]

- Optimizing linear counting queries under differential privacy [Li, Chao and Hay, Michael and Rastogi, Vibhor and Miklau, Gerome and McGregor, Andrew]

- The modernization of statistical disclosure limitation at the US Census Bureau [Aref N. Dajani, Amy D. Lauger, Phyllis E. Singer, Daniel Kifer, Jerome P. Reiter, Ashwin Machanavajjhala, Simson L. Garfinkel, Scot A. Dahl, Matthew Graham, Vishesh Karwa, Hang Kim, Philip Leclerc, Ian M. Schmutte, William N. Sexton, Lars Vilhuber, John M. Abowd]

Engineering

- Initial Implementation, to run on 2018 End to End test data.

Engagement

- Initial Media Outreach
- Presentations to researchers (ex: KDD)

Disclaimer: Assembled from census.gov, CNSTAT website, google scholar, and vague memory.

USCB Disclosure Avoidance System: 2019 -- 2020

Privacy-Preserving AI
P(D1/D2)
< A
Research
Engineering
Engagement
Privacy-Preserving AI
P(D1/D2)
< A
Research
Engineering
Engagement

Research

- Census topdown: Differentially private data, incremental schemas, and consistency with public knowledge [John Abowd, Robert Ashmead, Garfinkel Simson, Daniel Kifer, Philip Leclerc, Ashwin Machanavajjhala, William Sexton]
- Deploying differential privacy for the 2020 census of population and housing [Simson L. Garfinkel]
- Differential Privacy in the Real World: The 2018 End-to-End Census Test [John Abowed]

Engineering

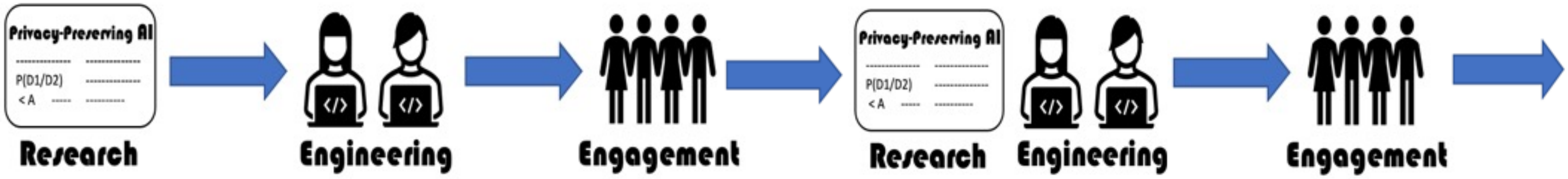
- Release of first public demonstration data product: **2010 Demonstration Data Products Baseline 2019-10-29**

Engagement

- CNSTAT Workshops and working groups

Disclaimer: Assembled from census.gov, CNSTAT website, google scholar, and vague memory.

USCB Disclosure Avoidance System: 2019 -- 2020



Research

- Census topdown: Differentially private data, incremental schemas, and consistency with public knowledge [John Abowd, Robert Ashmead, Garfinkel Simson, Daniel Kifer, Philip Leclerc, Ashwin Machanavajjhala, William Sexton]
- Deploying differential privacy for the 2020 census of population and housing [Simson L. Garfinkel]
- Differential Privacy in the Real World: The 2018 End-to-End Census Test [John Abowed]

Engineering

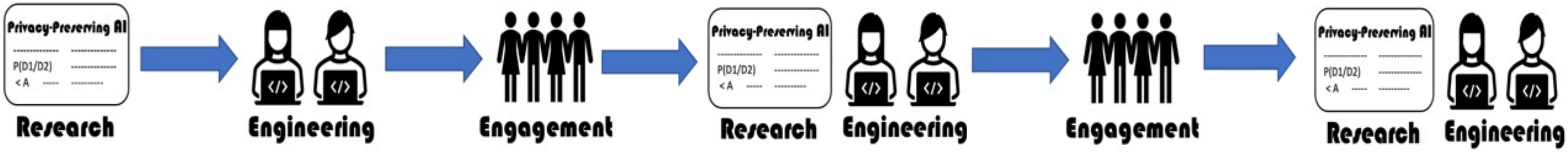
- Release of first public demonstration data product: **2010 Demonstration Data Products Baseline 2019-10-29**

Engagement

- CNSTAT Workshops and working groups
- Ongoing talks at data users' and privacy researchers' conferences.
- Ongoing media outreach
- Federal Register Notices (Public requests for comment)
- National Advisory Committee (NAC)
- Census Advisory Committees (CAC)

Disclaimer: Assembled from census.gov, CNSTAT website, google scholar, and vague memory.

USCB Disclosure Avoidance System: 2019 -- 2020



Research → Engineering → Engagement → Research Engineering → Engagement → Research Engineering

Research

- Census topdown: Differentially private data, incremental schemas, and consistency with public knowledge [John Abowd, Robert Ashmead, Garfinkel Simson, Daniel Kifer, Philip Leclerc, Ashwin Machanavajjhala, William Sexton]
- Deploying differential privacy for the 2020 census of population and housing [Simson L. Garfinkel]
- Differential Privacy in the Real World: The 2018 End-to-End Census Test [John Abowed]

Engineering

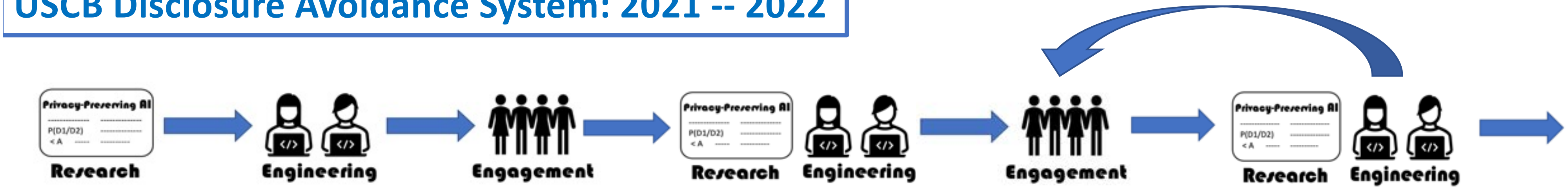
- Release of first public demonstration data product: **2010 Demonstration Data Products Baseline 2019-10-29**
- Release of second public demonstration data**: DAS Development Update 2020-11-16**
- Release of third public demonstration data: **DAS Development Update 2020-09-17**
- Release of fourth public demonstration data: **DAS Development Update 2020-05-27**

Engagement

- CNSTAT Workshops and working groups
- Ongoing talks at data users' and privacy researchers' conferences.
- Ongoing media outreach
- Federal Register Notices (Public requests for comment)
- National Advisory Committee (NAC)
- Census Advisory Committees (CAC)

Disclaimer: Assembled from census.gov, CNSTAT website, google scholar, and vague memory.

USCB Disclosure Avoidance System: 2021 -- 2022



Research

Engineering

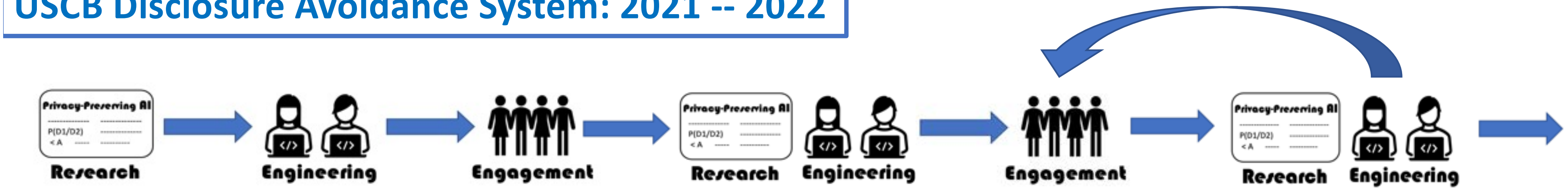
- Release of fifth public demonstration data product: **DAS Development Update 2021-04-28**
- Production: **Production Settings 2021-06-08**

Engagement

- CNSTAT Workshops and working groups
- Ongoing talks at data users' and privacy researchers' conferences.
- Ongoing media outreach
- Federal Register Notices (Public requests for comment)
- National Advisory Committee (NAC)
- Census Advisory Committees (CAC)

Disclaimer: Assembled from census.gov, CNSTAT website, google scholar, and vague memory.

USCB Disclosure Avoidance System: 2021 -- 2022



Research

- Bayesian and Frequentist Semantics for Common Variations of Differential Privacy: Applications to the 2020 Census.
- An uncertainty principle is a price of privacy-preserving microdata
- Data Science for Governing and Policy Making
- The 2020 census disclosure avoidance system topdown algorithm
- Exact Privacy Analysis of the Gaussian Sparse Histogram Mechanism
- Confidentiality protection in the 2020 US Census of population and housing
- Special Issue 2: Differential Privacy for the 2020 US Census
- Finding Needles in Haystacks: Multiple-Imputation Record Linkage Using Machine Learning
- Total error and variability measures for the quarterly workforce indicators and LEHD origin-destination employment statistics in OnTheMap
- Implementing differential privacy for the 2020 census

Engineering

- Release of fifth public demonstration data product: **DAS Development Update 2021-04-28**

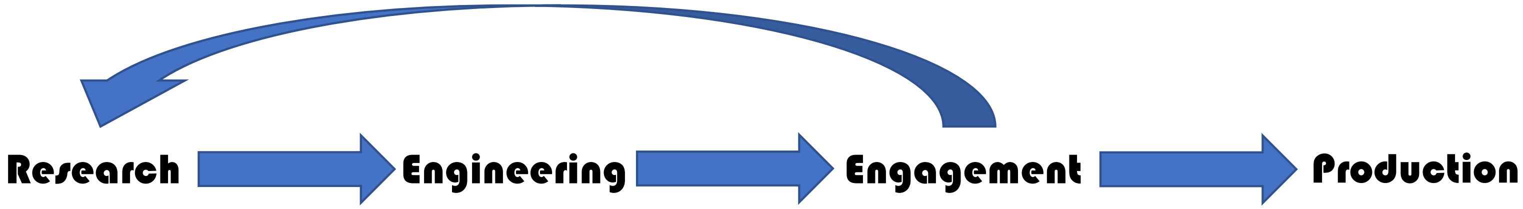
- Production: **Production Settings 2021-06-08**

Engagement

- CNSTAT Workshops and working groups
- Ongoing talks at data users' and privacy researchers' conferences.
- Ongoing media outreach
- Federal Register Notices (Public requests for comment)
- National Advisory Committee (NAC)
- Census Advisory Committees (CAC)

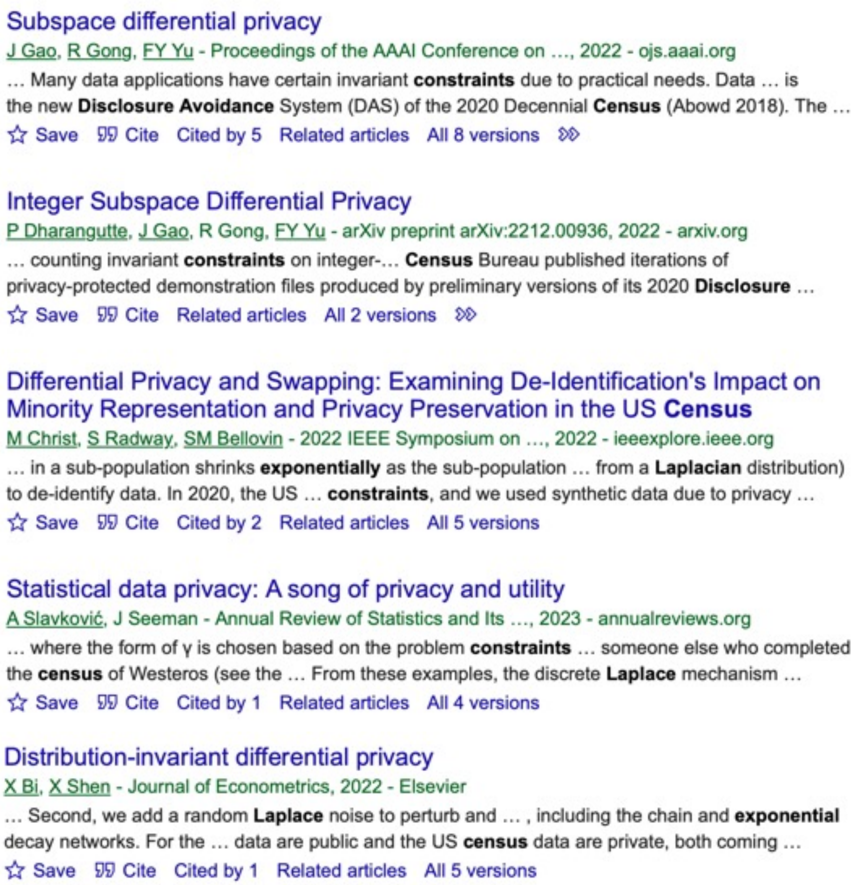
Disclaimer: Assembled from census.gov, CNSTAT website, google scholar, and vague memory.

2022 -- ??



Research → Engineering → Engagement → Production

Research



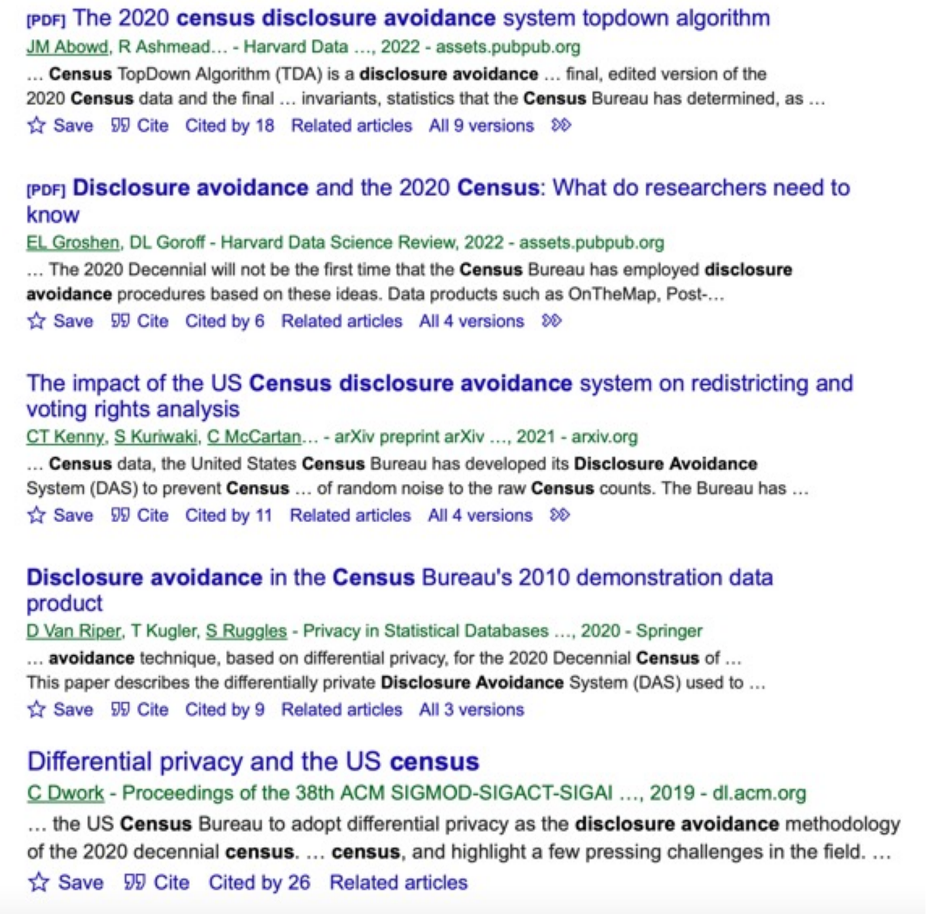
Subspace differential privacy
J Gao, R Gong, FY Yu - Proceedings of the AAAI Conference on …, 2022 - ojs.aaai.org
… Many data applications have certain invariant **constraints** due to practical needs. Data … is the new **Disclosure Avoidance** System (DAS) of the 2020 Decennial **Census** (Abowd 2018). The …
☆ Save ❞ Cite Cited by 5 Related articles All 8 versions »

Integer Subspace Differential Privacy
P Dharangutte, J Gao, R Gong, FY Yu - arXiv preprint arXiv:2212.00936, 2022 - arxiv.org
… counting invariant **constraints** on integer-… **Census** Bureau published iterations of privacy-protected demonstration files produced by preliminary versions of its 2020 **Disclosure** …
☆ Save ❞ Cite Related articles All 2 versions »

Differential Privacy and Swapping: Examining De-Identification's Impact on Minority Representation and Privacy Preservation in the US Census
M Christ, S Radway, SM Bellovin - 2022 IEEE Symposium on …, 2022 - ieeexplore.ieee.org
… in a sub-population shrinks **exponentially** as the sub-population … from a **Laplacian** distribution) to de-identify data. In 2020, the US … **constraints**, and we used synthetic data due to privacy …
☆ Save ❞ Cite Cited by 2 Related articles All 5 versions

Statistical data privacy: A song of privacy and utility
A Slavković, J Seeman - Annual Review of Statistics and Its …, 2023 - annualreviews.org
… where the form of γ is chosen based on the problem **constraints** … someone else who completed the **census** of Westeros (see the … From these examples, the discrete **Laplace** mechanism …
☆ Save ❞ Cite Cited by 1 Related articles All 4 versions

Distribution-invariant differential privacy
X Bi, X Shen - Journal of Econometrics, 2022 - Elsevier
… Second, we add a random **Laplace** noise to perturb and … , including the chain and **exponential** decay networks. For the … data are public and the US **census** data are private, both coming …
☆ Save ❞ Cite Cited by 1 Related articles All 5 versions



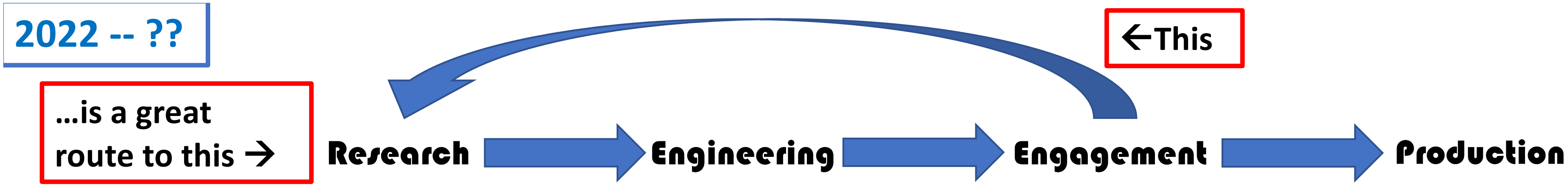
[PDF] **The 2020 census disclosure avoidance system topdown algorithm**
JM Abowd, R Ashmead… - Harvard Data …, 2022 - assets.pubpub.org
… **Census** TopDown Algorithm (TDA) is a **disclosure avoidance** … final, edited version of the 2020 **Census** data and the final … invariants, statistics that the **Census** Bureau has determined, as …
☆ Save ❞ Cite Cited by 18 Related articles All 9 versions »

[PDF] **Disclosure avoidance and the 2020 Census: What do researchers need to know**
EL Groshen, DL Goroff - Harvard Data Science Review, 2022 - assets.pubpub.org
… The 2020 Decennial will not be the first time that the **Census** Bureau has employed **disclosure avoidance** procedures based on these ideas. Data products such as OnTheMap, Post-…
☆ Save ❞ Cite Cited by 6 Related articles All 4 versions »

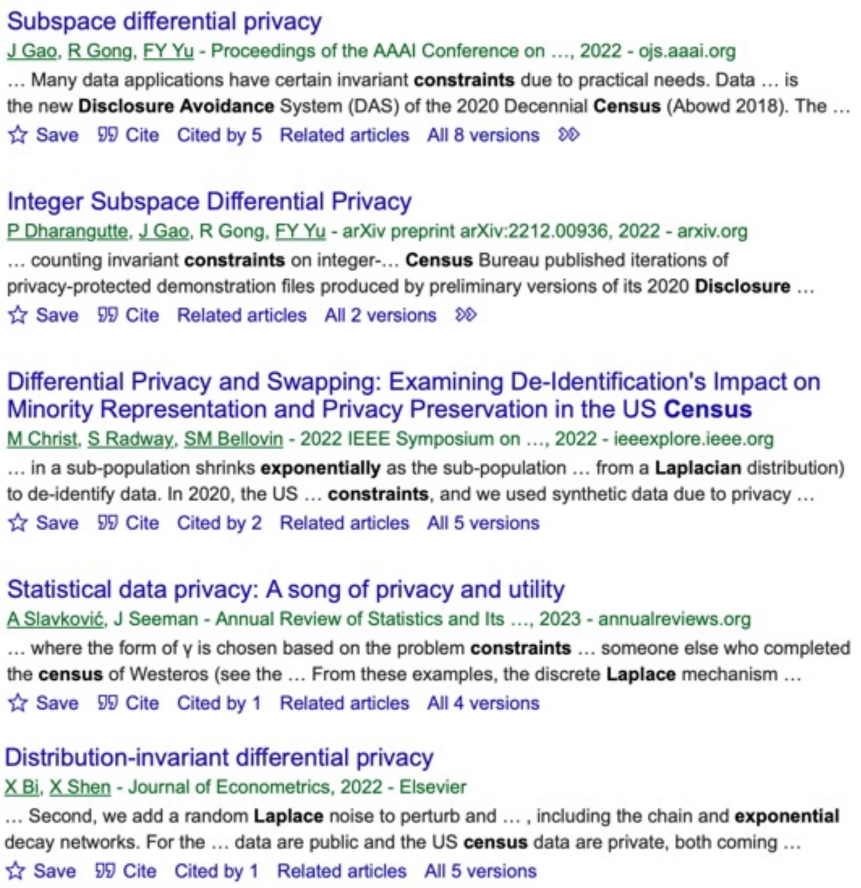
The impact of the US Census disclosure avoidance system on redistricting and voting rights analysis
CT Kenny, S Kuriwaki, C McCartan… - arXiv preprint arXiv …, 2021 - arxiv.org
… **Census** data, the United States **Census** Bureau has developed its **Disclosure Avoidance** System (DAS) to prevent **Census** … of random noise to the raw **Census** counts. The Bureau has …
☆ Save ❞ Cite Cited by 11 Related articles All 4 versions »

Disclosure avoidance in the Census Bureau's 2010 demonstration data product
D Van Riper, T Kugler, S Ruggles - Privacy in Statistical Databases …, 2020 - Springer
… **avoidance** technique, based on differential privacy, for the 2020 Decennial **Census** of … This paper describes the differentially private **Disclosure Avoidance** System (DAS) used to …
☆ Save ❞ Cite Cited by 9 Related articles All 3 versions

Differential privacy and the US census
C Dwork - Proceedings of the 38th ACM SIGMOD-SIGACT-SIGAI …, 2019 - dl.acm.org
… the US **Census** Bureau to adopt differential privacy as the **disclosure avoidance** methodology of the 2020 decennial **census**. … **census**, and highlight a few pressing challenges in the field. …
☆ Save ❞ Cite Cited by 26 Related articles



Research



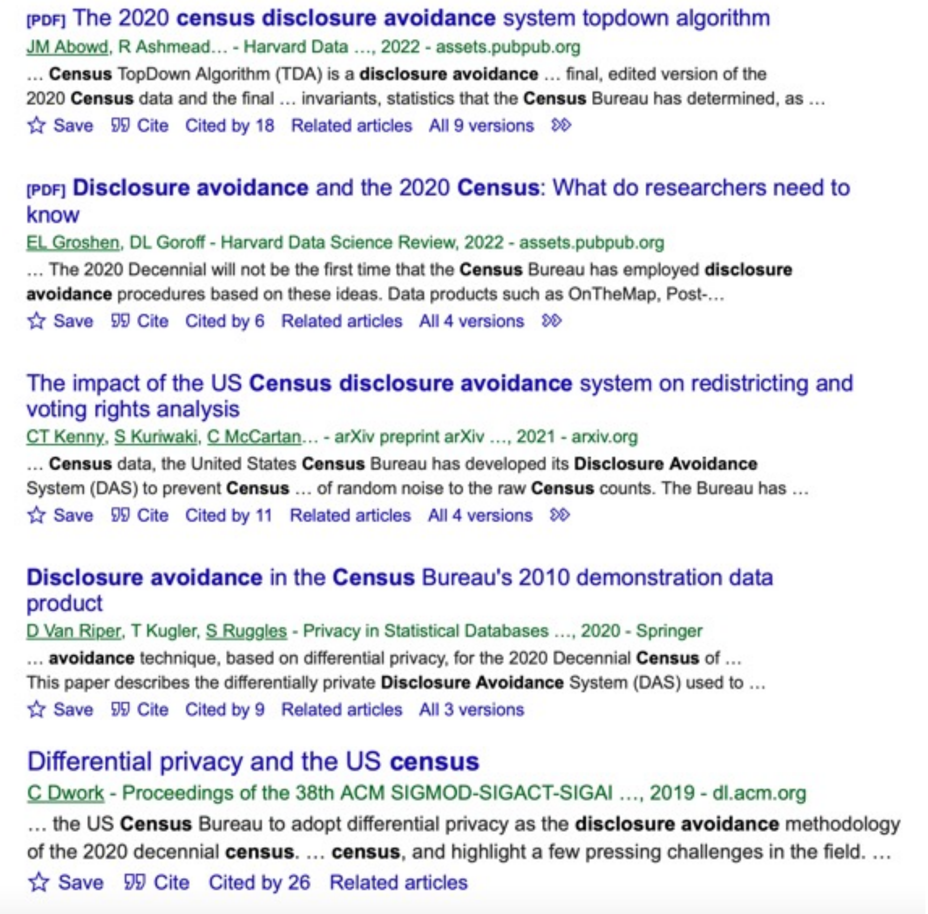
Subspace differential privacy
J Gao, R Gong, FY Yu - Proceedings of the AAAI Conference on …, 2022 - ojs.aaai.org
… Many data applications have certain invariant **constraints** due to practical needs. Data … is the new **Disclosure Avoidance** System (DAS) of the 2020 Decennial **Census** (Abowd 2018). The …
☆ Save 99 Cite Cited by 5 Related articles All 8 versions

Integer Subspace Differential Privacy
P Dharangutte, J Gao, R Gong, FY Yu - arXiv preprint arXiv:2212.00936, 2022 - arxiv.org
… counting invariant **constraints** on integer-… **Census** Bureau published iterations of privacy-protected demonstration files produced by preliminary versions of its 2020 **Disclosure** …
☆ Save 99 Cite Related articles All 2 versions

Differential Privacy and Swapping: Examining De-Identification's Impact on Minority Representation and Privacy Preservation in the US **Census**
M Christ, S Radway, SM Bellovin - 2022 IEEE Symposium on …, 2022 - ieeexplore.ieee.org
… in a sub-population shrinks **exponentially** as the sub-population … from a **Laplacian** distribution) to de-identify data. In 2020, the US … **constraints**, and we used synthetic data due to privacy …
☆ Save 99 Cite Cited by 2 Related articles All 5 versions

Statistical data privacy: A song of privacy and utility
A Slavković, J Seeman - Annual Review of Statistics and Its …, 2023 - annualreviews.org
… where the form of γ is chosen based on the problem **constraints** … someone else who completed the **census** of Westeros (see the … From these examples, the discrete **Laplace** mechanism …
☆ Save 99 Cite Cited by 1 Related articles All 4 versions

Distribution-invariant differential privacy
X Bi, X Shen - Journal of Econometrics, 2022 - Elsevier
… Second, we add a random **Laplace** noise to perturb and … , including the chain and **exponential** decay networks. For the … data are public and the US **census** data are private, both coming …
☆ Save 99 Cite Cited by 1 Related articles All 5 versions



[PDF] The 2020 **census disclosure avoidance** system topdown algorithm
JM Abowd, R Ashmead… - Harvard Data …, 2022 - assets.pubpub.org
… **Census** TopDown Algorithm (TDA) is a **disclosure avoidance** … final, edited version of the 2020 **Census** data and the final … invariants, statistics that the **Census** Bureau has determined, as …
☆ Save 99 Cite Cited by 18 Related articles All 9 versions

[PDF] **Disclosure avoidance** and the 2020 **Census**: What do researchers need to know
EL Groshen, DL Goroff - Harvard Data Science Review, 2022 - assets.pubpub.org
… The 2020 Decennial will not be the first time that the **Census** Bureau has employed **disclosure avoidance** procedures based on these ideas. Data products such as OnTheMap, Post-…
☆ Save 99 Cite Cited by 6 Related articles All 4 versions

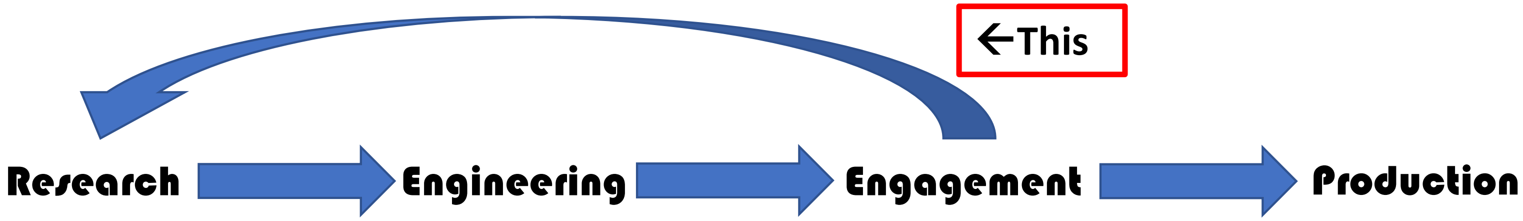
The impact of the US **Census disclosure avoidance** system on redistricting and voting rights analysis
CT Kenny, S Kuriwaki, C McCartan… - arXiv preprint arXiv …, 2021 - arxiv.org
… **Census** data, the United States **Census** Bureau has developed its **Disclosure Avoidance** System (DAS) to prevent **Census** … of random noise to the raw **Census** counts. The Bureau has …
☆ Save 99 Cite Cited by 11 Related articles All 4 versions

Disclosure avoidance in the **Census** Bureau's 2010 demonstration data product
D Van Riper, T Kugler, S Ruggles - Privacy in Statistical Databases …, 2020 - Springer
… **avoidance** technique, based on differential privacy, for the 2020 Decennial **Census** of … This paper describes the differentially private **Disclosure Avoidance** System (DAS) used to …
☆ Save 99 Cite Cited by 9 Related articles All 3 versions

Differential privacy and the US **census**
C Dwork - Proceedings of the 38th ACM SIGMOD-SIGACT-SIGAI …, 2019 - dl.acm.org
… the US **Census** Bureau to adopt differential privacy as the **disclosure avoidance** methodology of the 2020 decennial **census**. … **census**, and highlight a few pressing challenges in the field. …
☆ Save 99 Cite Cited by 26 Related articles

2018 -- 2023

…is a great route to this →



←This

But it's often not cheap, or easy to access

Engagement

Federal Committee on Statistical Methodology

IPUMS USA

American Community Survey Data Users Group

NORC at the University of Chicago

Committee on National Statistics (CNSTAT)

National Academies Sciences Engineering Medicine

Census Scientific Advisory Committee (CSAC)

National Advisory Committee (NAC)

Engineering

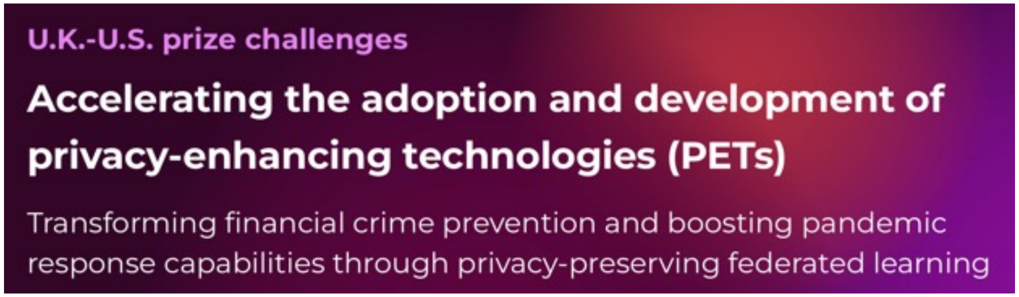
KNEXUS Research Corporation

MITRE

|galois|

TUMULT LABS

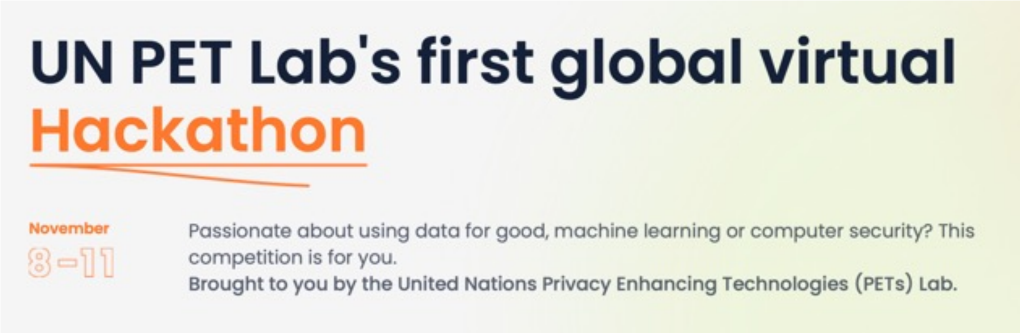
Hackathons and Challenges: An Easy and Accessible Path Through the Engineering and Engagement Loop



U.K.-U.S. prize challenges

Accelerating the adoption and development of privacy-enhancing technologies (PETs)

Transforming financial crime prevention and boosting pandemic response capabilities through privacy-preserving federated learning

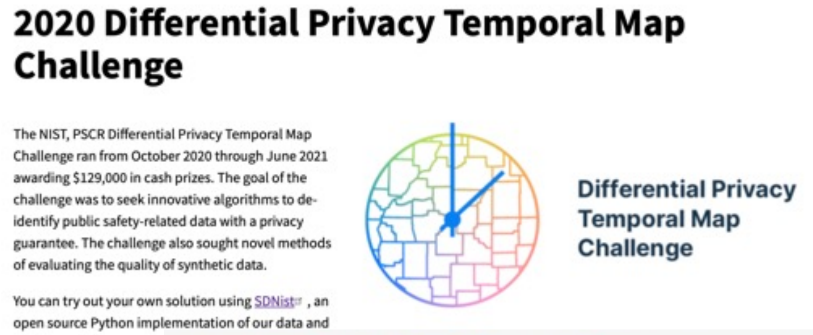


UN PET Lab's first global virtual Hackathon

November 8–11

Passionate about using data for good, machine learning or computer security? This competition is for you.

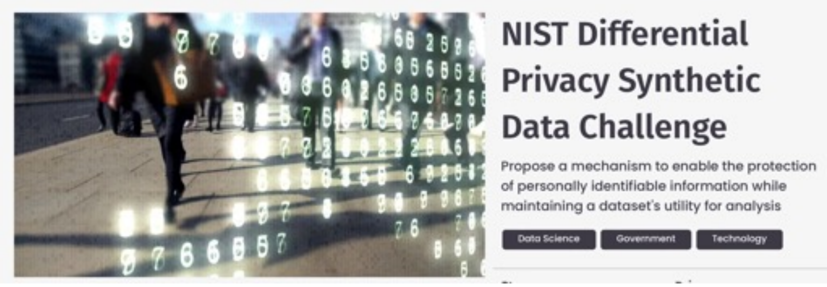
Brought to you by the United Nations Privacy Enhancing Technologies (PETs) Lab.



2020 Differential Privacy Temporal Map Challenge

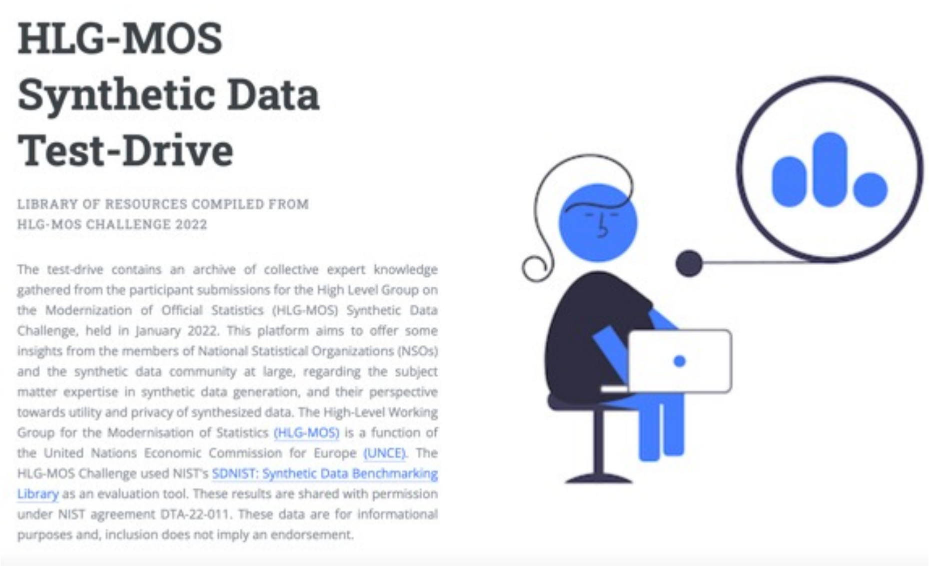
The NIST, PSCR Differential Privacy Temporal Map Challenge ran from October 2020 through June 2021 awarding $129,000 in cash prizes. The goal of the challenge was to seek innovative algorithms to de-identify public safety-related data with a privacy guarantee. The challenge also sought novel methods of evaluating the quality of synthetic data.

You can try out your own solution using SDNist , an open source Python implementation of our data and scoring metrics.



NIST Differential Privacy Synthetic Data Challenge

Propose a mechanism to enable the protection of personally identifiable information while maintaining a dataset's utility for analysis



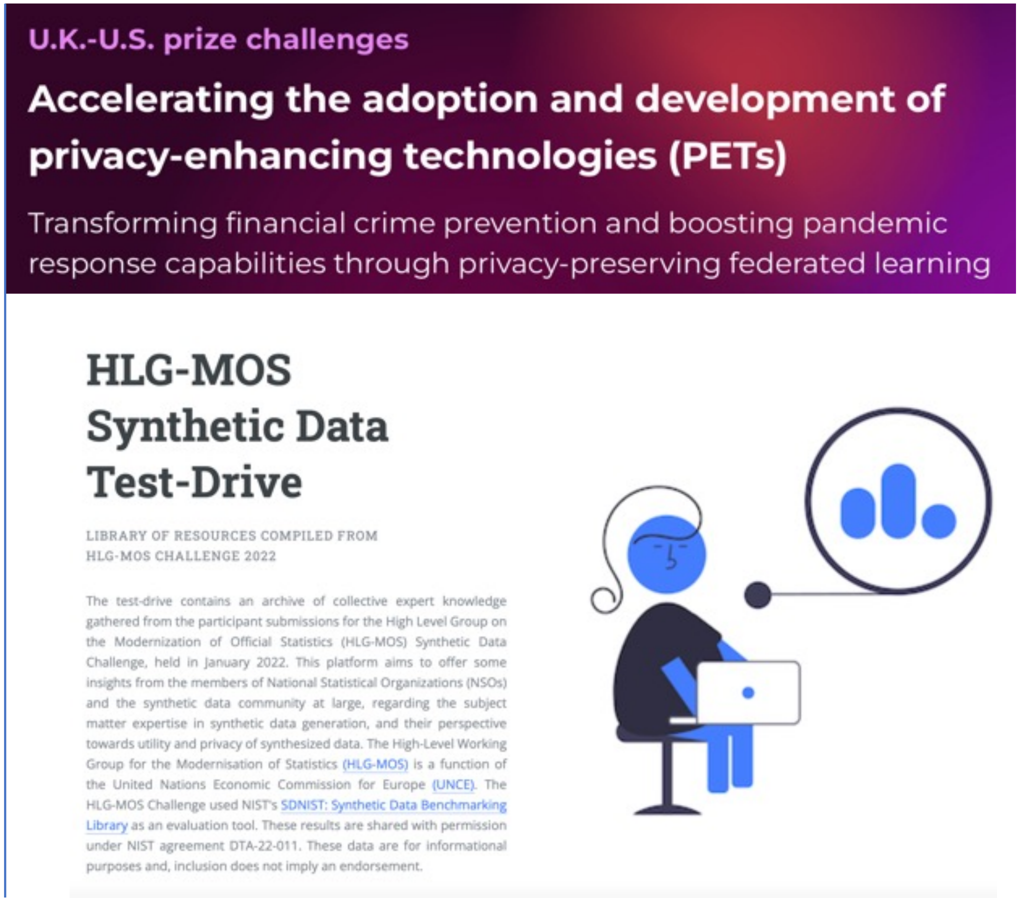
HLG-MOS Synthetic Data Test-Drive

LIBRARY OF RESOURCES COMPILED FROM HLG-MOS CHALLENGE 2022

The test-drive contains an archive of collective expert knowledge gathered from the participant submissions for the High Level Group on the Modernization of Official Statistics (HLG-MOS) Synthetic Data Challenge, held in January 2022. This platform aims to offer some insights from the members of National Statistical Organizations (NSOs) and the synthetic data community at large, regarding the subject matter expertise in synthetic data generation, and their perspective towards utility and privacy of synthesized data. The High-Level Working Group for the Modernisation of Statistics (HLG-MOS) is a function of the United Nations Economic Commission for Europe (UNCE). The HLG-MOS Challenge used NIST's SDNIST: Synthetic Data Benchmarking Library as an evaluation tool. These results are shared with permission under NIST agreement DTA-22-011. These data are for informational purposes and, inclusion does not imply an endorsement.

Hackathons and Challenges: An Easy and Accessible Path Through the Engineering and Engagement Loop

- Well-designed Challenges and Hackathons can serve as a bridge between researchers and users/stakeholders. This allows researchers to get the benefit of engineering their solutions for real world data problems, and engaging with scoring metrics that capture real world stakeholder concerns.
- And the research cycle works here too: Challenges and Hackathons do produce new Research Problems--- but a lot of times it's only us running things backstage, tucked behind an NDA, who get to see ***all*** of the new problems. Each team only sees their own.
- So we thought we might swap that around for you guys.



U.K.-U.S. prize challenges

Accelerating the adoption and development of privacy-enhancing technologies (PETs)

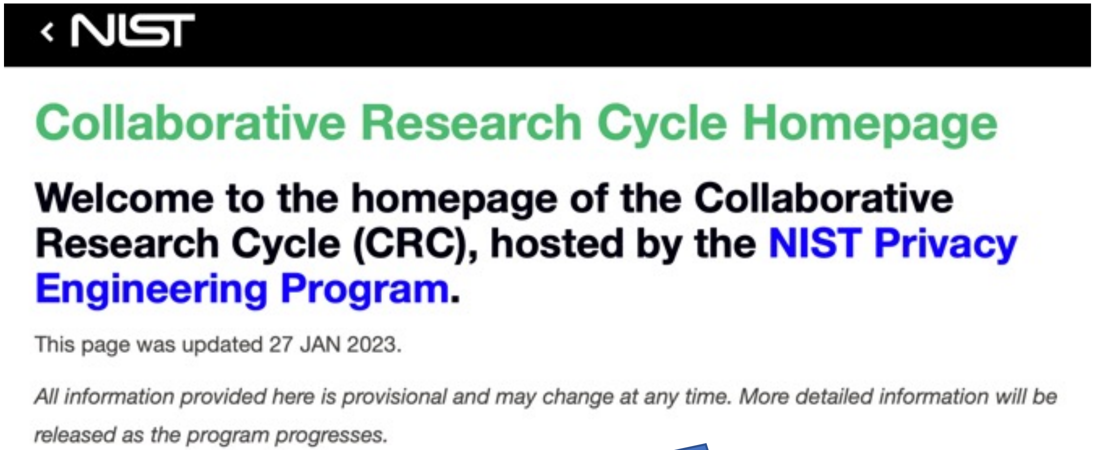
Transforming financial crime prevention and boosting pandemic response capabilities through privacy-preserving federated learning

HLG-MOS Synthetic Data Test-Drive

LIBRARY OF RESOURCES COMPILED FROM HLG-MOS CHALLENGE 2022

The test-drive contains an archive of collective expert knowledge gathered from the participant submissions for the High Level Group on the Modernization of Official Statistics (HLG-MOS) Synthetic Data Challenge, held in January 2022. This platform aims to offer some insights from the members of National Statistical Organizations (NSOs) and the synthetic data community at large, regarding the subject matter expertise in synthetic data generation, and their perspective towards utility and privacy of synthesized data. The High-Level Working Group for the Modernisation of Statistics (HLG-MOS) is a function of the United Nations Economic Commission for Europe (UNCE). The HLG-MOS Challenge used NIST's SDNIST: Synthetic Data Benchmarking Library as an evaluation tool. These results are shared with permission under NIST agreement DTA-22-011. These data are for informational purposes and, inclusion does not imply an endorsement.

Hackathons and Challenges: An Easy and Accessible Path Through the Engineering and Engagement Loop



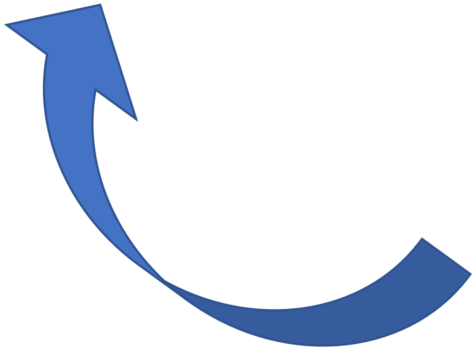
‹ NIST

Collaborative Research Cycle Homepage

Welcome to the homepage of the Collaborative Research Cycle (CRC), hosted by the NIST Privacy Engineering Program.

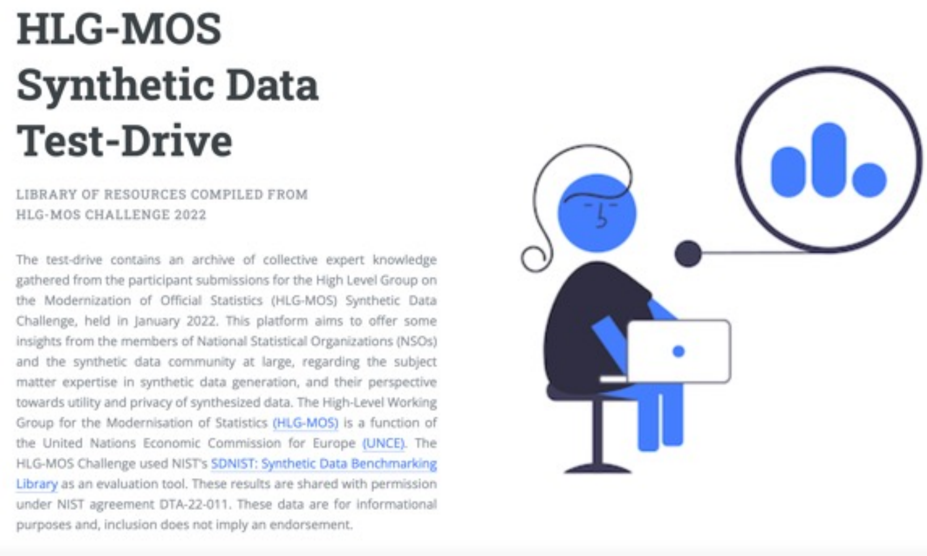
This page was updated 27 JAN 2023.

All information provided here is provisional and may change at any time. More detailed information will be released as the program progresses.



Google “Synthetic Data Test Drive”

- A lot of times it’s only us running things backstage, tucked behind an NDA, who get to see ***all*** of the new problems. Each team only sees their own.
- So we thought we might swap that around for you guys.



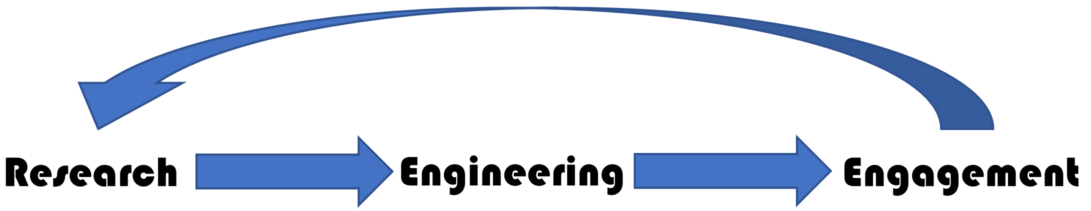
HLG-MOS Synthetic Data Test-Drive

LIBRARY OF RESOURCES COMPILED FROM HLG-MOS CHALLENGE 2022

The test-drive contains an archive of collective expert knowledge gathered from the participant submissions for the High Level Group on the Modernization of Official Statistics (HLG-MOS) Synthetic Data Challenge, held in January 2022. This platform aims to offer some insights from the members of National Statistical Organizations (NSOs) and the synthetic data community at large, regarding the subject matter expertise in synthetic data generation, and their perspective towards utility and privacy of synthesized data. The High-Level Working Group for the Modernisation of Statistics (HLG-MOS) is a function of the United Nations Economic Commission for Europe (UNCE). The HLG-MOS Challenge used NIST's SDNIST: Synthetic Data Benchmarking Library as an evaluation tool. These results are shared with permission under NIST agreement DTA-22-011. These data are for informational purposes and, inclusion does not imply an endorsement.

Introducing the Collaborative Research Cycle project:

- We want to try a challenge where we all work *together* to learn-- In this case, about data deidentification **Research**.

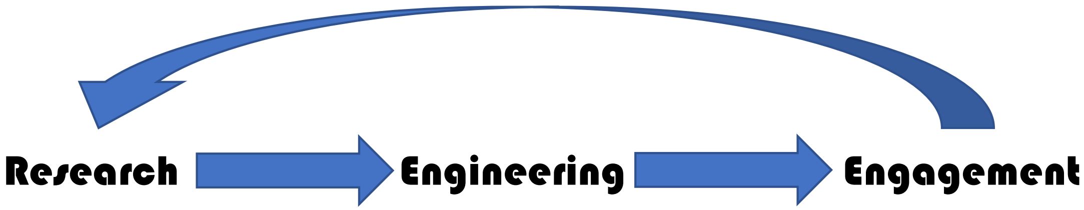


Research

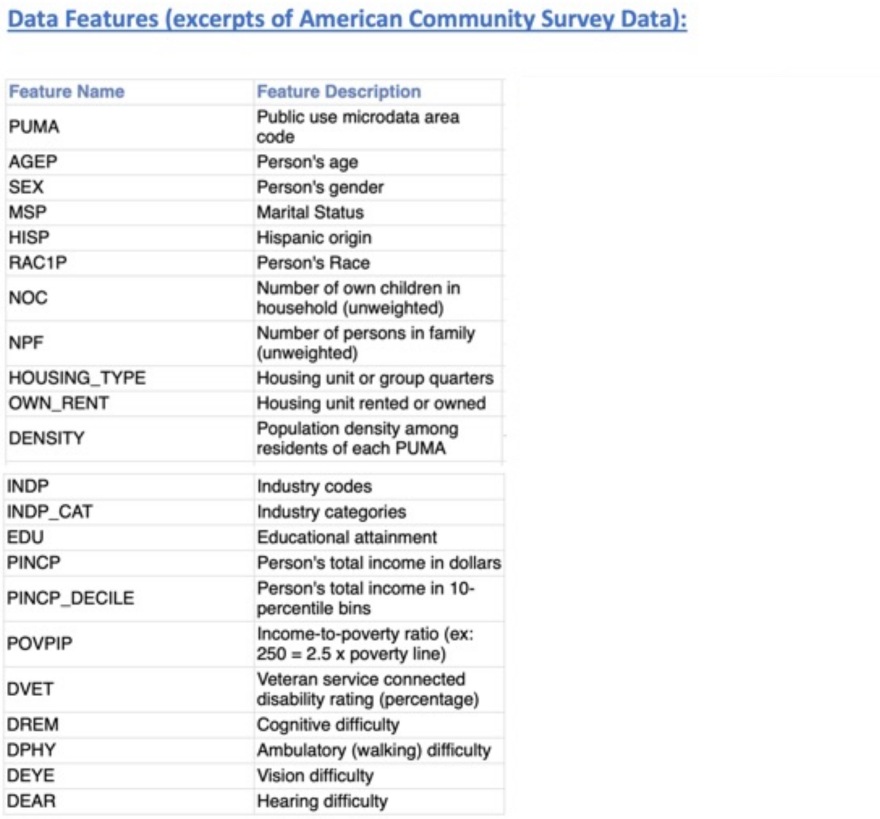
Engineering

Engagement

Introducing the Collaborative Research Cycle project:



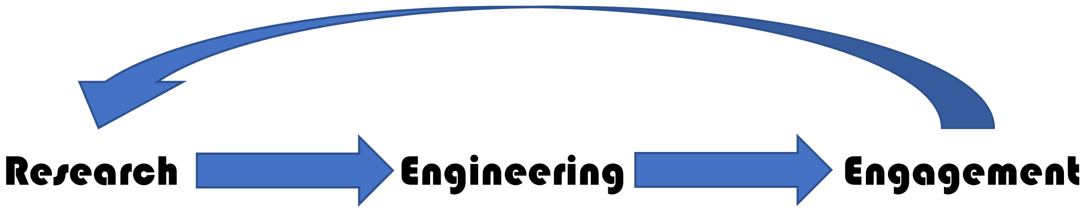
- We want to try a challenge where we all work *together* to learn-- In this case, about data deidentification algorithms.
- We took curated Diverse Community data excerpts from the ACS. Small enough to see what's happening, challenging enough to see the many interesting things that *can* happen.



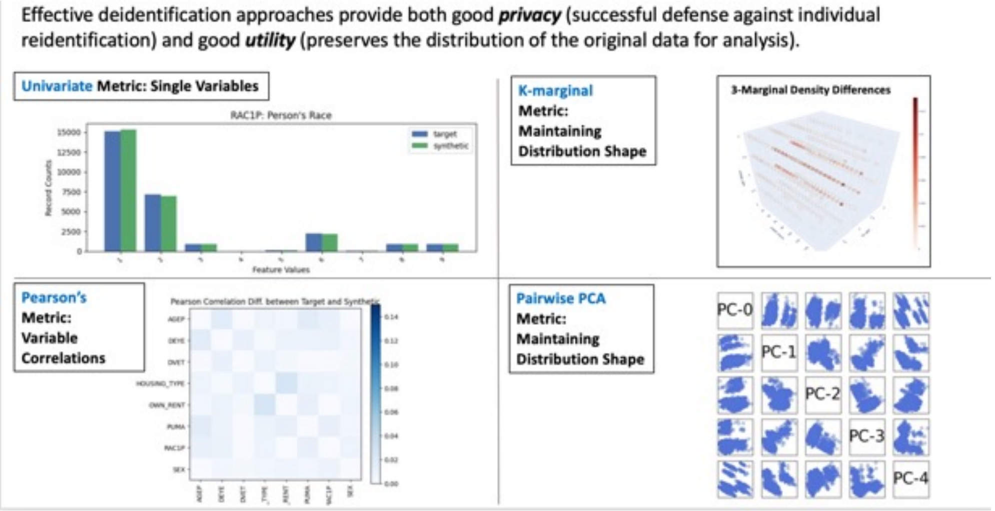
Data Features (excerpts of American Community Survey Data):

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

Introducing the Collaborative Research Cycle project:



- We want to try a challenge where we all work *together* to learn-- In this case, about data deidentification algorithms.
- We took curated Diverse Community data excerpts from the ACS. Small enough to see what's happening, challenging enough to see the many interesting things that *can* happen.
- We took metrics from the data experts themselves, the HLG-MOS Synthetic Data test drive.



Effective deidentification approaches provide both good ***privacy*** (successful defense against individual reidentification) and good ***utility*** (preserves the distribution of the original data for analysis).

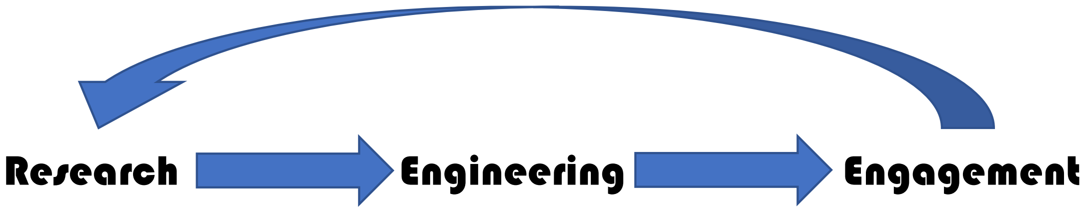
Univariate Metric: Single Variables

K-marginal Metric: Maintaining Distribution Shape

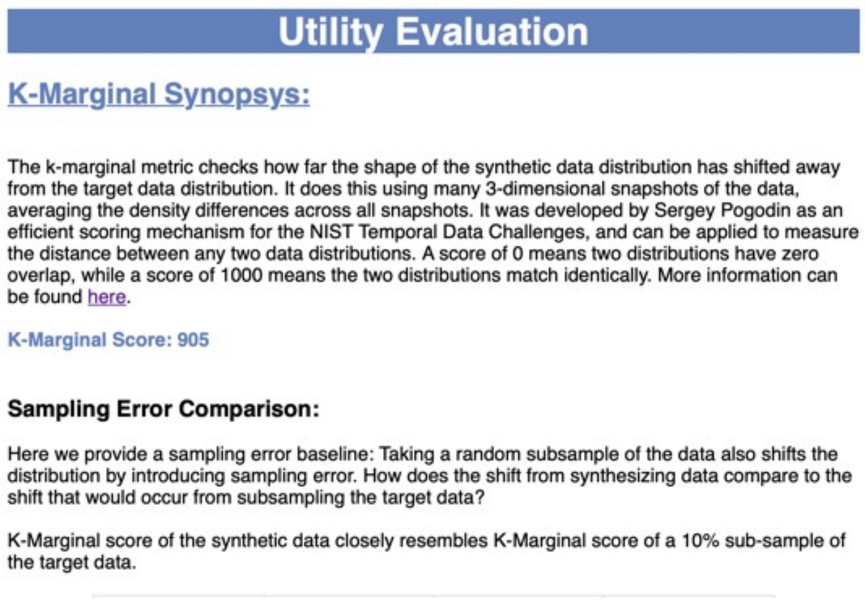
Pearson's Metric: Variable Correlations

Pairwise PCA Metric: Maintaining Distribution Shape

Introducing the Collaborative Research Cycle project:



- We want to try a challenge where we all work *together* to learn-- In this case, about data deidentification algorithms.
- We took curated Diverse Community data excerpts from the ACS. Small enough to see what's happening, challenging enough to see the many interesting things that *can* happen.
- We took metrics from the data experts themselves, the HLG-MOS Synthetic Data test drive.
- We want you to run your algorithm on our data, submit it, and we'll send you the evaluation results.
- Then we'll share with everyone. Review your deidentified data and compare it with others!
- And finally, tell us all what you find, in a tiny-paper workshop this fall.



Utility Evaluation

K-Marginal Synopsys:

The k-marginal metric checks how far the shape of the synthetic data distribution has shifted away from the target data distribution. It does this using many 3-dimensional snapshots of the data, averaging the density differences across all snapshots. It was developed by Sergey Pogodin 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 two distributions have zero overlap, while a score of 1000 means the two distributions match identically. More information can be found here.

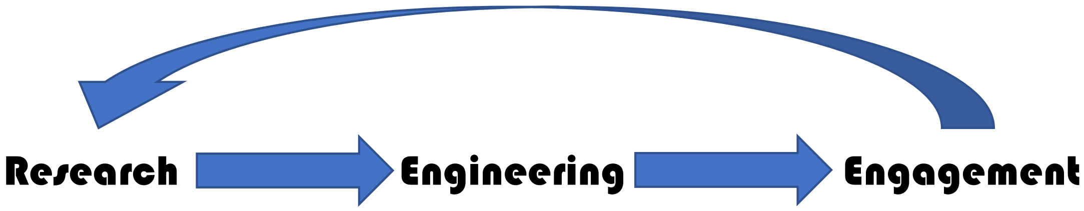
K-Marginal Score: 905

Sampling Error Comparison:

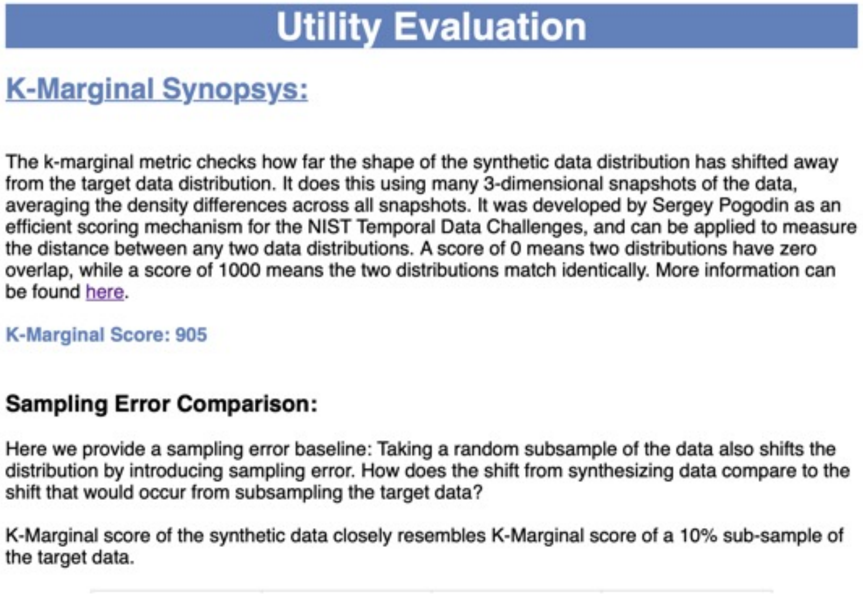
Here we provide a sampling error baseline: Taking a random subsample of the data also shifts the distribution by introducing sampling error. How does the shift from synthesizing data compare to the shift that would occur from subsampling the target data?

K-Marginal score of the synthetic data closely resembles K-Marginal score of a 10% sub-sample of the target data.

Introducing the Collaborative Research Cycle project:



- We’re kicking off today! Today is the first day you can register your team.
- ….but we got started a bit early, with our collaborators.
- To give you a better idea what all we’re on about, let’s show you what we have so far.



Utility Evaluation

K-Marginal Synopsys:

The k-marginal metric checks how far the shape of the synthetic data distribution has shifted away from the target data distribution. It does this using many 3-dimensional snapshots of the data, averaging the density differences across all snapshots. It was developed by Sergey Pogodin 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 two distributions have zero overlap, while a score of 1000 means the two distributions match identically. More information can be found here.

K-Marginal Score: 905

Sampling Error Comparison:

Here we provide a sampling error baseline: Taking a random subsample of the data also shifts the distribution by introducing sampling error. How does the shift from synthesizing data compare to the shift that would occur from subsampling the target data?

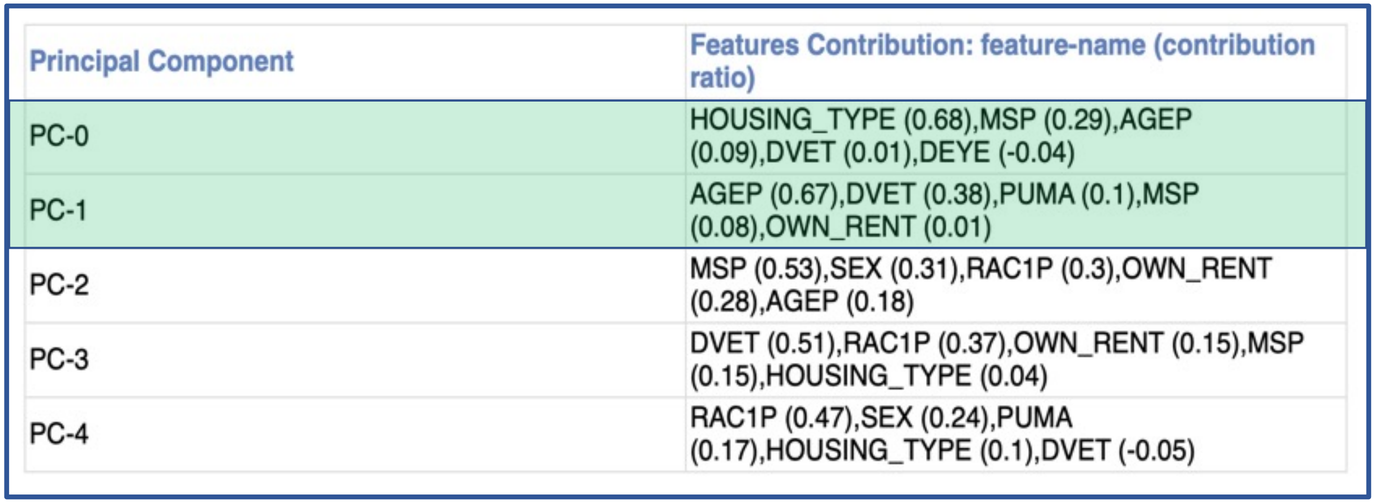
K-Marginal score of the synthetic data closely resembles K-Marginal score of a 10% sub-sample of the target data.

Higher Dimension Utility Metrics: The PCA Metric

The IPUMS International team proposed a new data visualization approach during the HLG-MOS Synthetic Data Test Drive.

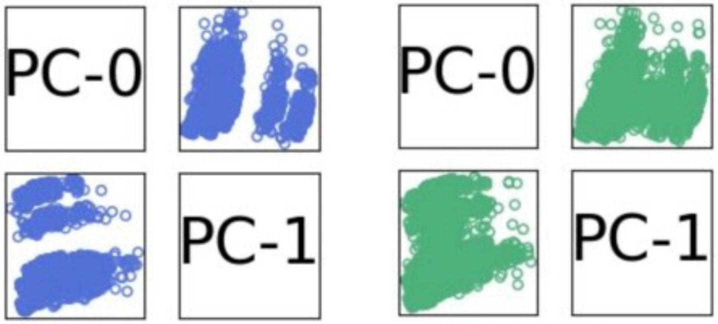
First, PCA is performed on the ground truth data, and then we use pairwise combinations of the top 5 components to get 2D scatterplots of the data.

This lets us see two dimensional impressions shapes formed by the high dimensional data distribution



Principal Component	Features Contribution: feature-name (contribution ratio)
PC-0	HOUSING_TYPE (0.68),MSP (0.29),AGEP (0.09),DVET (0.01),DEYE (-0.04)
PC-1	AGEP (0.67),DVET (0.38),PUMA (0.1),MSP (0.08),OWN_RENT (0.01)
PC-2	MSP (0.53),SEX (0.31),RAC1P (0.3),OWN_RENT (0.28),AGEP (0.18)
PC-3	DVET (0.51),RAC1P (0.37),OWN_RENT (0.15),MSP (0.15),HOUSING_TYPE (0.04)
PC-4	RAC1P (0.47),SEX (0.24),PUMA (0.17),HOUSING_TYPE (0.1),DVET (-0.05)

Original Data Distribution Shape



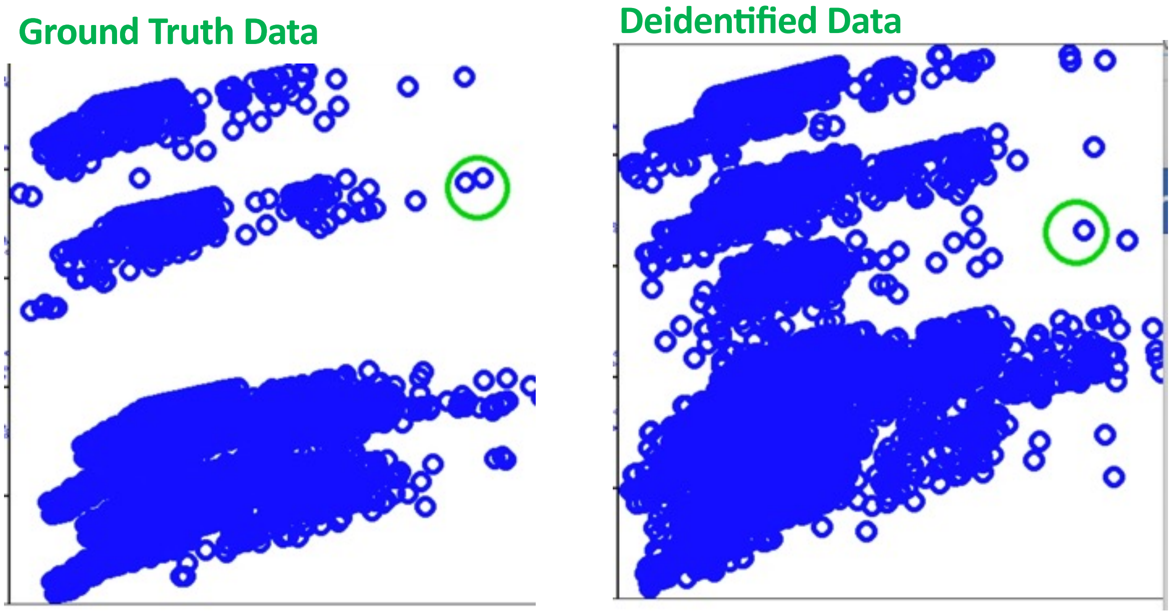
Deidentified Data Distribution Shape

How do Data Properties Impact Deidentification Techniques? GAN

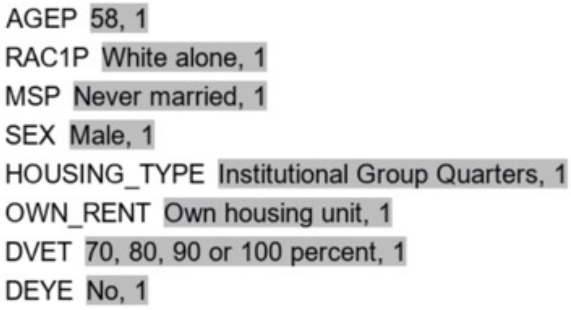
A GAN model trains a deep neural network to mimic the distribution of the input data, and then uses that network to produce a new data set that should keep the feature correlations from the original data.

We'll look at two GANS, neither differentially private. A basic GAN (the default setting from the SDVault library), and a Copula GAN (also from SDVault) a variant that should work better on tabular data.

PC0 [Housing Type, Marriage] x PC 1[Age, Veteran Status]:


Ground Truth Data

Deidentified Data



AGEP	58, 1
RAC1P	White alone, 1
MSP	Never married, 1
SEX	Male, 1
HOUSING_TYPE	Institutional Group Quarters, 1
OWN_RENT	Own housing unit, 1
DVET	70, 80, 90 or 100 percent, 1
DEYE	No, 1

What Happens?

The **SDV Copula GAN** graph here has a huge artifact that doesn't occur in the original data.

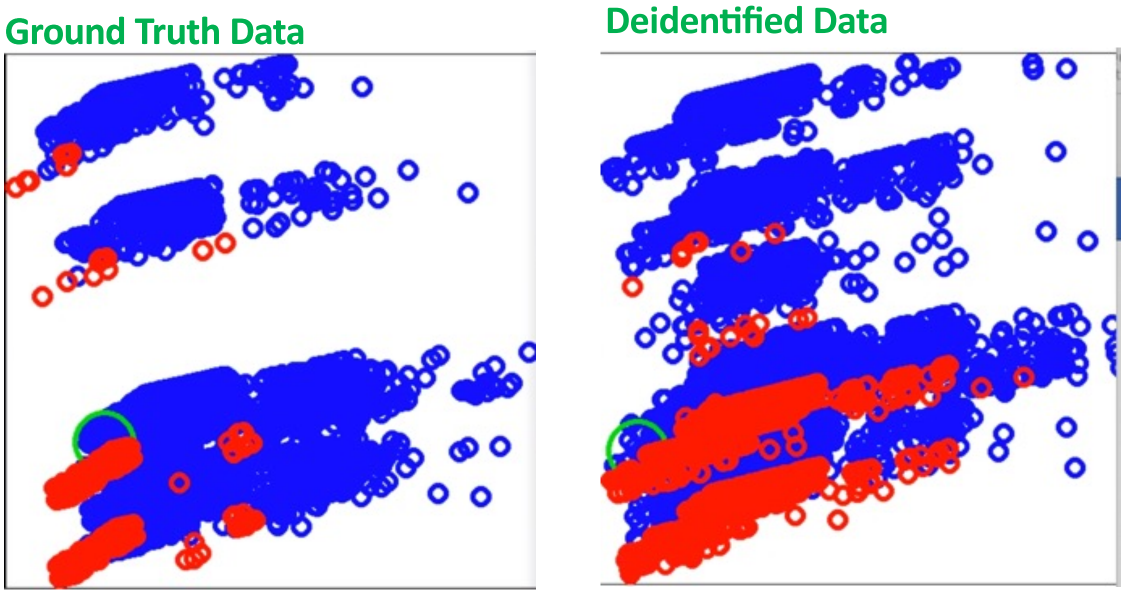
This is due to it missing consistency constraints– This person OWNs their own house, but they're also living in institutional group quarters.

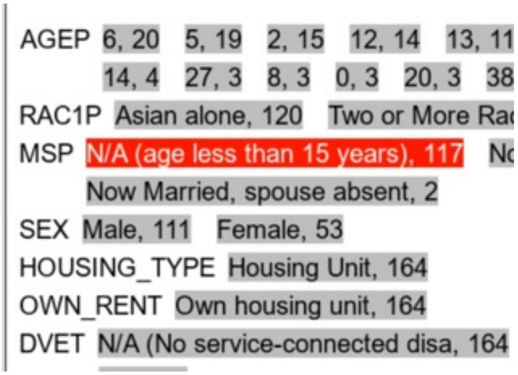
How do Data Properties Impact Deidentification Techniques? GAN

A GAN model trains a deep neural network to mimic the distribution of the input data, and then uses that network to produce a new data set that should keep the feature correlations from the original data.

We'll look at two GANS today, neither differentially private. A basic GAN (the default setting from the SDVault library), and a Copula GAN (also from SDVault) a variant that should work better on tabular data.

PC0 [Housing Type, Marriage] x PC 1[Age, Veteran Status]:




```
AGEP  6, 20   5, 19   2, 15   12, 14   13, 11
      14, 4   27, 3   8, 3   0, 3   20, 3   38
RAC1P  Asian alone, 120   Two or More Rac
MSP  N/A (age less than 15 years), 117   Nc
     Now Married, spouse absent, 2
SEX  Male, 111   Female, 53
HOUSING_TYPE  Housing Unit, 164
OWN_RENT  Own housing unit, 164
DVET  N/A (No service-connected disa, 164
```

What Happens?

The **SDV Copula GAN** graph here has a huge artifact that doesn't occur in the original data.

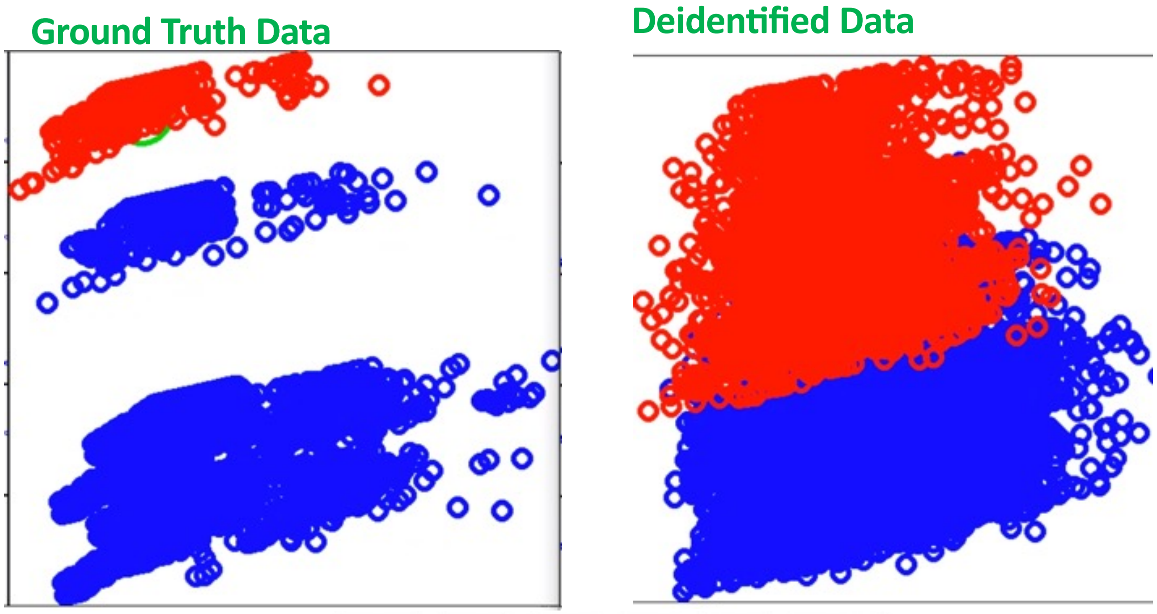
They are also generating Null marriage (under age 15) values for individuals of many age groups.

How do Data Properties Impact Deidentification Techniques? GAN

A GAN model trains a deep neural network to mimic the distribution of the input data, and then uses that network to produce a new data set that should keep the feature correlations from the original data.

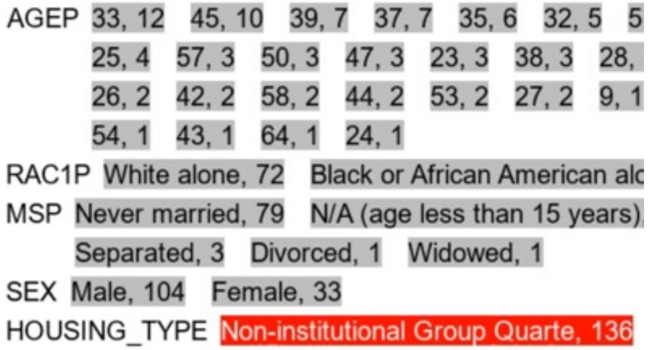
We'll look at two GANS today, neither differentially private. A basic GAN (the default setting from the SDVault library), and a Copula GAN (also from SDVault) a variant that should work better on tabular data.

PC0 [Housing Type, Marriage] x PC 1[Age, Veteran Status]:



Ground Truth Data

Deidentified Data



AGEP 33, 12 45, 10 39, 7 37, 7 35, 6 32, 5 5
25, 4 57, 3 50, 3 47, 3 23, 3 38, 3 28,
26, 2 42, 2 58, 2 44, 2 53, 2 27, 2 9, 1
54, 1 43, 1 64, 1 24, 1

RAC1P White alone, 72 Black or African American alc

MSP Never married, 79 N/A (age less than 15 years),
Separated, 3 Divorced, 1 Widowed, 1

SEX Male, 104 Female, 33

HOUSING_TYPE Non-institutional Group Quarte, 136

What Happens?

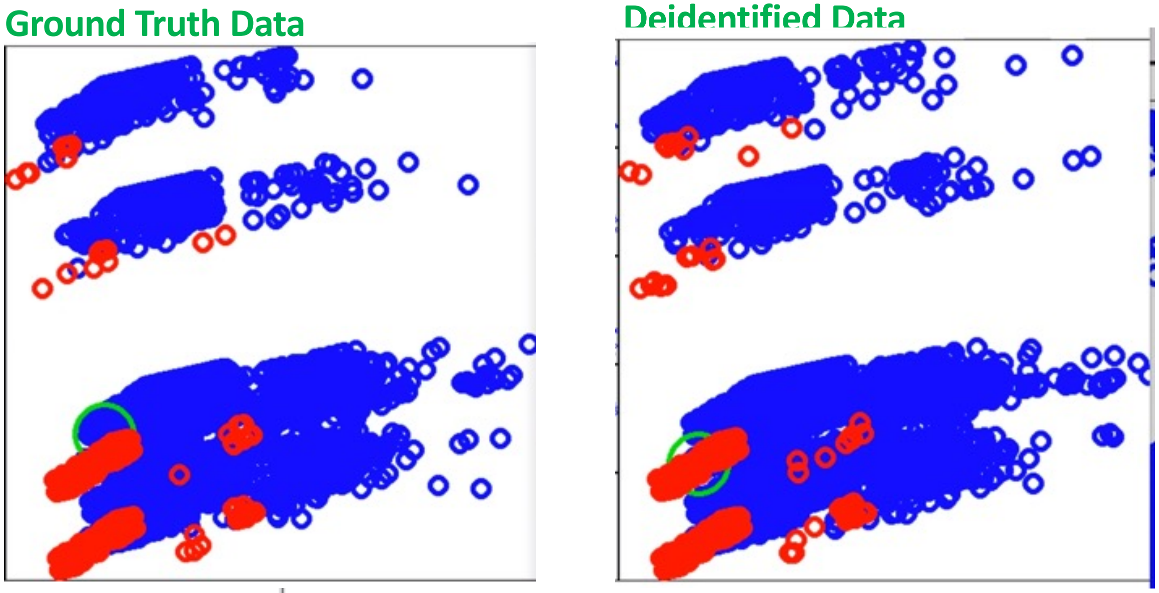
The **SDV** **Basic GAN** has very poor performance, for constraints and basic feature correlations

How do Data Properties Impact Deidentification Techniques? CART Model

A decision tree is a straightforward data modeling method for which allows you to predict the value of a variable based on the values of other variables. Different groups in the data are channeled down different paths in the tree, like building a flow chart.

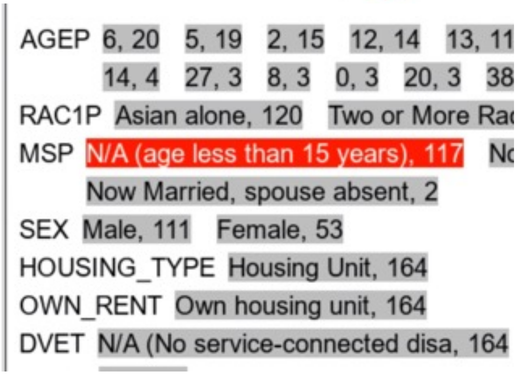
CART model synthesis doesn't provide formal privacy, but for larger feature sets it's unlikely to exactly reproduce real records.

PC0 [Housing Type, Marriage] x PC 1[Age, Veteran Status]:



Ground Truth Data

Deidentified Data



```
AGEP  6, 20   5, 19   2, 15   12, 14   13, 11
      14, 4   27, 3   8, 3   0, 3   20, 3   38
RAC1P  Asian alone, 120   Two or More Ra
MSP  N/A (age less than 15 years), 117   No
     Now Married, spouse absent, 2
SEX  Male, 111   Female, 53
HOUSING_TYPE  Housing Unit, 164
OWN_RENT  Own housing unit, 164
DVET  N/A (No service-connected disa, 164
```

What Happens?

The **CART-Tree** model data distribution looks very similar to the original data in their PCA graphs.

Feature relations are preserved and so are consistency constraints (in general).

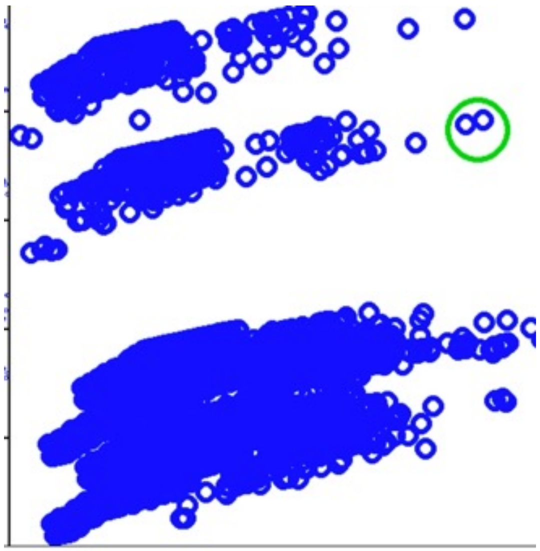
How do Data Properties Impact Deidentification Techniques? CART Model

A decision tree is a straightforward data modeling method for which allows you to predict the value of a variable based on the values of other variables. Different groups in the data are channeled down different paths in the tree, like building a flow chart.

CART model synthesis doesn't provide formal privacy, but for larger feature sets it's unlikely to exactly reproduce real records.

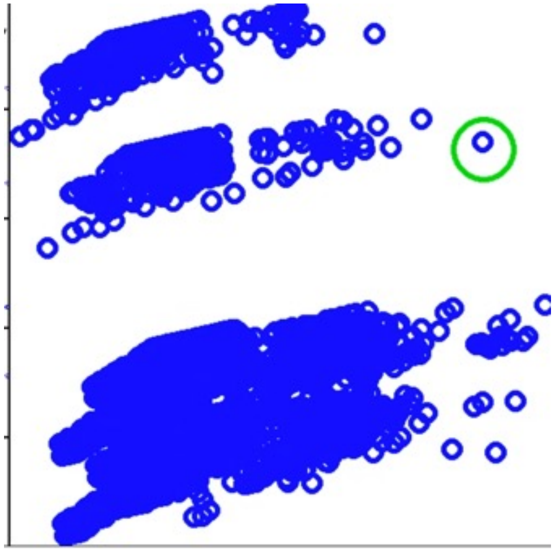
PC0 [Housing Type, Marriage] x PC 1[Age, Veteran Status]:

Ground Truth Data



AGEP 70, 1 56, 1
RAC1P Black or African American alon, 2
MSP Never married, 2
SEX Male, 2
HOUSING_TYPE Institutional Group Quarters, 2
OWN_RENT Group quarters, 2
DVET 70, 80, 90 or 100 percent, 2
DEYE Yes, 2

Deidentified Data



AGEP 68, 1
RAC1P White alone, 1
MSP Now Married, spouse absent, 1
SEX Male, 1
HOUSING_TYPE Institutional Group Quarters, 1
OWN_RENT Group quarters, 1
DVET 70, 80, 90 or 100 percent, 1
DEYE Yes, 1

What Happens?

The **CART-Tree** model data distribution looks very similar to the original data in their PCA graphs.

But notice that the individuals in the CART model graph are actually different than the individuals in the ground truth data.

MST

- Data is generated from a 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.
- Winner of the 2018 NIST Synthetic Data Challenge
- Generated with default settings, from OpenDP library's version of the Minuteman solution from the NIST Development Contest.
- Differentially Private, Epsilon 10
- Feature set: AGEP, SEX, MSP, RAC1P, HOUSING_TYPE, OWN_RENT, EDU, PINCP_DECILE, DVET, DEYE

Ryan McKenna, Gerome Miklau, Daniel Sheldon
[University of Massachusetts, Amherst, MA]

OpenDP

SmartNoise SDK: Tools on Tabular Data

WINNING THE NIST CONTEST: A SCALABLE AND GENERAL APPROACH TO DIFFERENTIALLY PRIVATE SYNTHETIC DATA

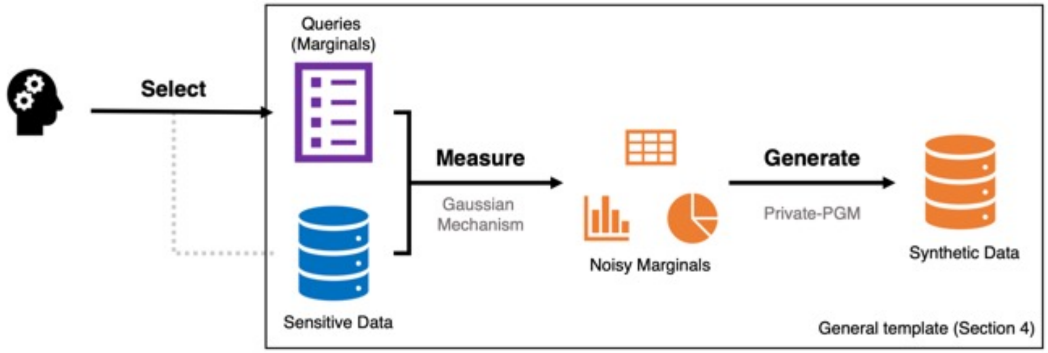
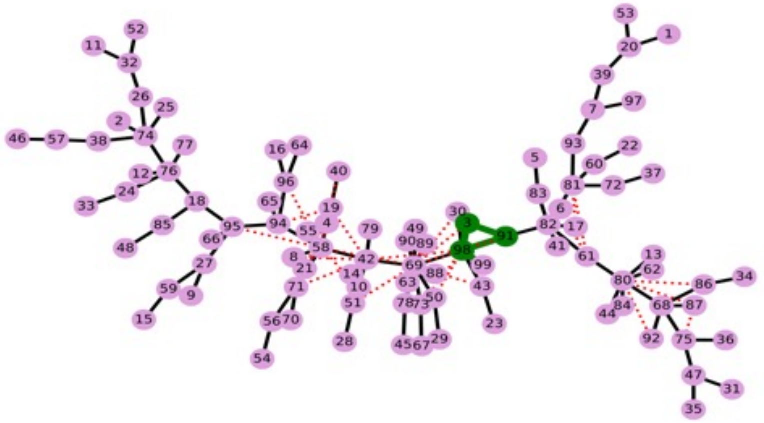
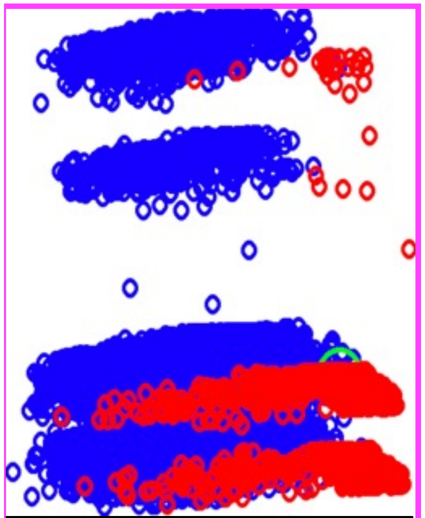
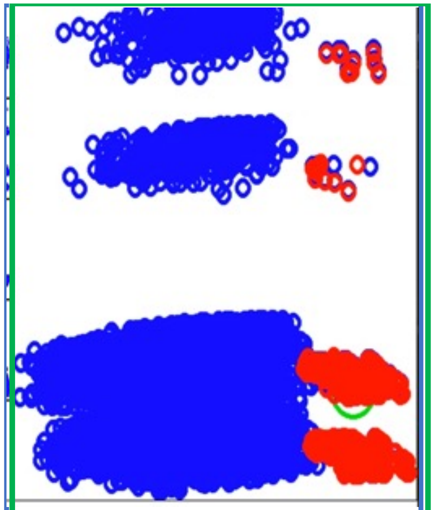


Figure 1. A general template for differentially private synthetic data generation. First, a collection of marginal queries is **selected**, either manually (e.g., by a domain expert) or automatically by an algorithm. Second, the Gaussian mechanism is used to **measure** those marginals while preserving differential privacy. Finally, `Private-PGM` is used to post-process the noisy marginals and **generate** a synthetic dataset that respects them.





OpenDP: MST



Target: Unmarried Child

ODP-MWEM

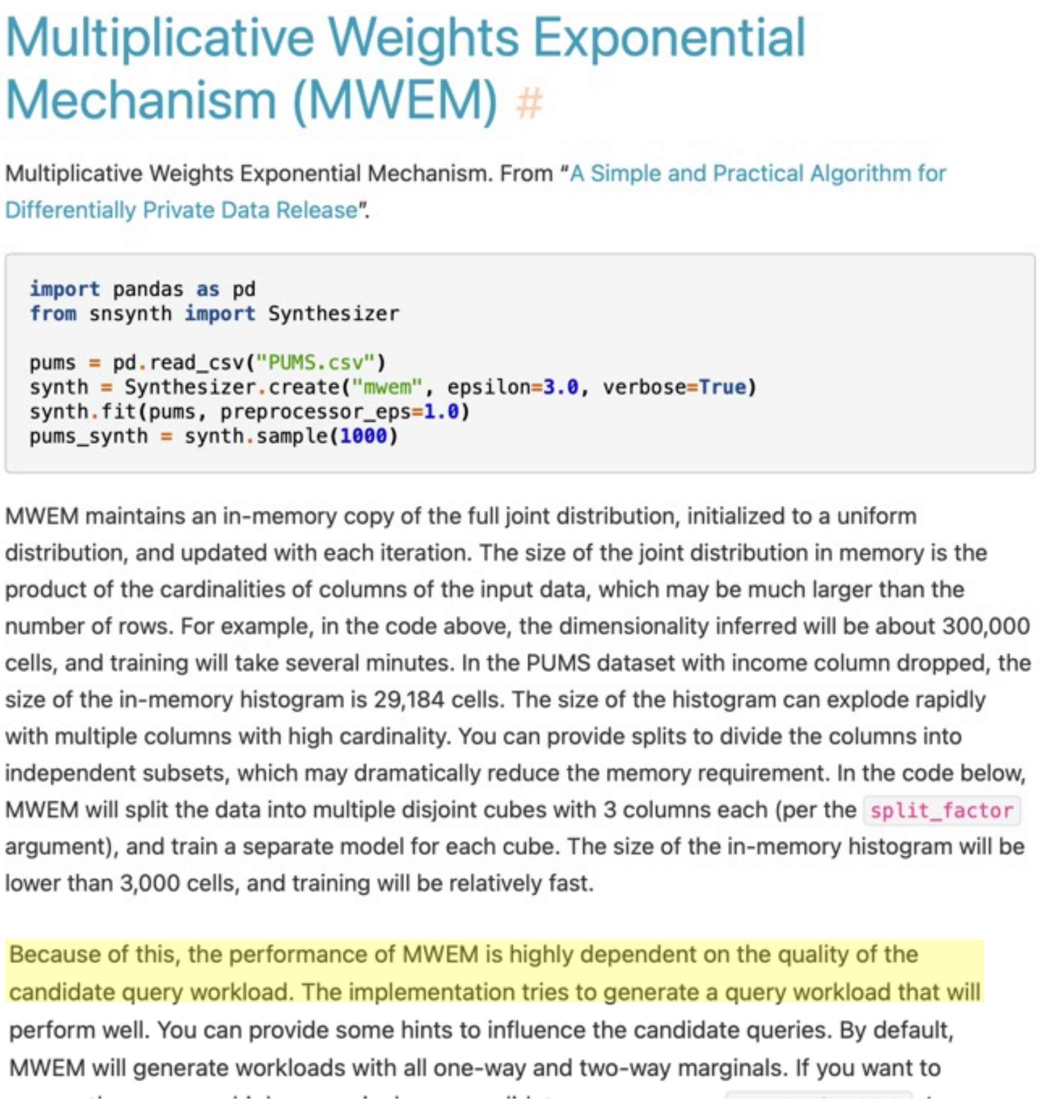
- Algorithm initializes synthetic data with random values and then iteratively refines its distribution to mimic noisy query results on ground truth data.
- Generated using the OpenDP library, with improvements to efficiency and query set!
- Differentially Private, Epsilon 10
- Feature set: AGEP, SEX, MSP, RAC1P, HOUSING_TYPE, OWN_RENT, EDU, PINCP_DECILE, DVET

Moritz Hardt [IBM Almaden Research]
Katrina Ligett [Caltech]
Frank McSherry [Microsoft Research]

Unnamed Open-DP Implementor [?]

OpenDP

SmartNoise SDK: Tools on Tabular Data



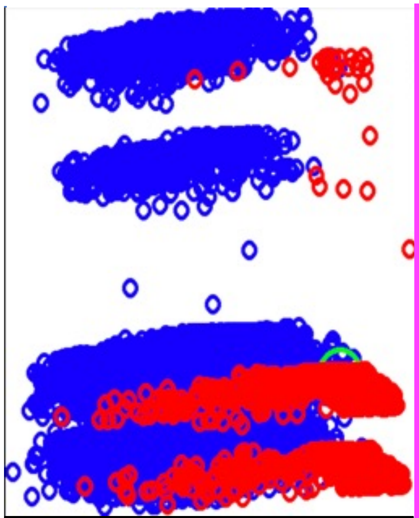
Multiplicative Weights Exponential Mechanism (MWEM)

Multiplicative Weights Exponential Mechanism. From "A Simple and Practical Algorithm for Differentially Private Data Release".

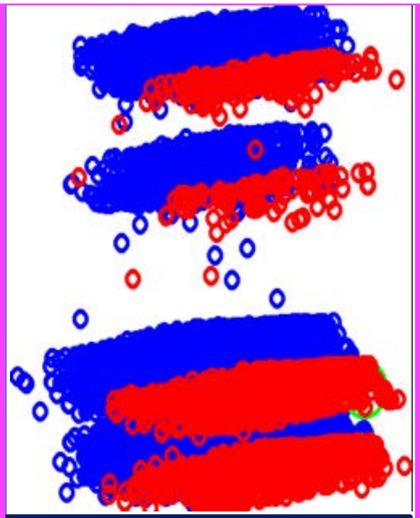
```
import pandas as pd
from snsynth import Synthesizer

pums = pd.read_csv("PUMS.csv")
synth = Synthesizer.create("mwem", epsilon=3.0, verbose=True)
synth.fit(pums, preprocessor_eps=1.0)
pums_synth = synth.sample(1000)
```

MWEM maintains an in-memory copy of the full joint distribution, initialized to a uniform distribution, and updated with each iteration. The size of the joint distribution in memory is the product of the cardinalities of columns of the input data, which may be much larger than the number of rows. For example, in the code above, the dimensionality inferred will be about 300,000 cells, and training will take several minutes. In the PUMS dataset with income column dropped, the size of the in-memory histogram is 29,184 cells. The size of the histogram can explode rapidly with multiple columns with high cardinality. You can provide splits to divide the columns into independent subsets, which may dramatically reduce the memory requirement. In the code below, MWEM will split the data into multiple disjoint cubes with 3 columns each (per the `split_factor` argument), and train a separate model for each cube. The size of the in-memory histogram will be lower than 3,000 cells, and training will be relatively fast.

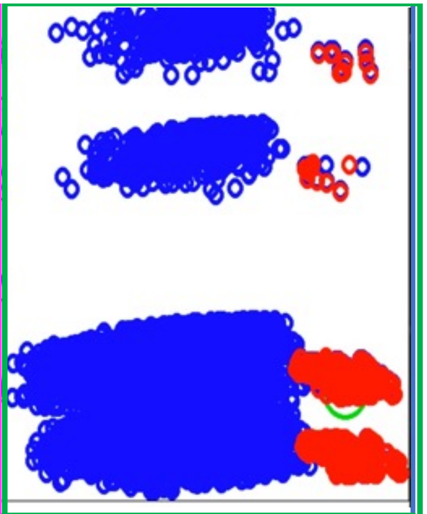
Because of this, the performance of MWEM is highly dependent on the quality of the candidate query workload. The implementation tries to generate a query workload that will perform well. You can provide some hints to influence the candidate queries. By default, MWEM will generate workloads with all one-way and two-way marginals. If you want to



OpenDP: MST



OpenDP: MWEM



Target: Unmarried Child

PATE-CTGAN

- Conditional tabular GAN using Private Aggregation of Teacher Ensembles.
- Generated from OpenDP library
- Differentially Private, Epsilon 10
- Feature set: AGEP, SEX, MSP, RAC1P, HOUSING_TYPE, OWN_RENT, EDU, PINCP_DECILE, DVET, DEYE

PATE-GAN:
James Jordon [University of Oxford], Jinsung Yoon∗ [UCLA], Mihaela van der Schaar [University of Cambridge, UCLA, Alan Turing Institute]

CT-GAN:
Lei Xu [MIT], Alfredo Cuesta-Infante [Universidad Rey Juan Carlos Móstoles], Maria Skoularidou [University of Cambridge], Kalyan Veeramachaneni [MIT]

OpenDP

License MIT

SmartNoise SDK: Tools on Tabular Data

Modeling Tabular Data using Conditional GAN

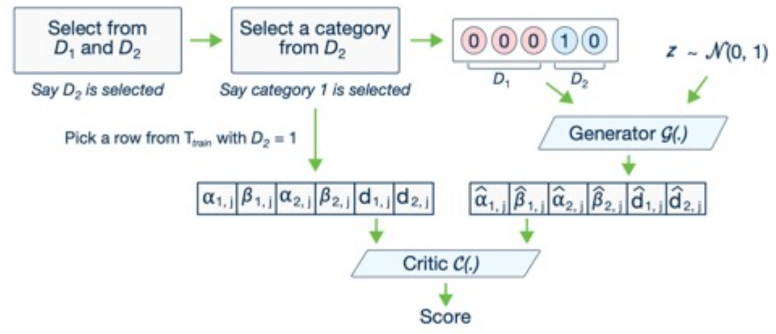
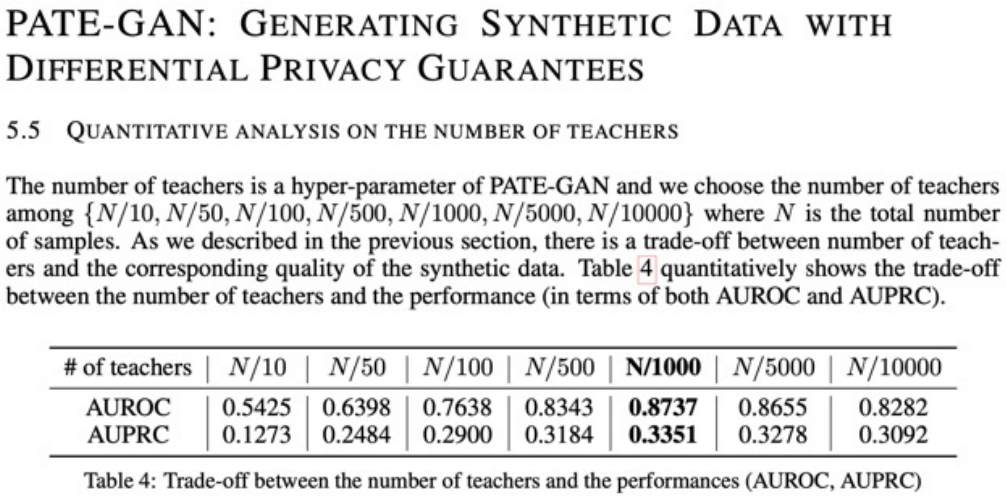

Select from D_1 and D_2
Say D_2 is selected
Select a category from D_2
Say category 1 is selected
0 0 0 1 0
D_1
D_2
$z \sim \mathcal{N}(0, 1)$
Generator $\mathcal{G}(.)$
Pick a row from T_{train} with $D_2 = 1$
$\alpha_{1,j}$ $\beta_{1,j}$ $\alpha_{2,j}$ $\beta_{2,j}$ $d_{1,j}$ $d_{2,j}$
$\hat{\alpha}_{1,j}$ $\hat{\beta}_{1,j}$ $\hat{\alpha}_{2,j}$ $\hat{\beta}_{2,j}$ $\hat{d}_{1,j}$ $\hat{d}_{2,j}$
Critic $\mathcal{C}(.)$
Score

Figure 2: CTGAN model. The conditional generator can generate synthetic rows conditioned on one of the discrete columns. With training-by-sampling, the *cond* and training data are sampled according to the log-frequency of each category, thus CTGAN can evenly explore all possible discrete values.



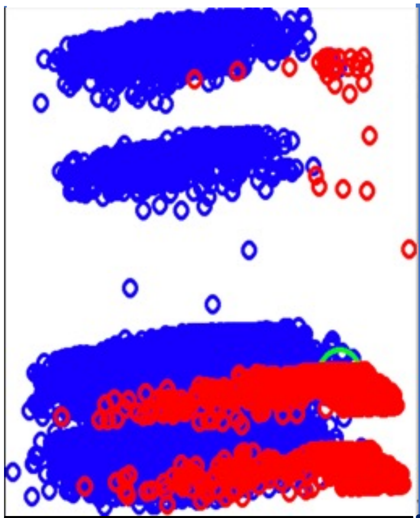
PATE-GAN: Generating Synthetic Data with Differential Privacy Guarantees

5.5 Quantitative analysis on the number of teachers

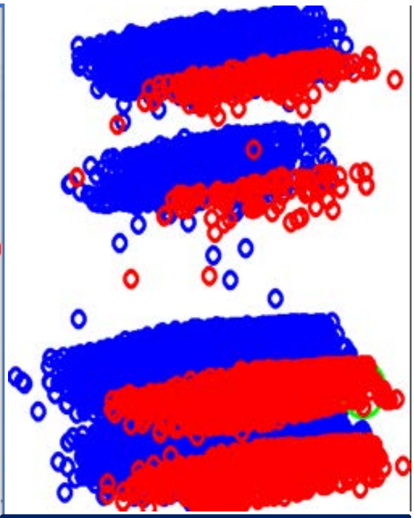
The number of teachers is a hyper-parameter of PATE-GAN and we choose the number of teachers among $\{N/10, N/50, N/100, N/500, N/1000, N/5000, N/10000\}$ where N is the total number of samples. As we described in the previous section, there is a trade-off between number of teachers and the corresponding quality of the synthetic data. Table 4 quantitatively shows the trade-off between the number of teachers and the performance (in terms of both AUROC and AUPRC).

# of teachers	$N/10$	$N/50$	$N/100$	$N/500$	**N/1000**	$N/5000$	$N/10000$
AUROC	0.5425	0.6398	0.7638	0.8343	**0.8737**	0.8655	0.8282
AUPRC	0.1273	0.2484	0.2900	0.3184	**0.3351**	0.3278	0.3092

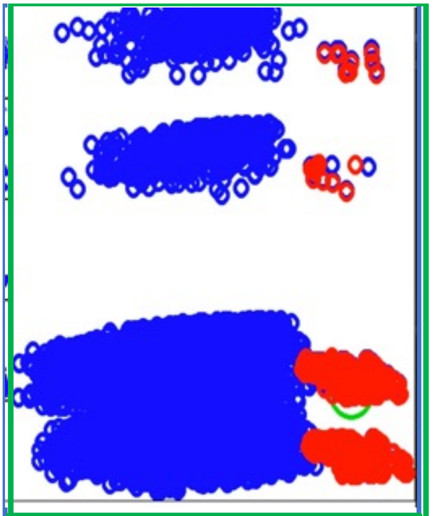
Table 4: Trade-off between the number of teachers and the performances (AUROC, AUPRC)



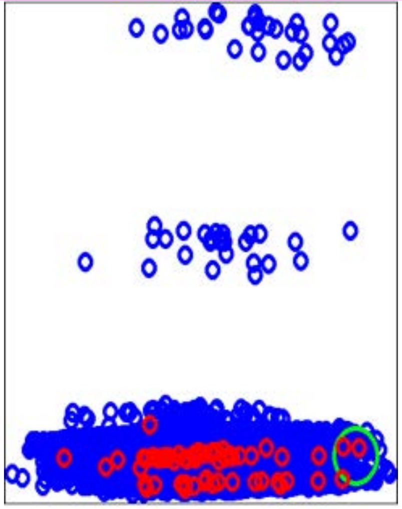
OpenDP: MST



OpenDP: MWEM



Target: Unmarried Child



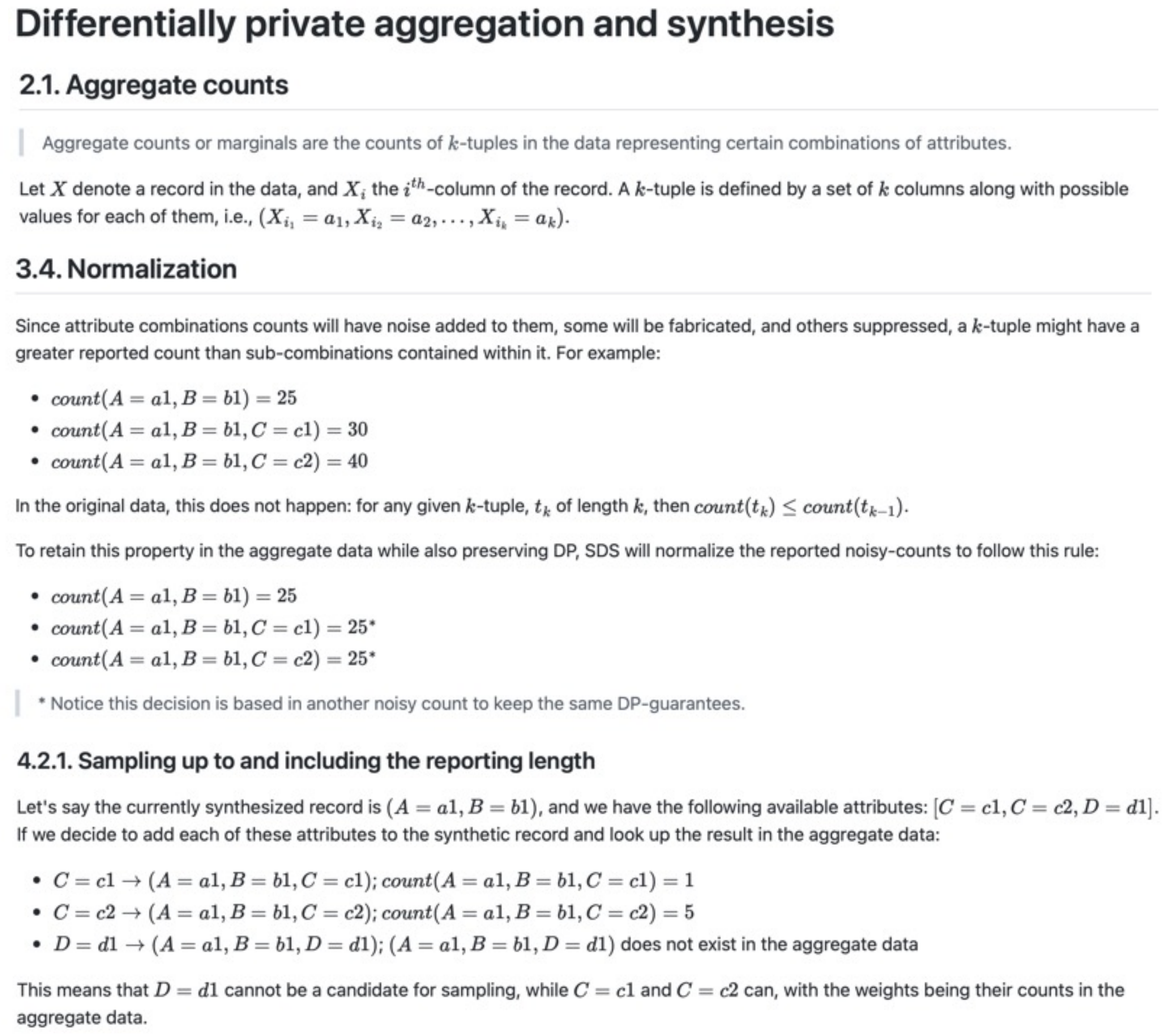
OpenDP: PATE-CTGAN

PAC-Synth

- Algorithm creates an internally consistent set of overlapping marginal counts (A, B, A∩B), and then samples new records from that distribution while maintaining consistency. (records with null values were removed)
- Generated from OpenDP library
- Differentially Private, Epsilon 10
- Feature set: AGEP, SEX, MSP, RAC1P, HOUSING_TYPE, OWN_RENT, EDU, PINCP_DECILE, DVET, DEYE

Sivakanth Gopi, Sepideh Mahabadi, and Sergey Yekhanin [Microsoft Research]

OpenDP SmartNoise SDK: Tools on Tabular Data



Differentially private aggregation and synthesis

2.1. Aggregate counts

> Aggregate counts or marginals are the counts of k-tuples in the data representing certain combinations of attributes.

Let X denote a record in the data, and X_i the i^{th}-column of the record. A k-tuple is defined by a set of k columns along with possible values for each of them, i.e., $(X_{i_1} = a_1, X_{i_2} = a_2, \ldots, X_{i_k} = a_k)$.

3.4. Normalization

Since attribute combinations counts will have noise added to them, some will be fabricated, and others suppressed, a k-tuple might have a greater reported count than sub-combinations contained within it. For example:

- $count(A = a1, B = b1) = 25$
- $count(A = a1, B = b1, C = c1) = 30$
- $count(A = a1, B = b1, C = c2) = 40$

In the original data, this does not happen: for any given k-tuple, t_k of length k, then $count(t_k) \leq count(t_{k-1})$.

To retain this property in the aggregate data while also preserving DP, SDS will normalize the reported noisy-counts to follow this rule:

- $count(A = a1, B = b1) = 25$
- $count(A = a1, B = b1, C = c1) = 25^*$
- $count(A = a1, B = b1, C = c2) = 25^*$

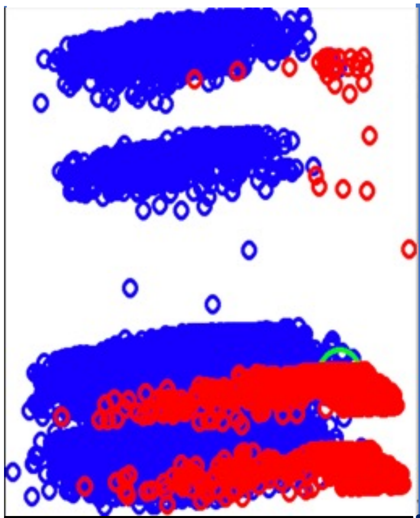
> * Notice this decision is based in another noisy count to keep the same DP-guarantees.

4.2.1. Sampling up to and including the reporting length

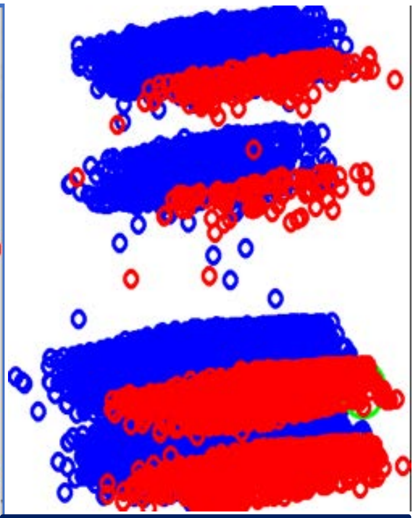
Let's say the currently synthesized record is $(A = a1, B = b1)$, and we have the following available attributes: $[C = c1, C = c2, D = d1]$. If we decide to add each of these attributes to the synthetic record and look up the result in the aggregate data:

- $C = c1 \rightarrow (A = a1, B = b1, C = c1)$; $count(A = a1, B = b1, C = c1) = 1$
- $C = c2 \rightarrow (A = a1, B = b1, C = c2)$; $count(A = a1, B = b1, C = c2) = 5$
- $D = d1 \rightarrow (A = a1, B = b1, D = d1)$; $(A = a1, B = b1, D = d1)$ does not exist in the aggregate data

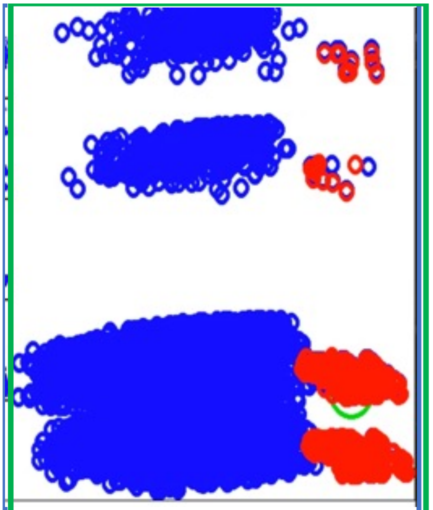
This means that $D = d1$ cannot be a candidate for sampling, while $C = c1$ and $C = c2$ can, with the weights being their counts in the aggregate data.



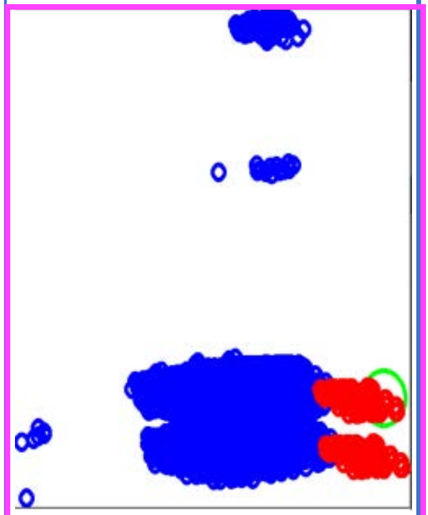
OpenDP: MST



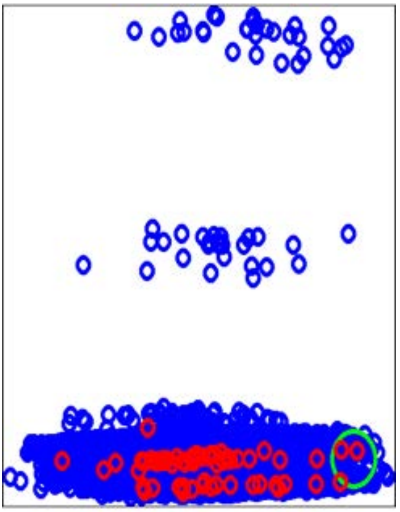
OpenDP: MWEM



Target: Unmarried Child



OpenDP: PAC-SYNTH



OpenDP: PATE-CTGAN

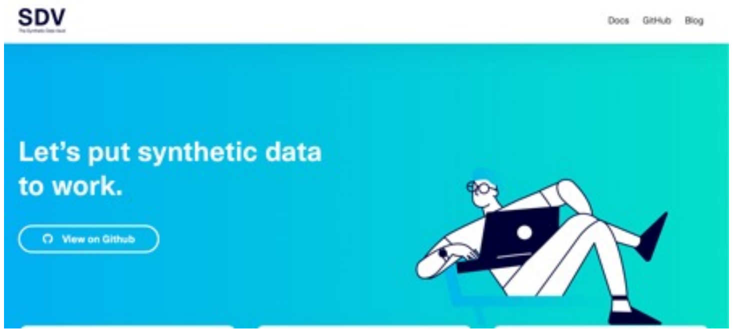
SDV-Basic & SDV-Copula

- Basic GAN (FAST_ML) and Copula GAN synthetic data generation.
- Data was generated using the Synthetic Data Vault open source python library: https://sdv.dev/SDV/
- Is not differentially private
- Is popular: 682K downloads
- Feature set: Full 22 features from the Diverse Community Benchmarks

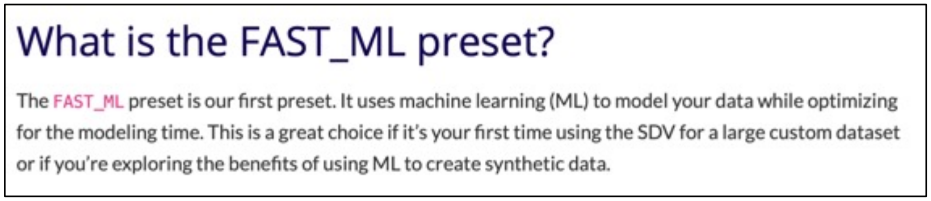
The Synthetic Data Vault Project was first created at MIT's Data to AI Lab in 2016. After 4 years of research and traction with enterprise, we created DataCebo in 2020 with the goal of growing the project. Today, DataCebo is the proud developer of SDV, the largest ecosystem for synthetic data generation & evaluation.

682 Thousand downloads

datacebo
SDV
The Synthetic Data Vault



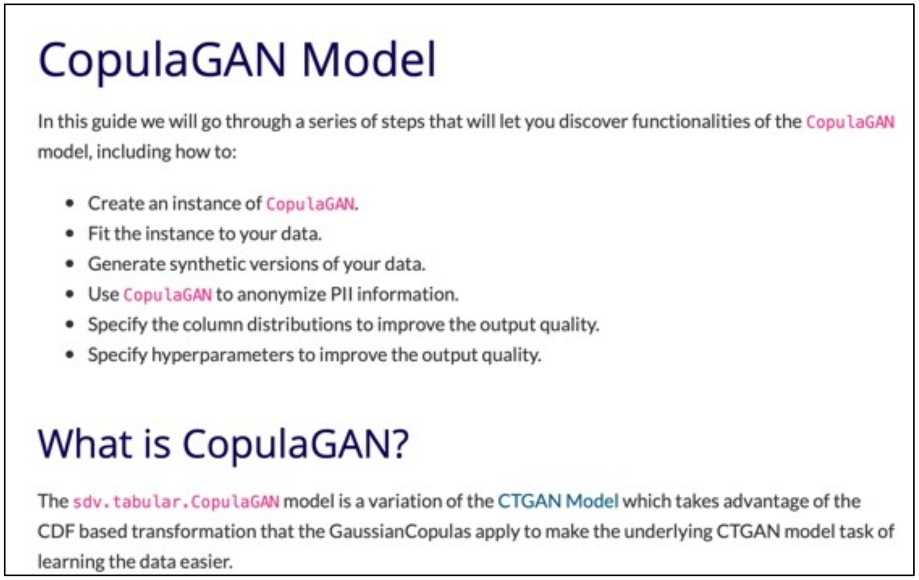
Basic Setting:



What is the FAST_ML preset?

The FAST_ML preset is our first preset. It uses machine learning (ML) to model your data while optimizing for the modeling time. This is a great choice if it's your first time using the SDV for a large custom dataset or if you're exploring the benefits of using ML to create synthetic data.

Copula Setting:



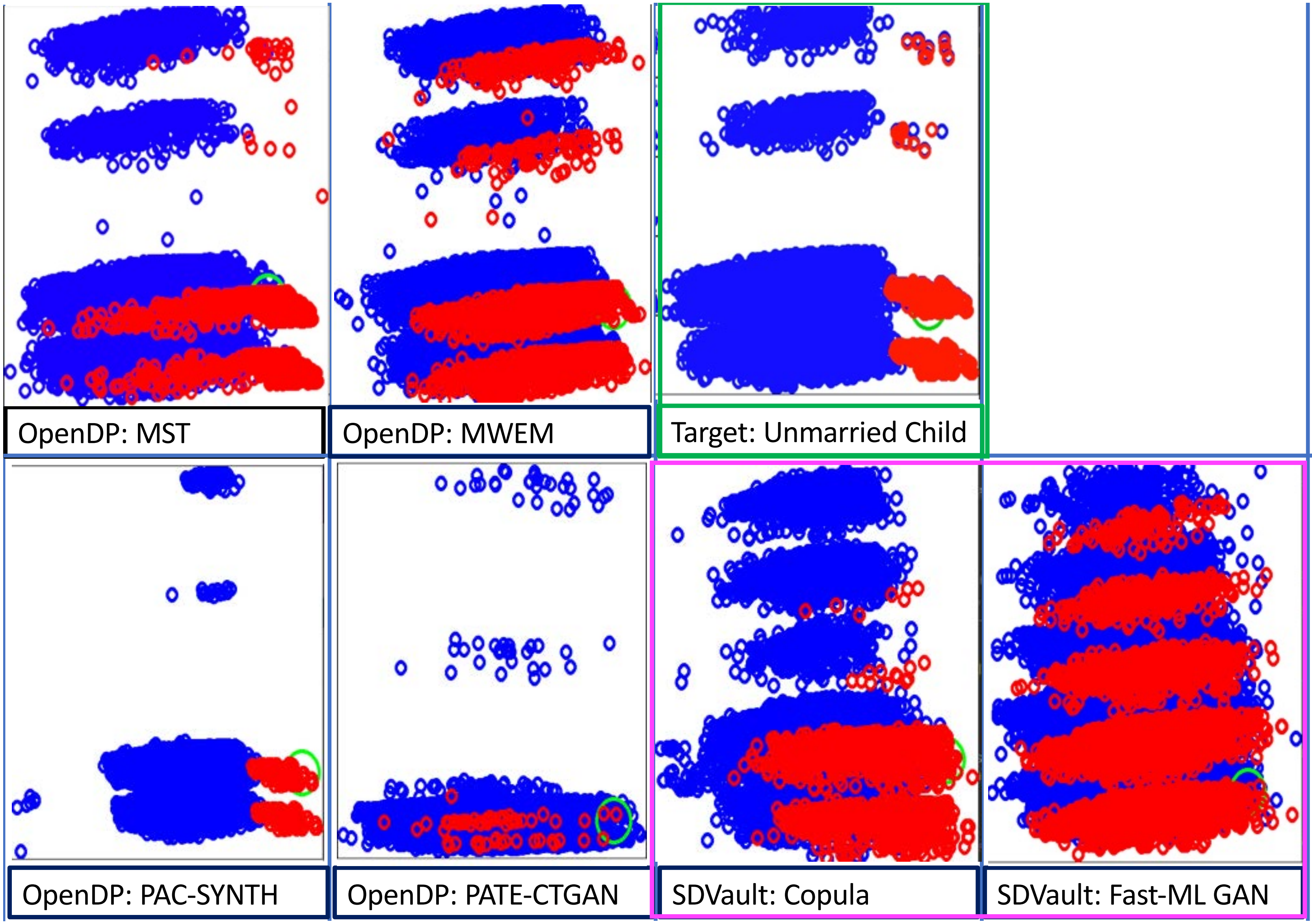
CopulaGAN Model

In this guide we will go through a series of steps that will let you discover functionalities of the CopulaGAN model, including how to:

- Create an instance of CopulaGAN.
- Fit the instance to your data.
- Generate synthetic versions of your data.
- Use CopulaGAN to anonymize PII information.
- Specify the column distributions to improve the output quality.
- Specify hyperparameters to improve the output quality.

What is CopulaGAN?

The sdv.tabular.CopulaGAN model is a variation of the CTGAN Model which takes advantage of the CDF based transformation that the GaussianCopulas apply to make the underlying CTGAN model task of learning the data easier.



OpenDP: MST

OpenDP: MWEM

Target: Unmarried Child

OpenDP: PAC-SYNTH

OpenDP: PATE-CTGAN

SDVault: Copula

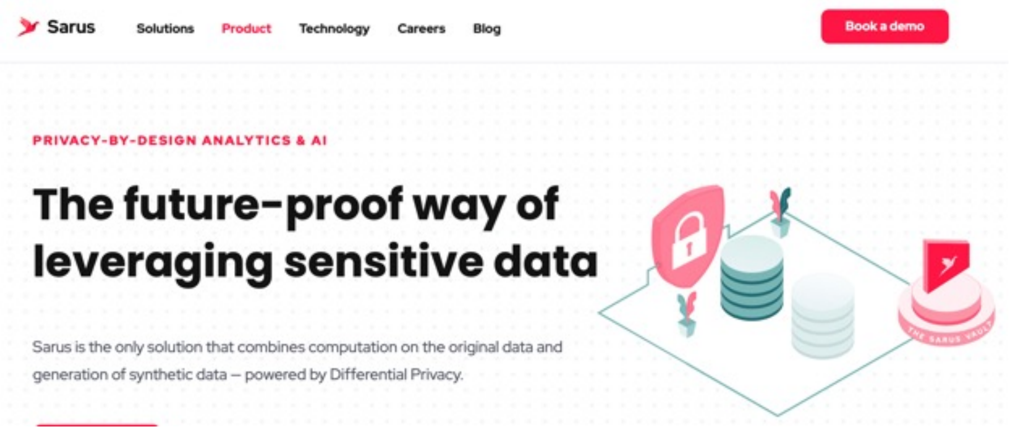
SDVault: Fast-ML GAN

Sarus

- Data is generated using a Transformer network (similar to LLM). New records are generated iteratively, one column at a time.
- Data Submitted to the CRC by Nicolas Grislain and Luca Canale [Sarus].
- Is differentially private, Epsilon = 10
- Feature set: AGEP, SEX, MSP, RAC1P, HOUSING_TYPE, OWN_RENT, EDU, PINCP_DECILE, DVET, DEYE

Generative Modeling of Complex Data
Luca Canale, Nicolas Grislain, Grégoire Lothe, Johan Leduc [Sarus Technology]

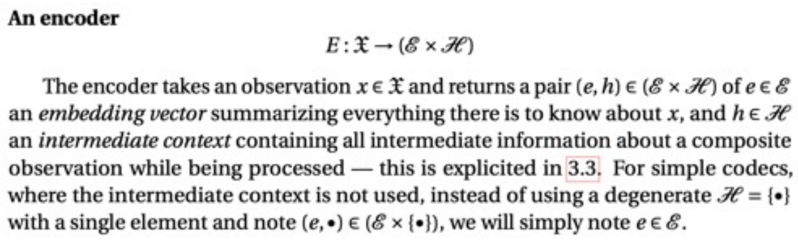
Sarus



PRIVACY-BY-DESIGN ANALYTICS & AI

The future-proof way of leveraging sensitive data

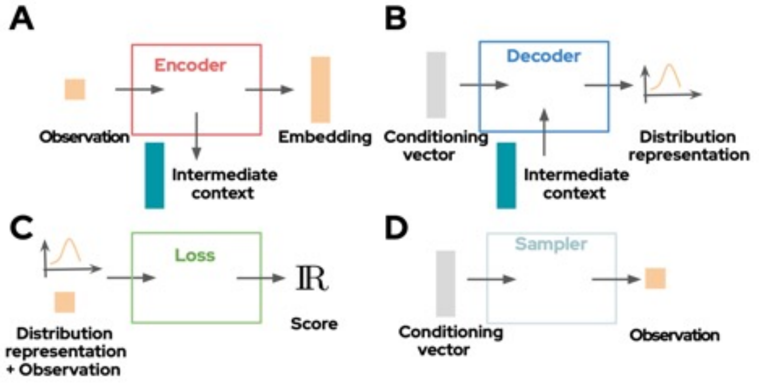
Sarus is the only solution that combines computation on the original data and generation of synthetic data – powered by Differential Privacy.

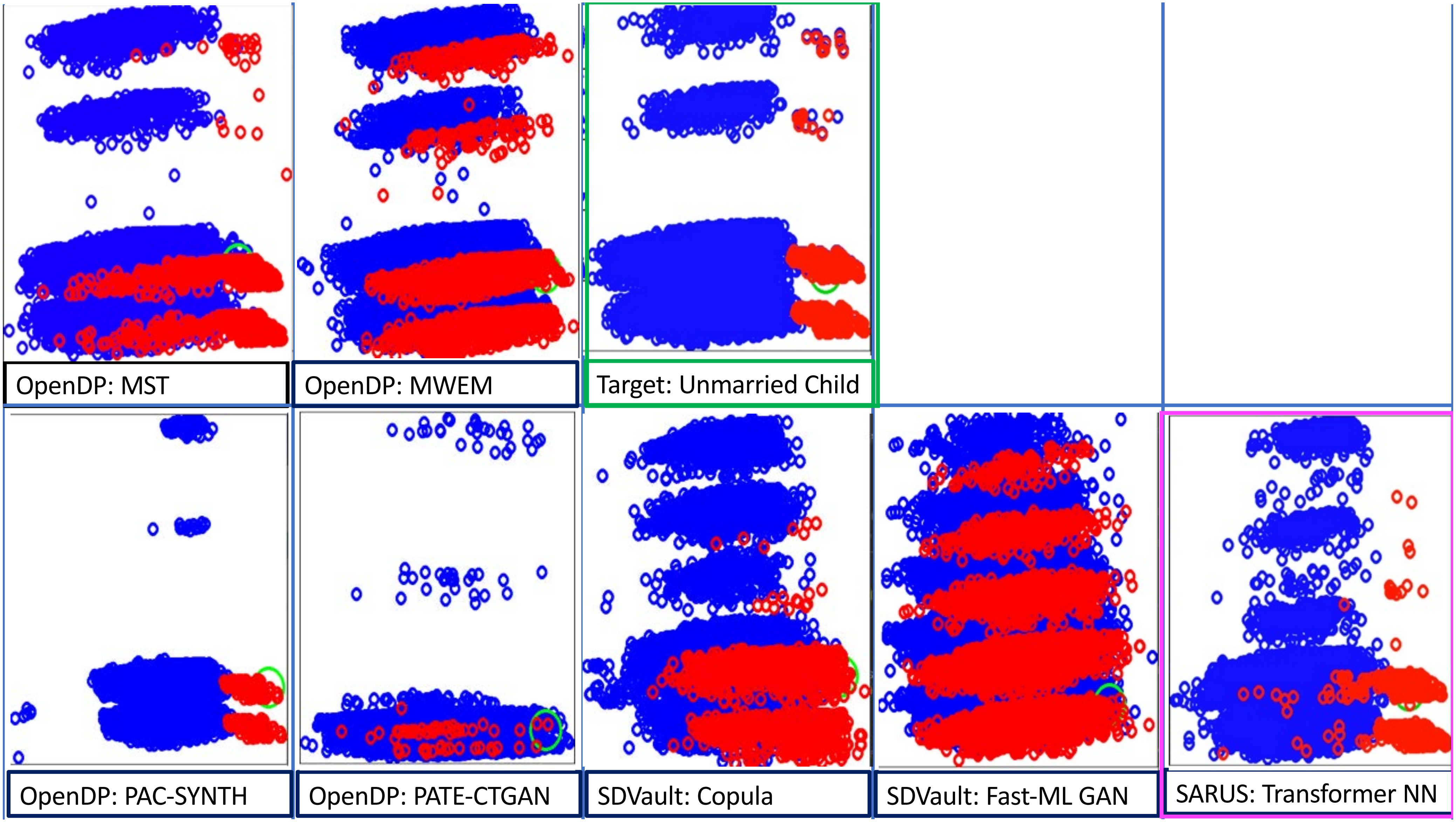


An encoder

$$E : \mathfrak{X} \rightarrow (\mathscr{E} \times \mathscr{H})$$

The encoder takes an observation $x \in \mathfrak{X}$ and returns a pair $(e, h) \in (\mathscr{E} \times \mathscr{H})$ of $e \in \mathscr{E}$ an *embedding vector* summarizing everything there is to know about x, and $h \in \mathscr{H}$ an *intermediate context* containing all intermediate information about a composite observation while being processed — this is explicited in 3.3. For simple codecs, where the intermediate context is not used, instead of using a degenerate $\mathscr{H} = \{\bullet\}$ with a single element and note $(e, \bullet) \in (\mathscr{E} \times \{\bullet\})$, we will simply note $e \in \mathscr{E}$.





OpenDP: MST

OpenDP: MWEM

Target: Unmarried Child

OpenDP: PAC-SYNTH

OpenDP: PATE-CTGAN

SDVault: Copula

SDVault: Fast-ML GAN

SARUS: Transformer NN

CART-Trees

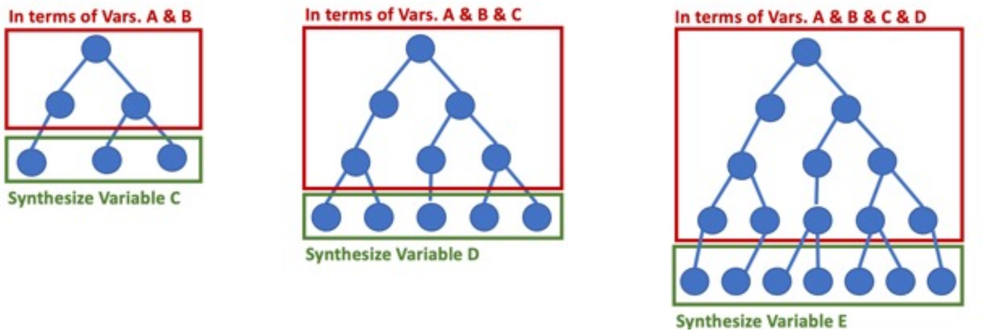
- New records are generated one feature at a time, using a sequence of decision trees that select plausible new values for each feature based on the values synthesized for previous features.
- Is a popular solution for higher dimensional government data: Scottish Longitudinal Survey, StatCan, CDC, Knexus: Census ACS/ABS/AHS, Urban Institute: IRS.
- Data was generated using the open-source R Synthpop library, from Gillian Raab and Beata Nowok. https://www.synthpop.org.uk/get-started.html
- Is not differentially private
- Feature set: Full 22 features from CRC benchmark

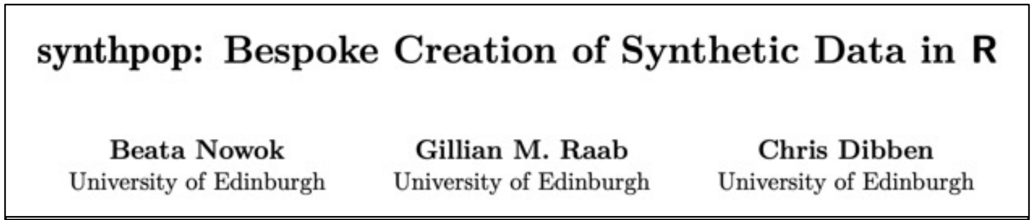
Relevant Papers:
Several, see sidebar →

ABOUT SYNTHPOP

Submitting Library Credit:
R synthpop, default settings

Decision Trees (CART Models):

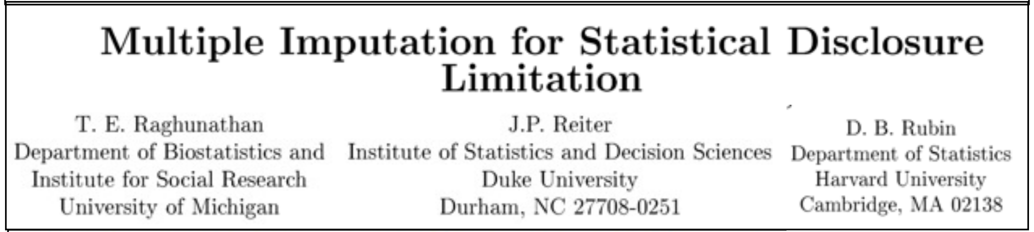



synthpop: Bespoke Creation of Synthetic Data in R

Beata Nowok
University of Edinburgh

Gillian M. Raab
University of Edinburgh

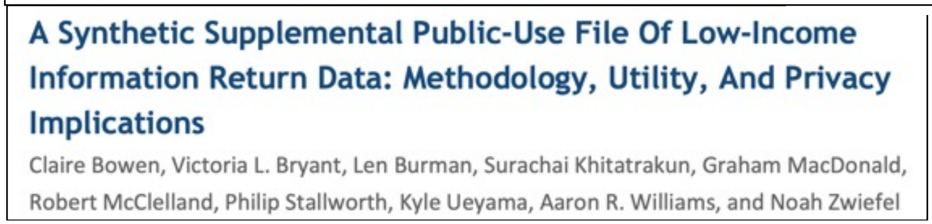
Chris Dibben
University of Edinburgh


Multiple Imputation for Statistical Disclosure Limitation

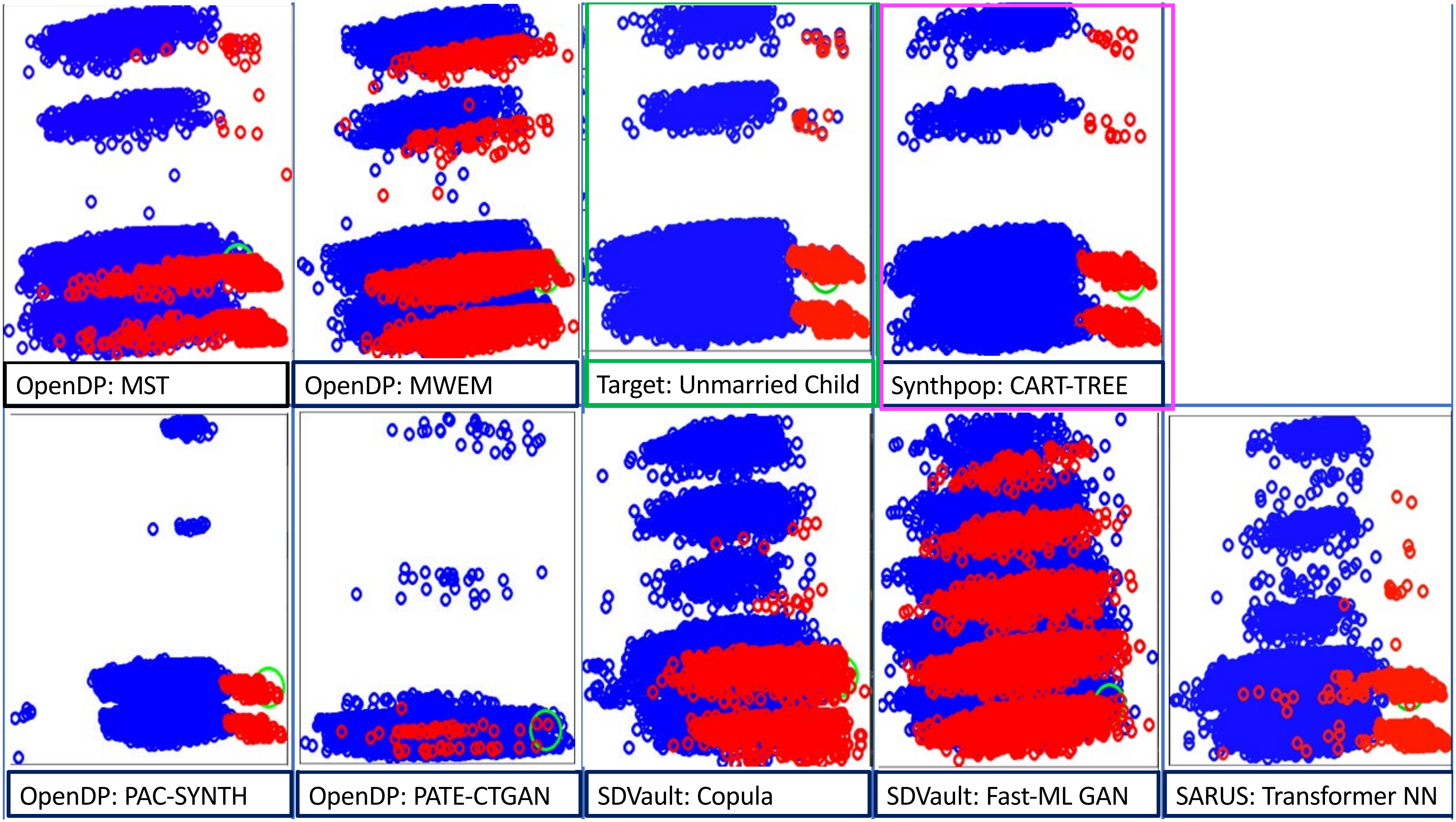
T. E. Raghunathan
Department of Biostatistics and
Institute for Social Research
University of Michigan

J.P. Reiter
Institute of Statistics and Decision Sciences
Duke University
Durham, NC 27708-0251

D. B. Rubin
Department of Statistics
Harvard University
Cambridge, MA 02138


A Synthetic Supplemental Public-Use File Of Low-Income Information Return Data: Methodology, Utility, And Privacy Implications

Claire Bowen, Victoria L. Bryant, Len Burman, Surachai Khitatrakun, Graham MacDonald, Robert McClelland, Philip Stallworth, Kyle Ueyama, Aaron R. Williams, and Noah Zwiefel


OpenDP: MST
OpenDP: MWEM
Target: Unmarried Child
Synthpop: CART-TREE
OpenDP: PAC-SYNTH
OpenDP: PATE-CTGAN
SDVault: Copula
SDVault: Fast-ML GAN
SARUS: Transformer NN

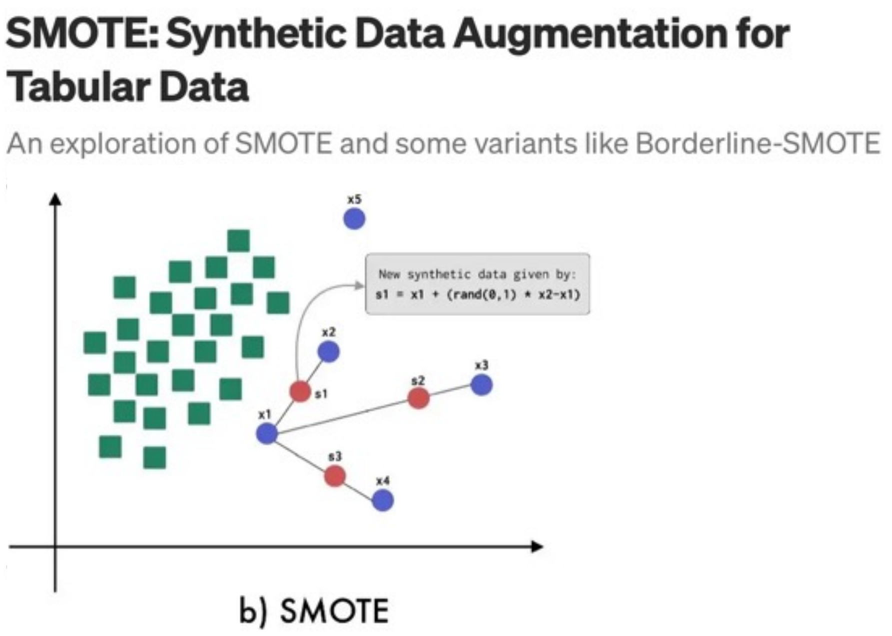
SMOTE

- New data is generated using linear interpolation on ground truth data to find new, plausible points that fill the same distribution of the original data.
- Data generator was implemented, tuned, and submitted to the CRC by Yan Zhao of UT Dallas.
- Is not differentially private
- Feature set: AGEP, SEX, MSP, RAC1P, HOUSING_TYPE, OWN_RENT, EDU, PINCP_DECILE, DVET, DEYE

Blog/Image Credit: Fernando López [Towards Data Science]

Original Paper: SMOTE: Synthetic Minority Over-sampling Technique (2002)
Nitesh V. Chawla [University of South Florida], Kevin W. Bowyer [University of Notre Dame], Lawrence O. Hall [University of South Florida], W. Philip Kegelmeyer [Sandia National Laboratories]

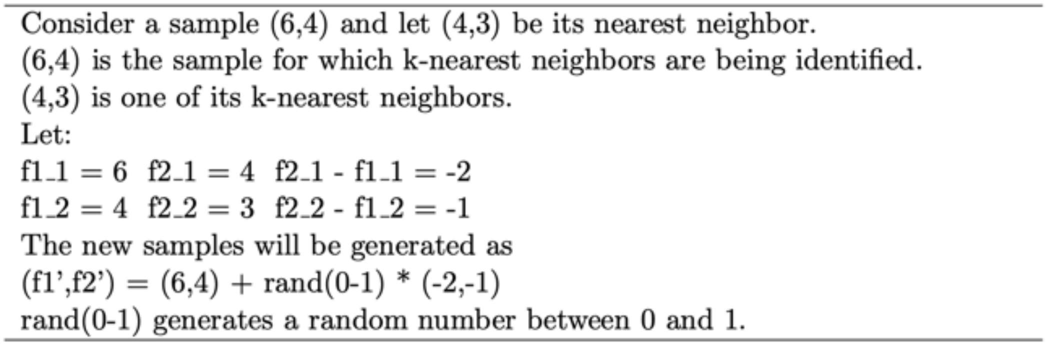
Submitting Researcher Credit:
Yan Zhao [UT Dallas]



SMOTE: Synthetic Data Augmentation for Tabular Data

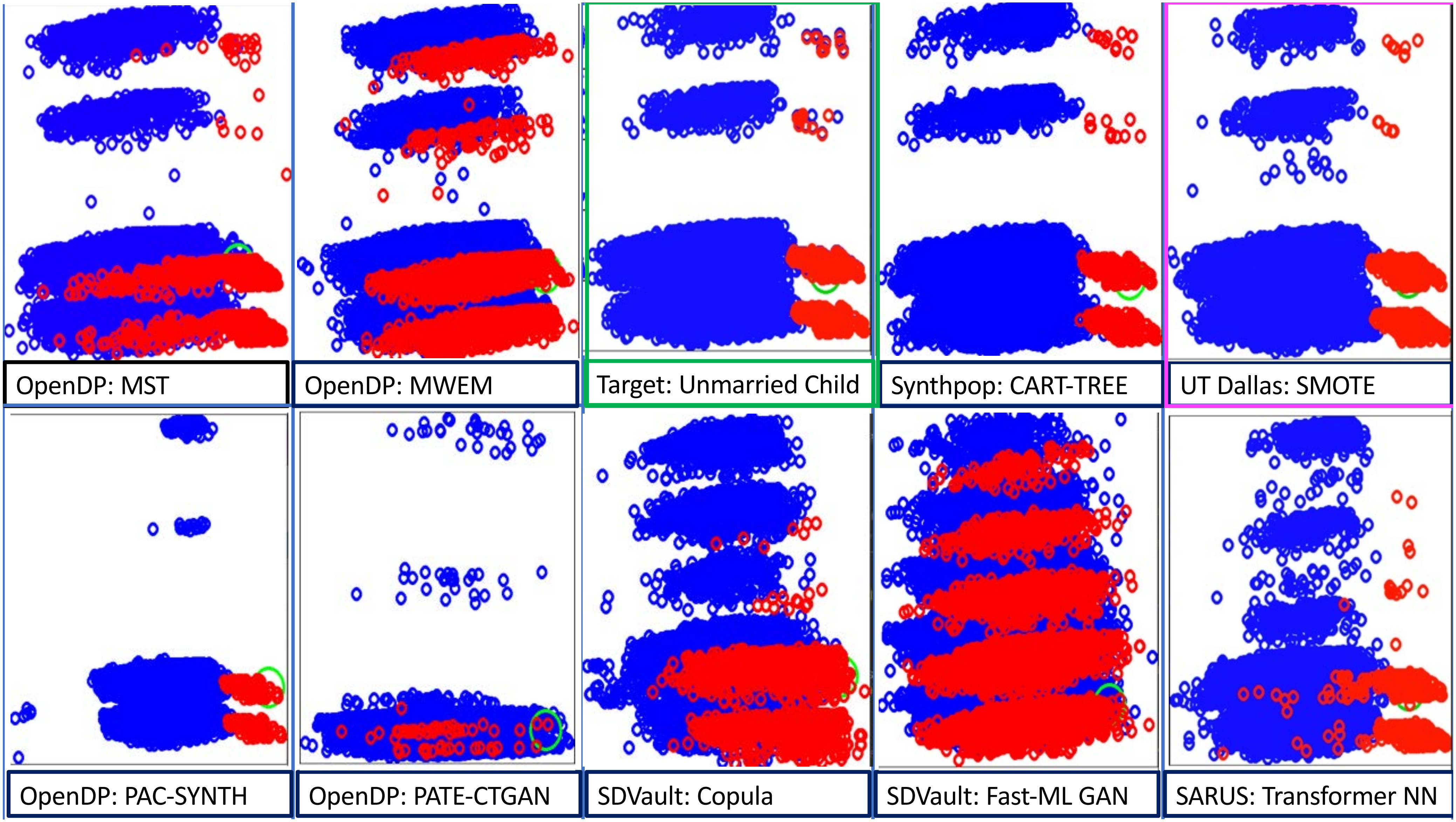
An exploration of SMOTE and some variants like Borderline-SMOTE

b) SMOTE



Consider a sample (6,4) and let (4,3) be its nearest neighbor.
(6,4) is the sample for which k-nearest neighbors are being identified.
(4,3) is one of its k-nearest neighbors.
Let:
f1_1 = 6 f2_1 = 4 f2_1 - f1_1 = -2
f1_2 = 4 f2_2 = 3 f2_2 - f1_2 = -1
The new samples will be generated as
(f1',f2') = (6,4) + rand(0-1) * (-2,-1)
rand(0-1) generates a random number between 0 and 1.

Table 1: Example of generation of synthetic examples (SMOTE).



OpenDP: MST

OpenDP: MWEM

Target: Unmarried Child

Synthpop: CART-TREE

UT Dallas: SMOTE

OpenDP: PAC-SYNTH

OpenDP: PATE-CTGAN

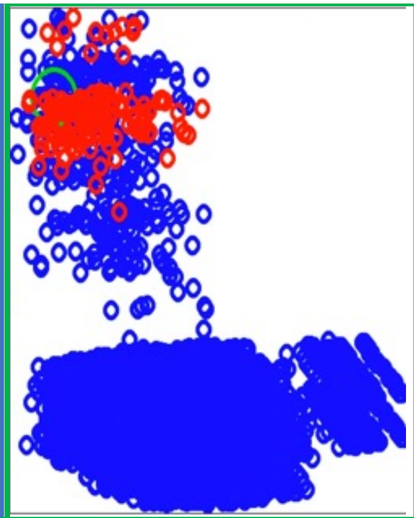
SDVault: Copula

SDVault: Fast-ML GAN

SARUS: Transformer NN

OpenDP: MST

OpenDP: MWEM



Target >70% Disabled Vet

Synthpop: CART-TREE

UT Dallas: SMOTE

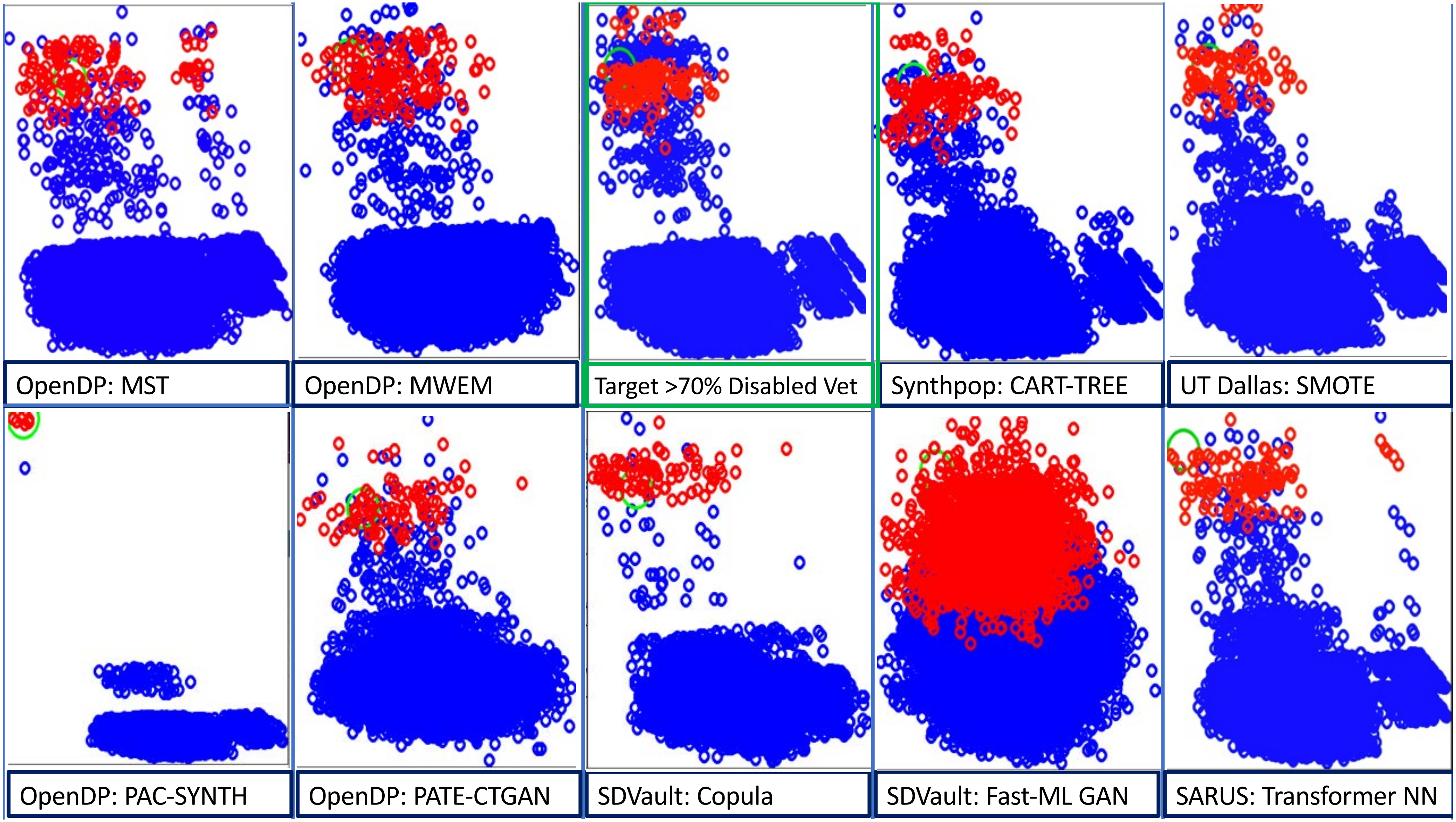
OpenDP: PAC-SYNTH

OpenDP: PATE-CTGAN

SDVault: Copula

SDVault: Fast-ML GAN

SARUS: Transformer NN



OpenDP: MST

OpenDP: MWEM

Target >70% Disabled Vet

Synthpop: CART-TREE

UT Dallas: SMOTE

OpenDP: PAC-SYNTH

OpenDP: PATE-CTGAN

SDVault: Copula

SDVault: Fast-ML GAN

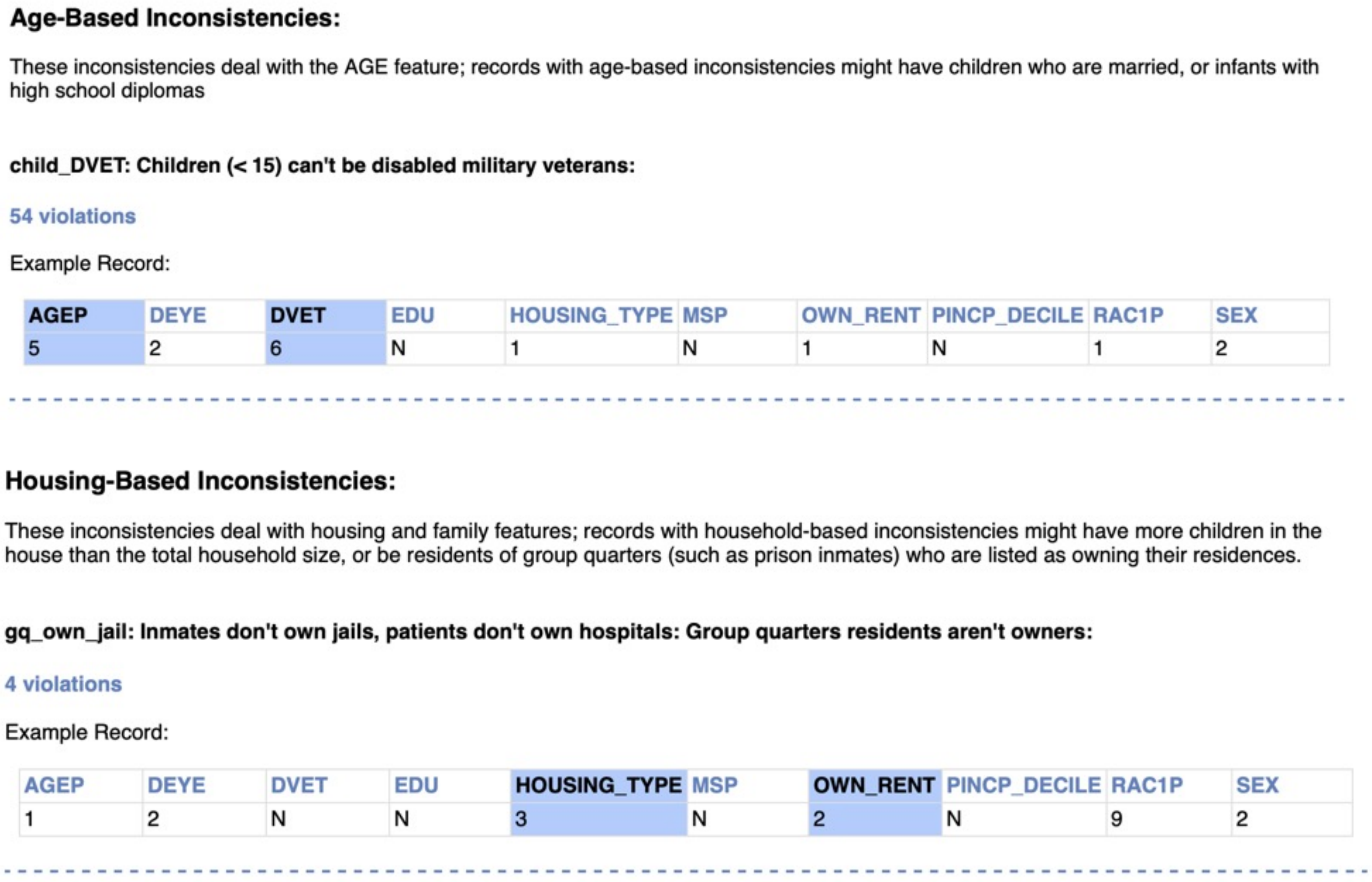
SARUS: Transformer NN

Utility Metrics: The Consistency Checks Metric

In real world survey data, it's common for records features to have deterministic dependencies on each other.

In fact, this is almost impossible to avoid (ex: knowing someone's AGE= 5, necessarily tells you something about their income and education).

But these relationships may be tricky for some model to pick up on automatically.



Age-Based 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

child_DVET: Children (< 15) can't be disabled military veterans:

54 violations

Example Record:

AGEP	DEYE	DVET	EDU	HOUSING_TYPE	MSP	OWN_RENT	PINCP_DECILE	RAC1P	SEX
5	2	6	N	1	N	1	N	1	2

Housing-Based Inconsistencies:

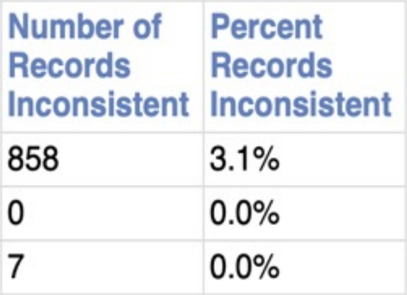
These inconsistencies deal with housing and family features; records with household-based inconsistencies might have more children in the house than the total household size, or be residents of group quarters (such as prison inmates) who are listed as owning their residences.

gq_own_jail: Inmates don't own jails, patients don't own hospitals: Group quarters residents aren't owners:

4 violations

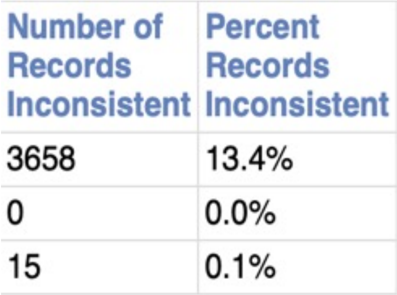
Example Record:

AGEP	DEYE	DVET	EDU	HOUSING_TYPE	MSP	OWN_RENT	PINCP_DECILE	RAC1P	SEX
1	2	N	N	3	N	2	N	9	2



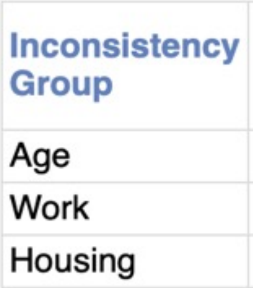
Number of Records Inconsistent	Percent Records Inconsistent
858	3.1%
0	0.0%
7	0.0%

OpenDP: MST



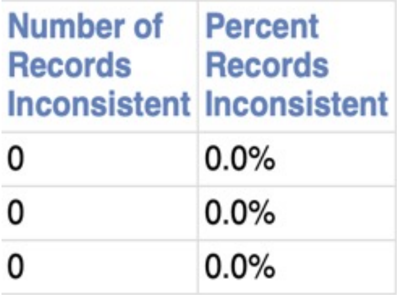
Number of Records Inconsistent	Percent Records Inconsistent
3658	13.4%
0	0.0%
15	0.1%

OpenDP: MWEM



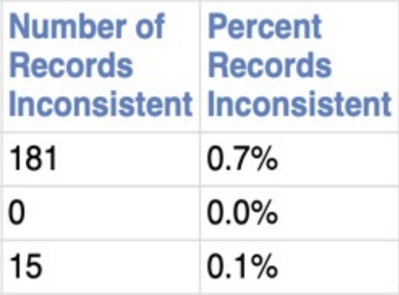
Inconsistency Group
Age
Work
Housing

Record Inconsistencies



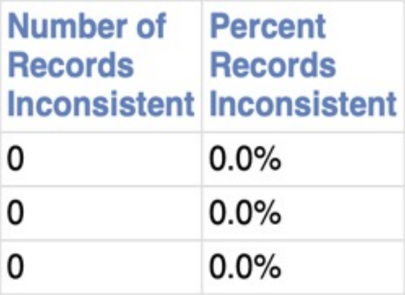
Number of Records Inconsistent	Percent Records Inconsistent
0	0.0%
0	0.0%
0	0.0%

Synthpop: CART-TREE



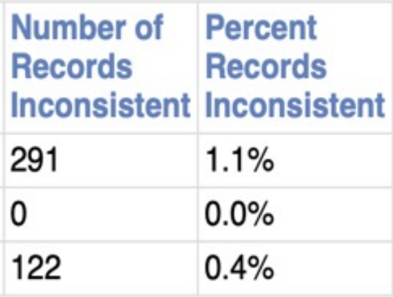
Number of Records Inconsistent	Percent Records Inconsistent
181	0.7%
0	0.0%
15	0.1%

UT Dallas: SMOTE



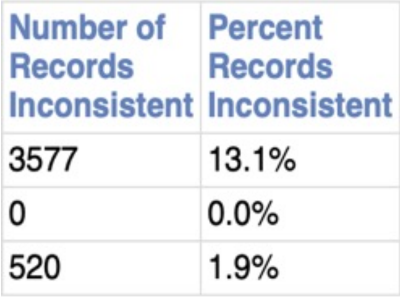
Number of Records Inconsistent	Percent Records Inconsistent
0	0.0%
0	0.0%
0	0.0%

OpenDP: PAC-SYNTH



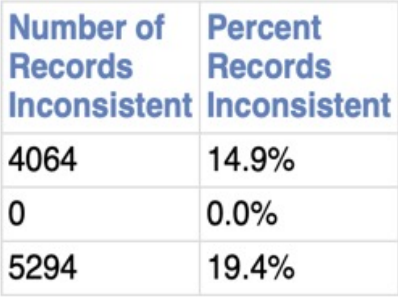
Number of Records Inconsistent	Percent Records Inconsistent
291	1.1%
0	0.0%
122	0.4%

OpenDP: PATE-CTGAN



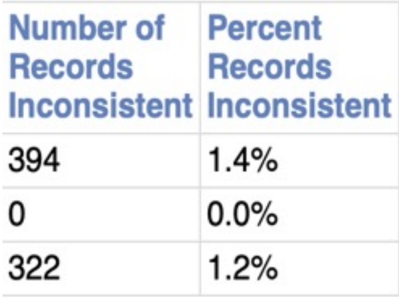
Number of Records Inconsistent	Percent Records Inconsistent
3577	13.1%
0	0.0%
520	1.9%

SDVault: Copula



Number of Records Inconsistent	Percent Records Inconsistent
4064	14.9%
0	0.0%
5294	19.4%

SDVault: Fast-ML GAN



Number of Records Inconsistent	Percent Records Inconsistent
394	1.4%
0	0.0%
322	1.2%

SARUS: Transformer NN

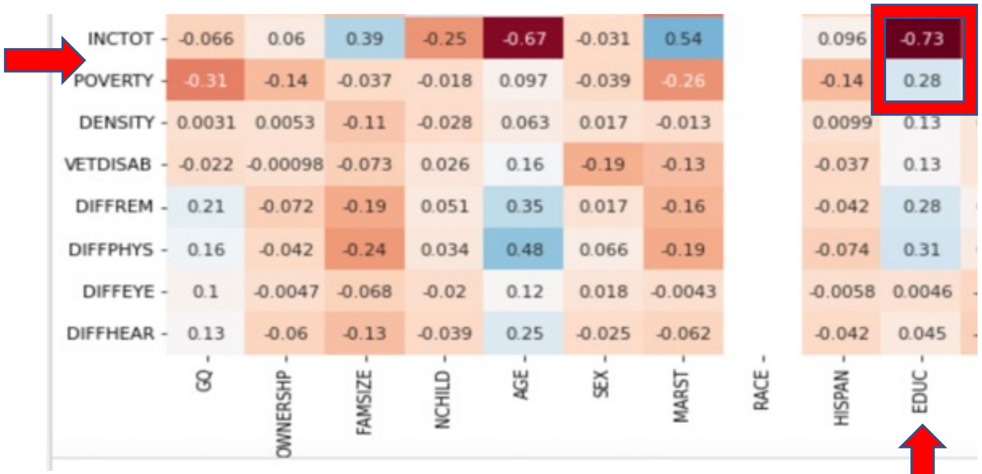
Utility Metrics: The Education/Income Regression Metric

Our work with analysts at the US Census Bureau highlighted the importance of retaining good regression performance on the privatized data.

But regression is tricky–you're fitting a line (or a curve) to a potentially complicated data distribution.

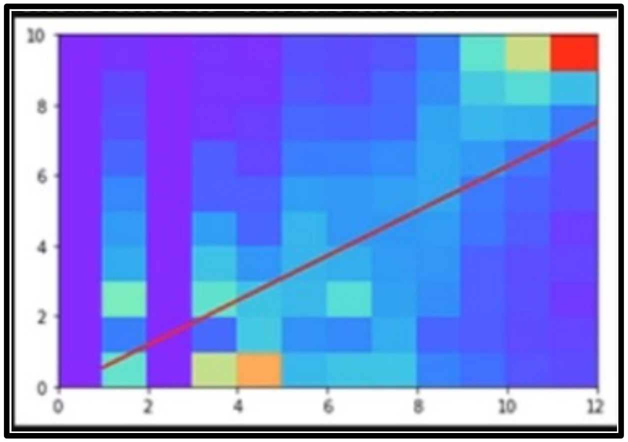
This metric takes two features that we know have different correlations for different demographic subgroups and explores how that impacts a simple linear regression

The correlation patterns between educational attainment and financial features differs for different racial groups.

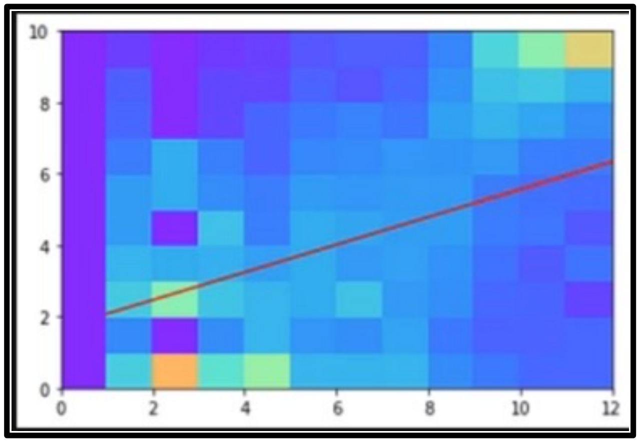


	GQ	OWNERSHP	FAMSIZE	NCHILD	AGE	SEX	MARST	RACE	HISPAN	EDUC
INCTOT	-0.066	0.06	0.39	-0.25	-0.67	-0.031	0.54		0.096	-0.73
POVERTY	-0.31	-0.14	-0.037	-0.018	0.097	-0.039	-0.26		-0.14	0.28
DENSITY	0.0031	0.0053	-0.11	-0.028	0.063	0.017	-0.013		0.0099	0.13
VETDISAB	-0.022	-0.00098	-0.073	0.026	0.16	-0.19	-0.13		-0.037	0.13
DIFFREM	0.21	-0.072	-0.19	0.051	0.35	0.017	-0.16		-0.042	0.28
DIFFPHYS	0.16	-0.042	-0.24	0.034	0.48	0.066	-0.19		-0.074	0.31
DIFFEYE	0.1	-0.0047	-0.068	-0.02	0.12	0.018	-0.0043		-0.0058	0.0046
DIFFHEAR	0.13	-0.06	-0.13	-0.039	0.25	-0.025	-0.062		-0.042	0.045

A simple linear regression between educational attainment (x-axis) and income decile (y-axis), with a normalized heatmap showing the underlying distribution.



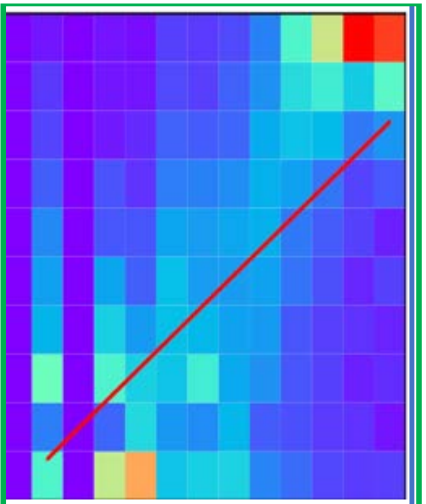
Original Data Distribution Shape



Deidentified Data Distribution Shape

OpenDP: MST

OpenDP: MWEM



Target Data: Edu. vs Inc.

Synthpop: CART-TREE

UT Dallas: SMOTE

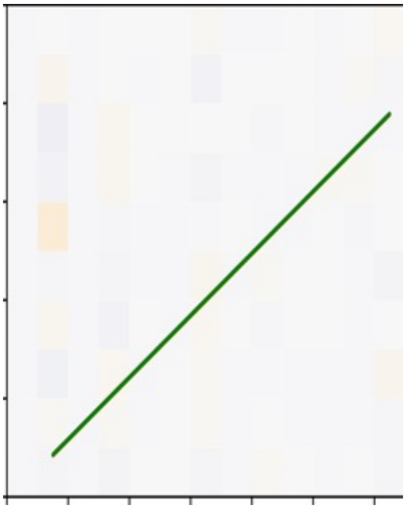
OpenDP: PAC-SYNTH

OpenDP: PATE-CTGAN

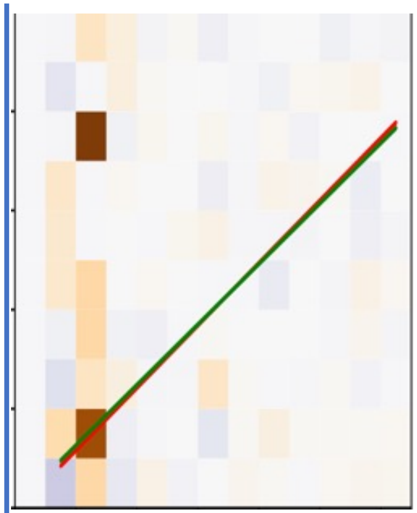
SDVault: Copula

SDVault: Fast-ML GAN

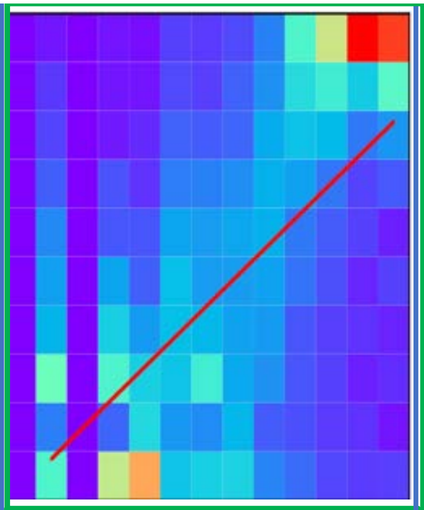
SARUS: Transformer NN



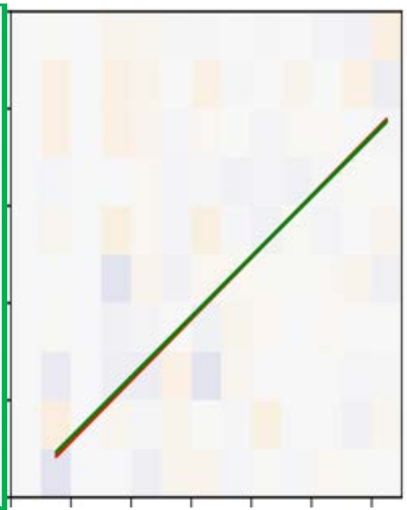
OpenDP: MST



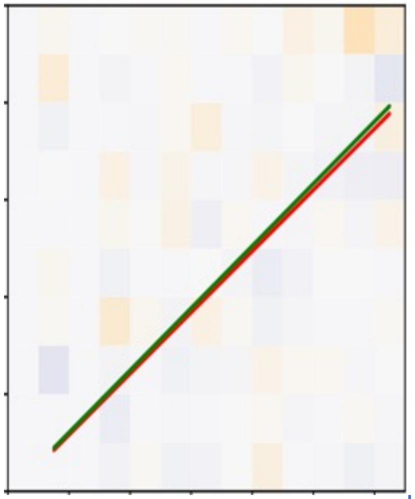
OpenDP: MWEM



Target Data: Edu. vs Inc.



Synthpop: CART-TREE



UT Dallas: SMOTE

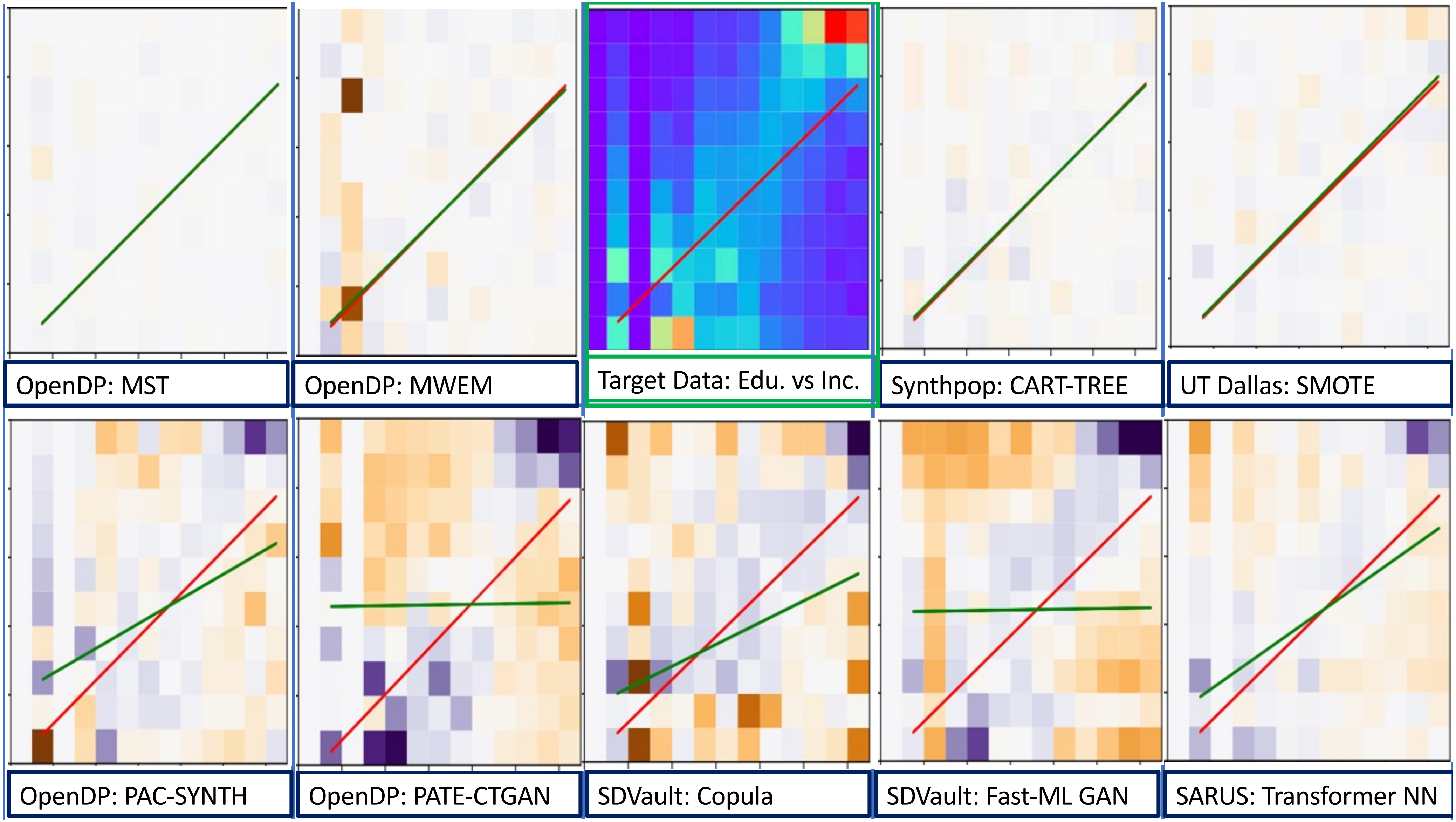
OpenDP: PAC-SYNTH

OpenDP: PATE-CTGAN

SDVault: Copula

SDVault: Fast-ML GAN

SARUS: Transformer NN



OpenDP: MST

OpenDP: MWEM

Target Data: Edu. vs Inc.

Synthpop: CART-TREE

UT Dallas: SMOTE

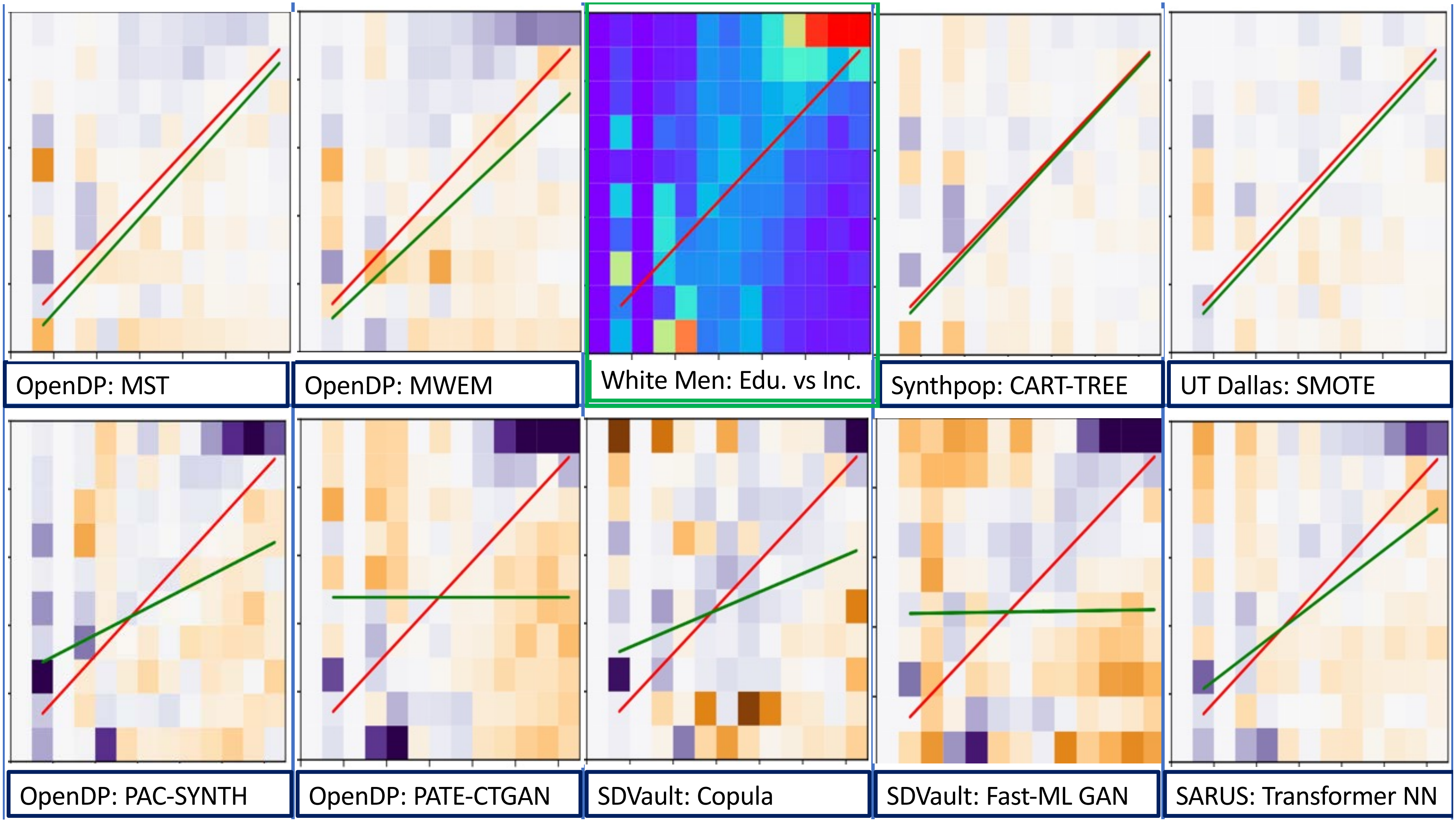
OpenDP: PAC-SYNTH

OpenDP: PATE-CTGAN

SDVault: Copula

SDVault: Fast-ML GAN

SARUS: Transformer NN



OpenDP: MST

OpenDP: MWEM

White Men: Edu. vs Inc.

Synthpop: CART-TREE

UT Dallas: SMOTE

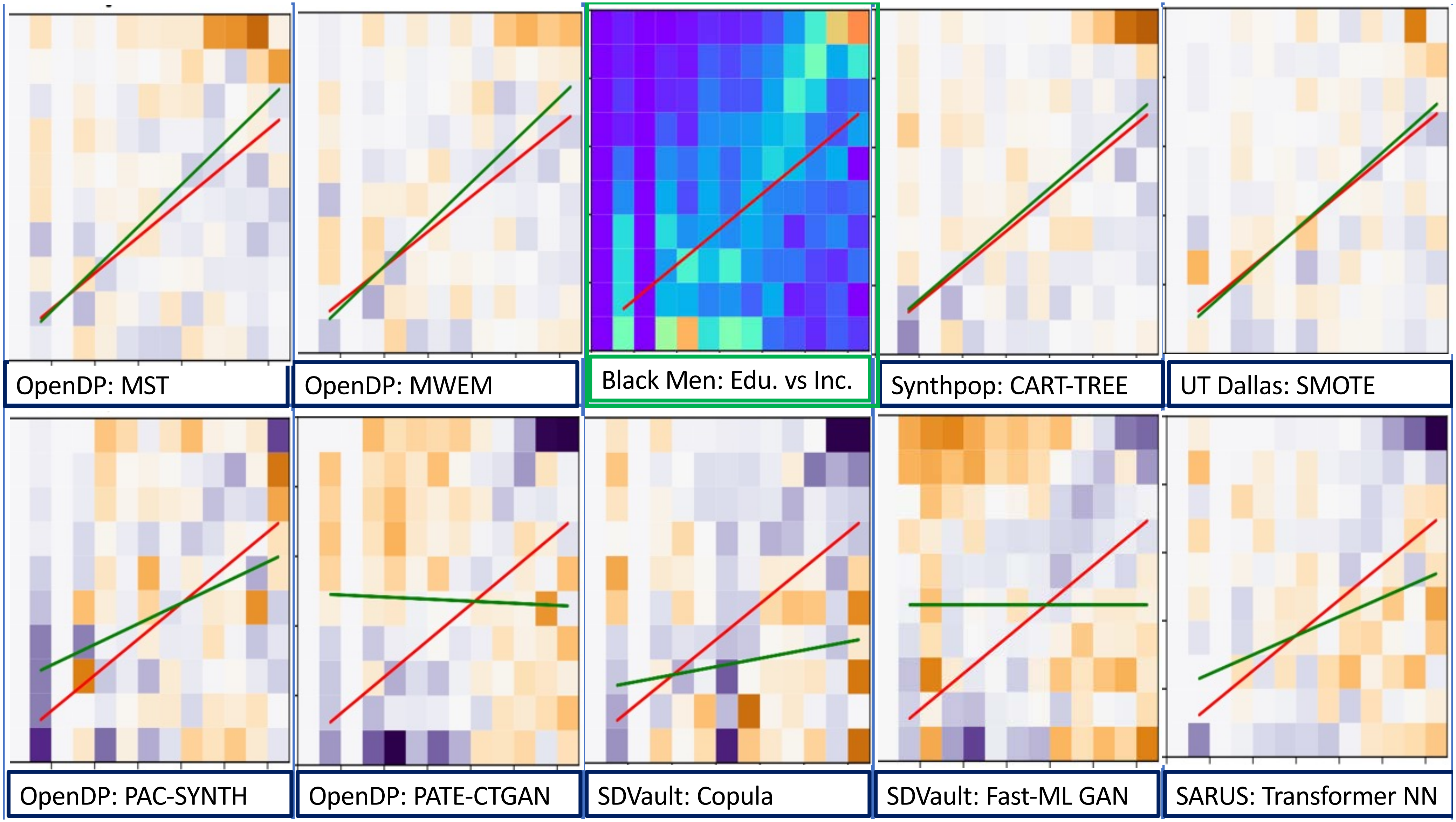
OpenDP: PAC-SYNTH

OpenDP: PATE-CTGAN

SDVault: Copula

SDVault: Fast-ML GAN

SARUS: Transformer NN



OpenDP: MST

OpenDP: MWEM

Black Men: Edu. vs Inc.

Synthpop: CART-TREE

UT Dallas: SMOTE

OpenDP: PAC-SYNTH

OpenDP: PATE-CTGAN

SDVault: Copula

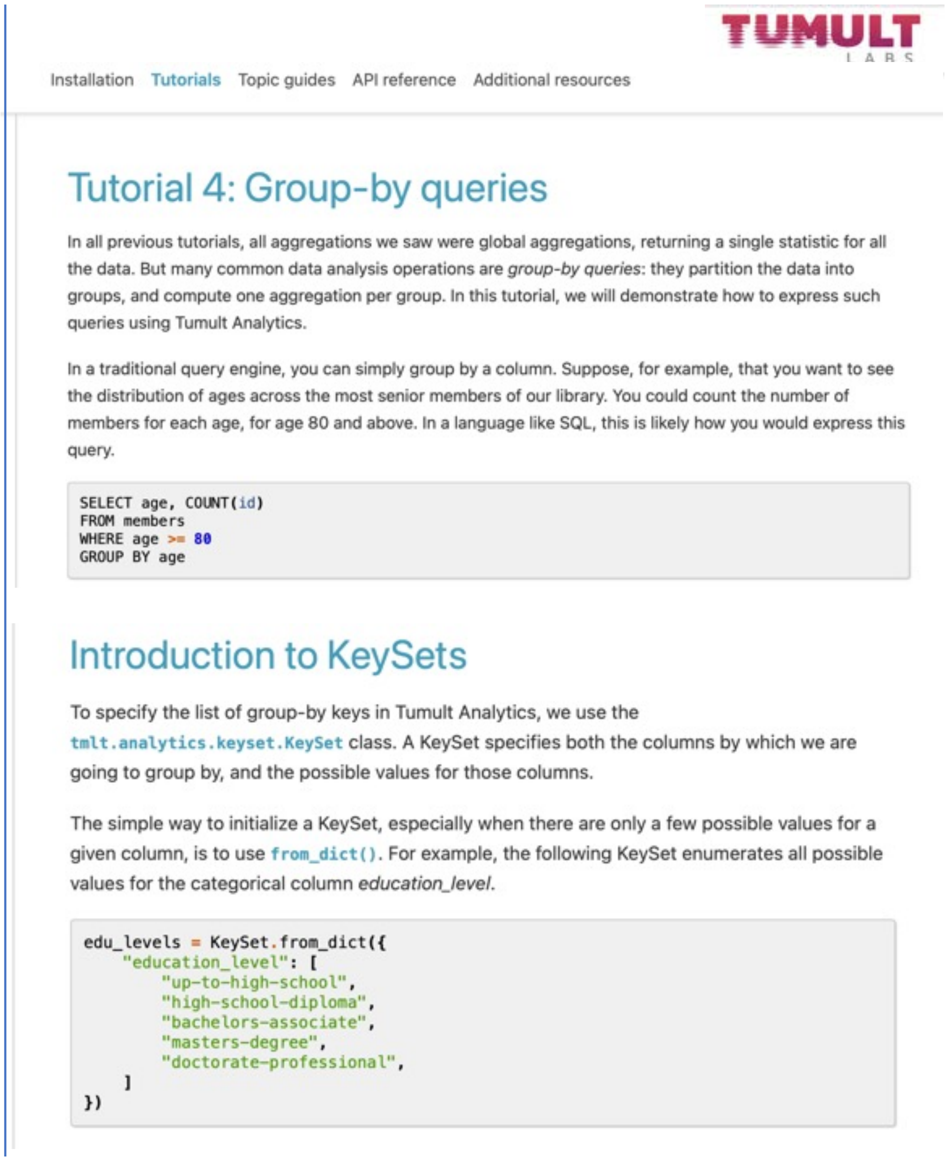
SDVault: Fast-ML GAN

SARUS: Transformer NN

Tumult Group-by

- Generated from a DP group-by counting query (and then unrolled from a histogram to deidentified), using the Tumult Analytics open-source library: https://docs.tmlt.dev/analytics/latest/installation.html
- Differentially Private, Epsilon 10
- We didn't *quite* have enough memory to run all 9 features for comparison with the other approaches. Feature set: SEX, MSP, RAC1P, HOUSING_TYPE, OWN_RENT, EDU, PINCP_DECILE, DVET

ϵKTELO: A Framework for Defining Differentially Private Computations
Dan Zhang, Ryan McKenna, Ios Kotsogiannis, George Bissias, Michael Hay, Ashwin Machanavajjhala, Gerome Miklau



Tutorial 4: Group-by queries

In all previous tutorials, all aggregations we saw were global aggregations, returning a single statistic for all the data. But many common data analysis operations are *group-by queries*: they partition the data into groups, and compute one aggregation per group. In this tutorial, we will demonstrate how to express such queries using Tumult Analytics.

In a traditional query engine, you can simply group by a column. Suppose, for example, that you want to see the distribution of ages across the most senior members of our library. You could count the number of members for each age, for age 80 and above. In a language like SQL, this is likely how you would express this query.

```
SELECT age, COUNT(id)
FROM members
WHERE age >= 80
GROUP BY age
```

Introduction to KeySets

To specify the list of group-by keys in Tumult Analytics, we use the `tmlt.analytics.keyset.KeySet` class. A KeySet specifies both the columns by which we are going to group by, and the possible values for those columns.

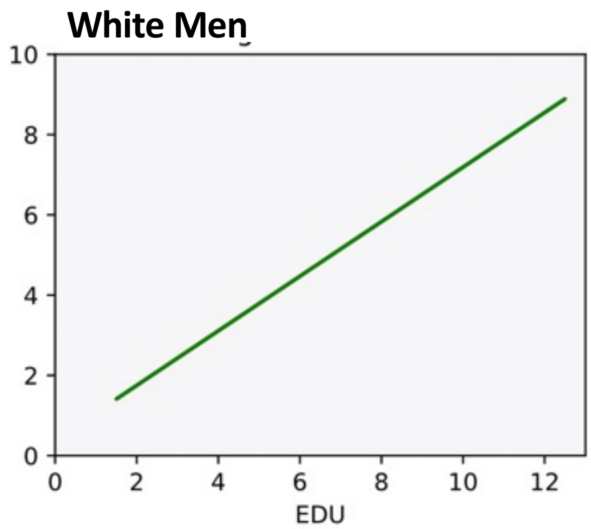
The simple way to initialize a KeySet, especially when there are only a few possible values for a given column, is to use `from_dict()`. For example, the following KeySet enumerates all possible values for the categorical column *education_level*.

```
edu_levels = KeySet.from_dict({
    "education_level": [
        "up-to-high-school",
        "high-school-diploma",
        "bachelors-associate",
        "masters-degree",
        "doctorate-professional",
    ]
})
```

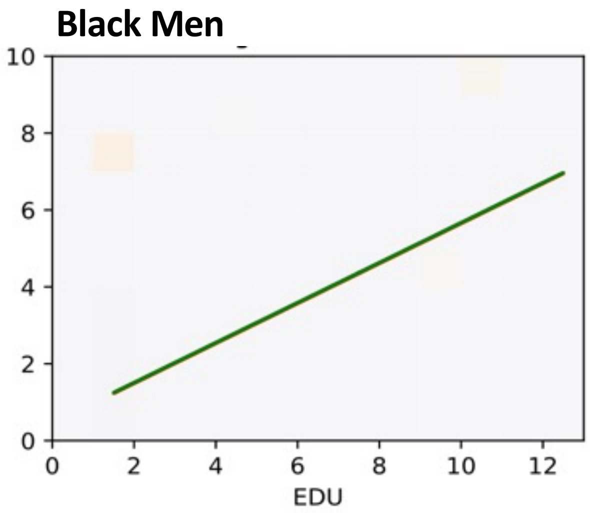
Tumult Group-by

- Generated from a DP group-by counting query (and then unrolled from a histogram to deidentified), using the Tumult Analytics open-source library: https://docs.tmlt.dev/analytics/latest/installation.html
- Differentially Private, Epsilon 10
- We didn't *quite* have enough memory to run all 9 features for comparison with the other approaches. Feature set: SEX, MSP, RAC1P, HOUSING_TYPE, OWN_RENT, EDU, PINCP_DECILE, DVET

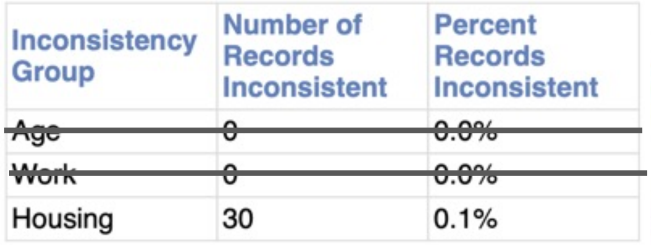
ϵKTELO: A Framework for Defining Differentially Private Computations
Dan Zhang, Ryan McKenna, Ios Kotsogiannis, George Bissias, Michael Hay, Ashwin Machanavajjhala, Gerome Miklau



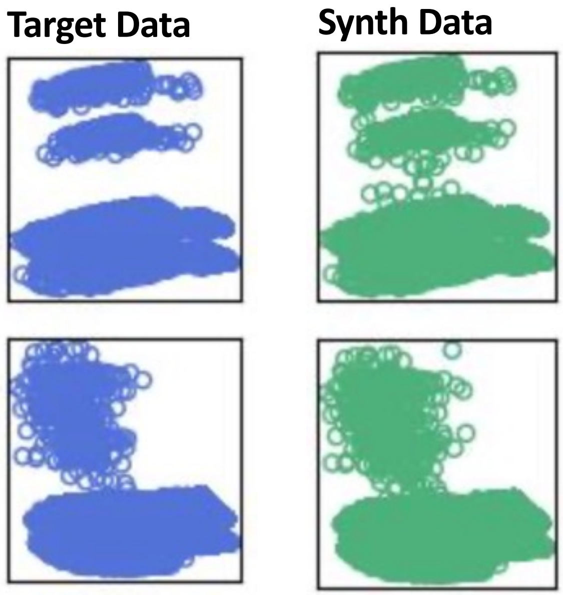
White Men



Black Men



Inconsistency Group	Number of Records Inconsistent	Percent Records Inconsistent
~~Age~~	~~0~~	~~0.0%~~
~~Work~~	~~0~~	~~0.0%~~
Housing	30	0.1%



Target Data

Synth Data

There’s More: Other Metrics in the Report

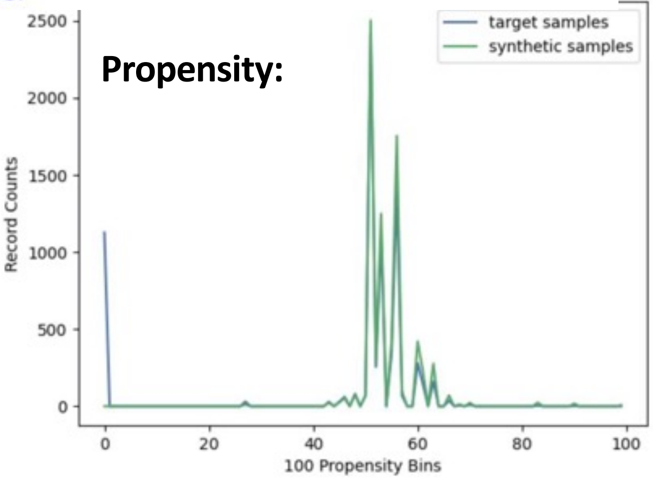
- Here we show some of the other metrics that are included in the report tool. This data was contributed by Matt Williams [RTI], using the R NPBayesImputeCat library.
- The **k-marginal** metric (from the NIST challenges) provides an overall edit distance between the real and deidentified distribution. 1000 is a perfect match.
- **Propensity** metrics train a classifier to predict whether a record is real or synthetic– the closer the lines are, the more confused the classifier, and better utility.
- The **univariates** and pairwise **Pearson Correlation** metric helps identify which features are not being modeled as well.
- The ”**Apparent Match**” metric measures privacy as a simple reidentification risk. There were **0** apparent matches in this data.

Bayesian Estimation of Discrete Multivariate Latent Structure Models With Structural Zeros

Daniel Manrique-Vallier & Jerome P. Reiter

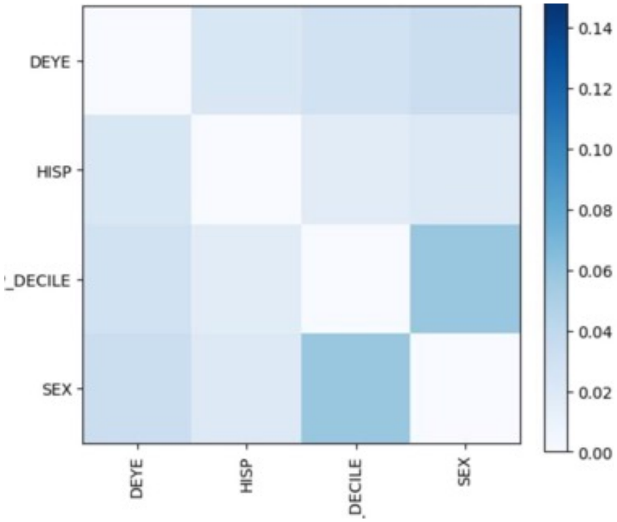
K-Marginal:

K-Marginal Score: 905

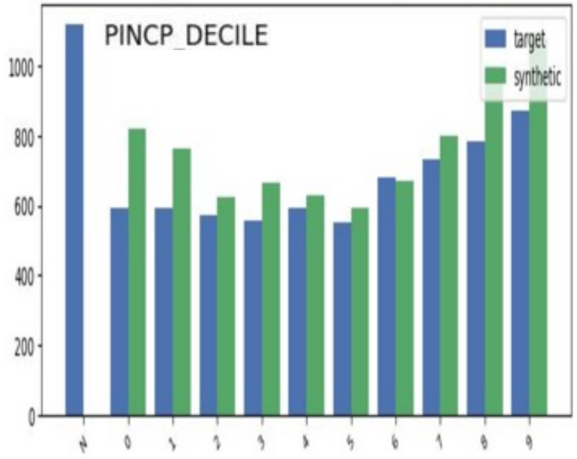


Propensity:

Pearson’s Correlation:



Univariate:



There’s More: Other Metrics in the Report

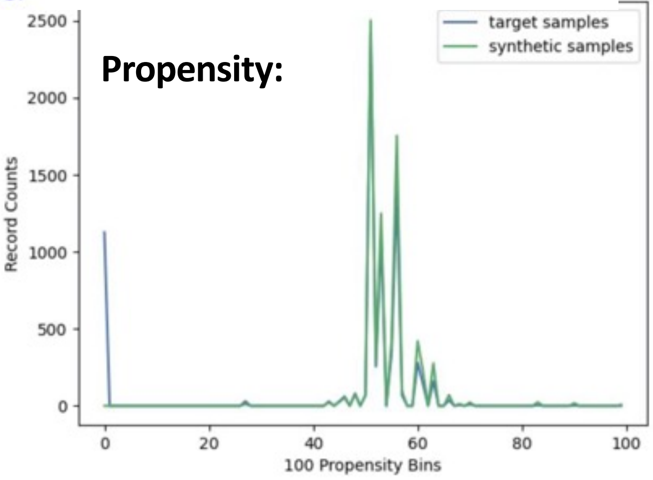
- Here we show some of the other metrics that are included in the report tool. This data was contributed by Matt Williams [RTI], using the R NPBayesImputeCat library.
- The **k-marginal** metric (from the NIST challenges) provides an overall edit distance between the real and deidentified distribution. 1000 is a perfect match.
- **Propensity** metrics train a classifier to predict whether a record is real or synthetic– the closer the lines are, the more confused the classifier, and better utility.
- The **univariates** and pairwise **Pearson Correlation** metric helps identify which features are not being modeled as well.
- The ”**Apparent Match**” metric measures privacy as a simple reidentification risk. There were **0** apparent matches in this data.

Bayesian Estimation of Discrete Multivariate Latent Structure Models With Structural Zeros

Daniel Manrique-Vallier & Jerome P. Reiter

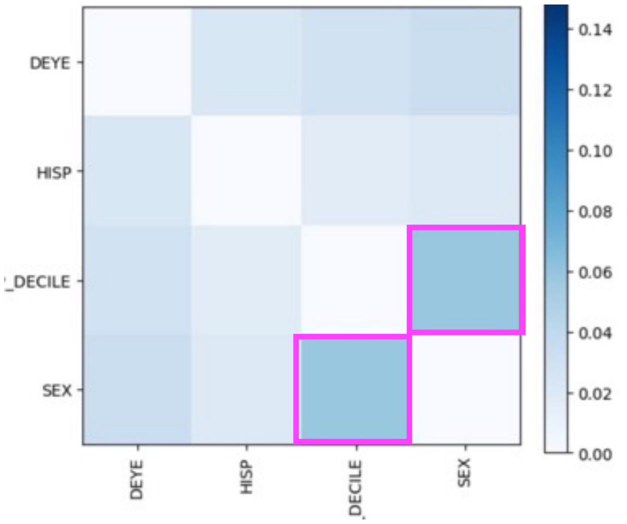
K-Marginal:

K-Marginal Score: 905

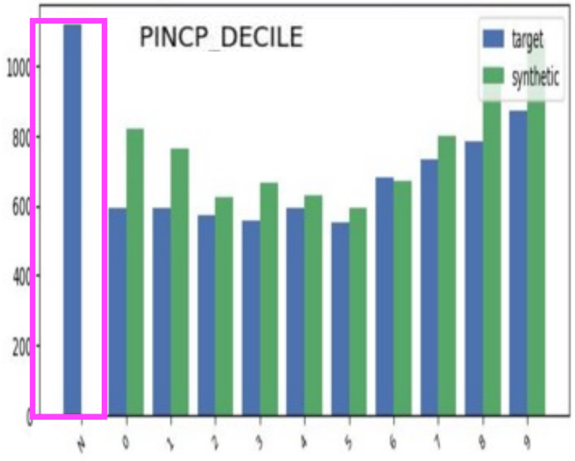


Propensity:

Pearson’s Correlation:



Univariate:

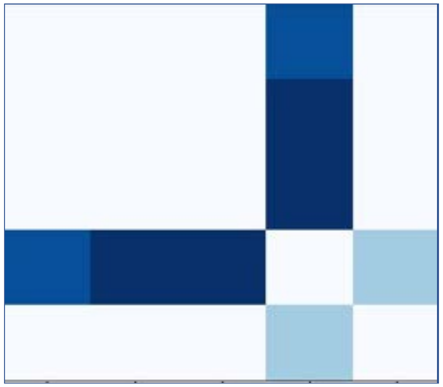


Comparing DP to Traditional Disclosure Avoidance

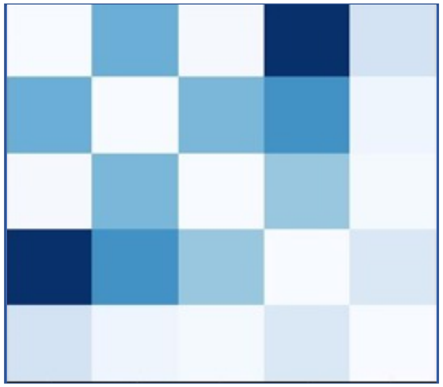
- The benchmark data and report tool can be used to compare ***any deidentification technique***, not just synthetic data
- Cuong Tran, Keyu Zhu, Saswat Das are exploring how traditional anonymization techniques like Swapping and Cell Suppression compare to DP Histograms. Older techniques work to remove or alter records that are more vulnerable to reidentification, but this may have disparate impact on diverse communities.
- They've also made randomized variants of older techniques with formal privacy.
- Looking at the Pearson Correlation Metric, we see how different approaches respond to the diverse community data.

Privacy and Bias Analysis of Disclosure Avoidance Systems

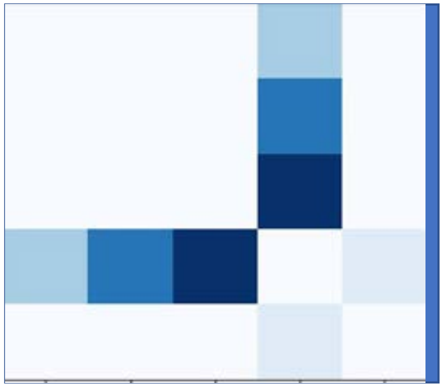
Keyu Zhu, Ferdinando Fioretto, Pascal Van Hentenryck, Saswat Das, Christine Task



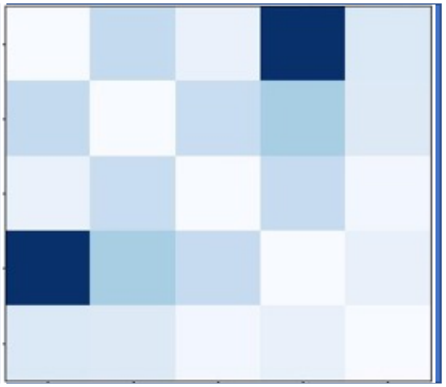
Traditional Swapping



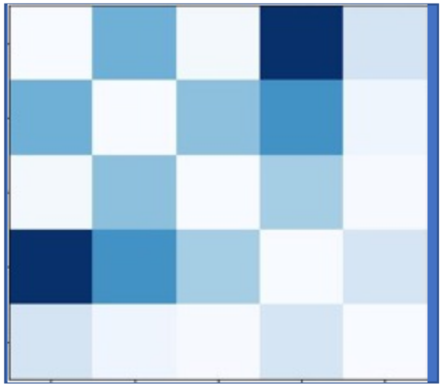
Traditional Cell Supp.



DP Swapping



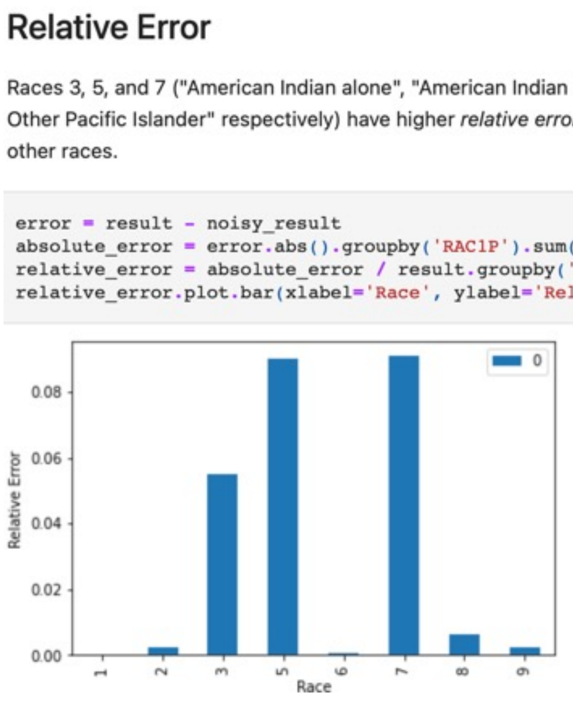
DP Histogram



DP Cell Suppression

Just A Basic Laplace Histogram

- Joe Near [University of Vermont] demonstrates the benefits benchmark data on the tiny scale too. This is just a simple Laplace Histogram in Jupyter Notebook, looking at a two histogram of Race (RAC1P) and Geography (PUMA). Eps. = 10.
- He notes that the relative error is much worse for some groups than others.
- Let's stop to consider the definition of the RAC1P feature. The Native (AIANHN) demographic is split into finer-grained labels than other demographic groups, and this also splits the population group into much smaller counts. In the MA data set, some subgroups are so small we erase them when we provide privacy.
- For more clear, accessible explanations of Differential Privacy, its variants, and its implementation on real data, you should check out **Joe Near and Chike Abuah's book--Programming Differential Privacy** https://programming-dp.com



Relative Error

Races 3, 5, and 7 ("American Indian alone", "American Indian and Alaska Native tribes" and "Native Hawaiian and Other Pacific Islander" respectively) have higher *relative error* (L1 error over the PUMA histogram) compared to other races.

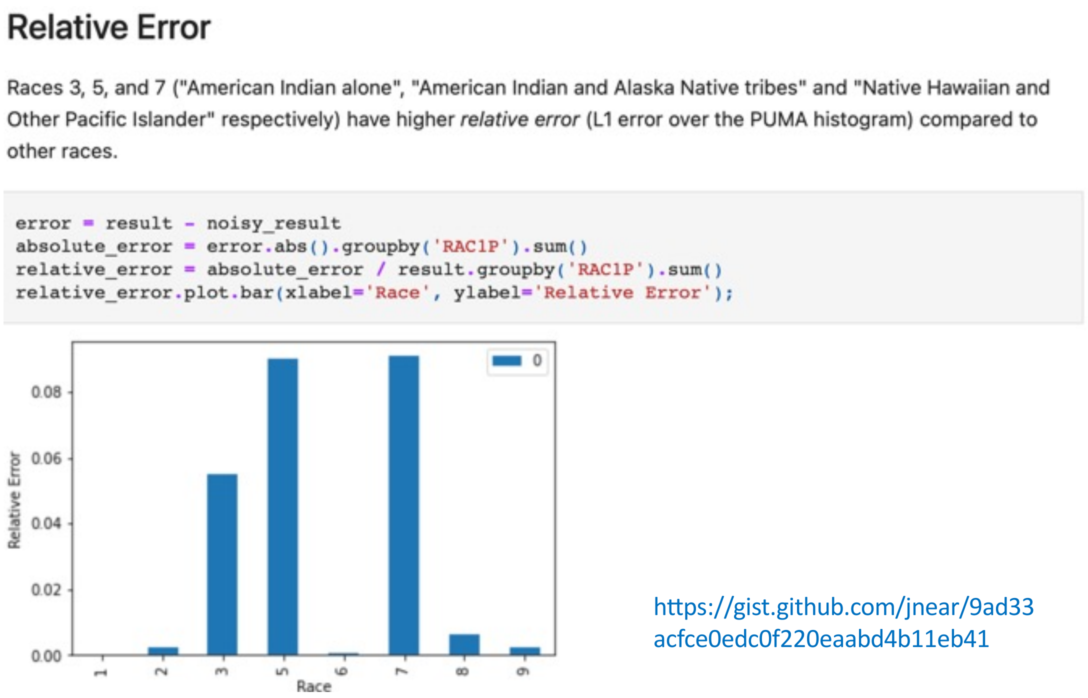
```
error = result - noisy_result
absolute_error = error.abs().groupby('RAC1P').sum()
relative_error = absolute_error / result.groupby('RAC1P').sum()
relative_error.plot.bar(xlabel='Race', ylabel='Relative Error');
```

https://gist.github.com/jnear/9ad33acfce0edc0f220eaabd4b11eb41

RAC1P Code	Code Description
1	White alone
2	Black or African American alone
3	American Indian alone
4	Alaska Native alone
5	American Indian and Alaska Native tribes specified; or American Indian or Alaska Native, not specified and no other races
6	Asian alone
7	Native Hawaiian and Other Pacific Islander alone
8	Some Other Race alone
9	Two or More Races

Just A Basic Laplace Histogram

- Joe Near [University of Vermont] demonstrates the benefits benchmark data on the tiny scale too. This is just a simple Laplace Histogram in Jupyter Notebook, looking at a two histogram of Race (RAC1P) and Geography (PUMA). Eps. = 10.
- He notes that the relative error is much worse for some groups than others.
- Let's stop to consider the definition of the RAC1P feature. The Native (AIANHN) demographic is split into finer-grained labels than other demographic groups, and this also splits the population group into much smaller counts. In the MA data set, some subgroups are so small we erase them when we provide privacy.
- For more clear, accessible explanations of Differential Privacy, its variants, and its implementation on real data, you should check out **Joe Near and Chike Abuah's book--Programming Differential Privacy** https://programming-dp.com

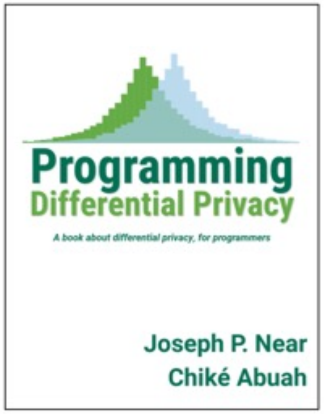


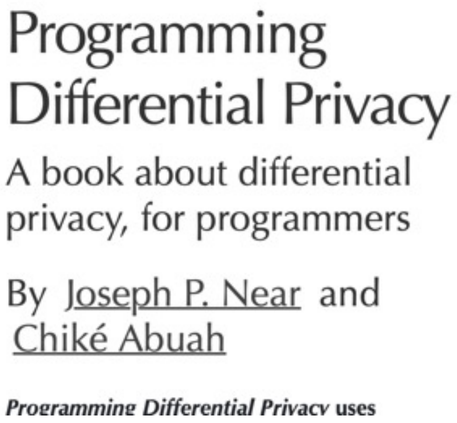
Relative Error

Races 3, 5, and 7 ("American Indian alone", "American Indian and Alaska Native tribes" and "Native Hawaiian and Other Pacific Islander" respectively) have higher *relative error* (L1 error over the PUMA histogram) compared to other races.

```
error = result - noisy_result
absolute_error = error.abs().groupby('RAC1P').sum()
relative_error = absolute_error / result.groupby('RAC1P').sum()
relative_error.plot.bar(xlabel='Race', ylabel='Relative Error');
```

https://gist.github.com/jnear/9ad33acfce0edc0f220eaabd4b11eb41





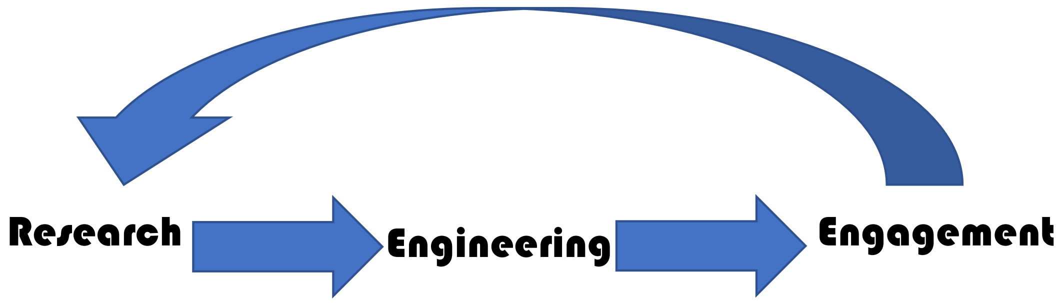
Programming Differential Privacy

A book about differential privacy, for programmers

By Joseph P. Near and Chiké Abuah

Programming Differential Privacy uses

So, got an Open Problem in mind yet?



Research

Engineering

Engagement

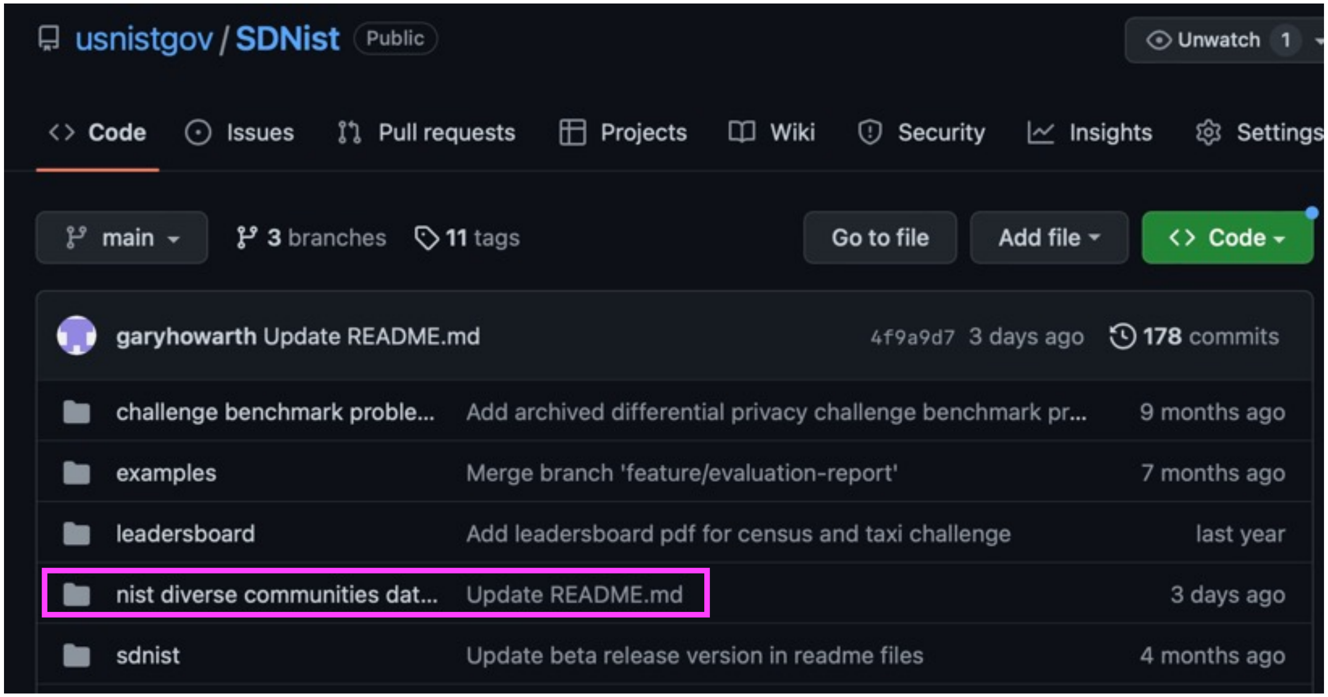
Here’s some that have been bugging us:

- **Consistency:** Not only why do some methods automatically catch record consistency and others don’t. Why do different methods do better on *different* feature categories? How does epsilon impact this? How can we improve results?
- **Through the looking glass:** Why (and how, and when) do data modeling techniques magnify some parts of the population and shrink others? What’s happening when GANs introduce artifacts, or marginal methods reduce diversity?
- **Equity and bias:** What happens when a minority demographic has a different pattern of feature correlations than the majority (see regression metric)? Which methods are more or less expressive for retaining these differences and why?
- **Granularity:** What happens when a feature definition gives fine-grained information on a particular demographic group, causing only that group to be more sparsely spread out across the data space? (Consider RAC!P and AIANHN, or DVET and military veterans). Which privacy approaches handle this more or less gracefully? How can we improve?
- **More exciting feature sets:** All of the experiments here were done with less than half the features in the data. How do these results change when you’re running on 15 features? On all 22? What if you consider weights, or household joins?
- **The thing you mention to me when you come up to ask how to register your team during the break:** I’m looking forwards to hearing about it!

Important Stuff: Data

- It’s easy to get at the data, to try it yourself.

Google “SDNIST”



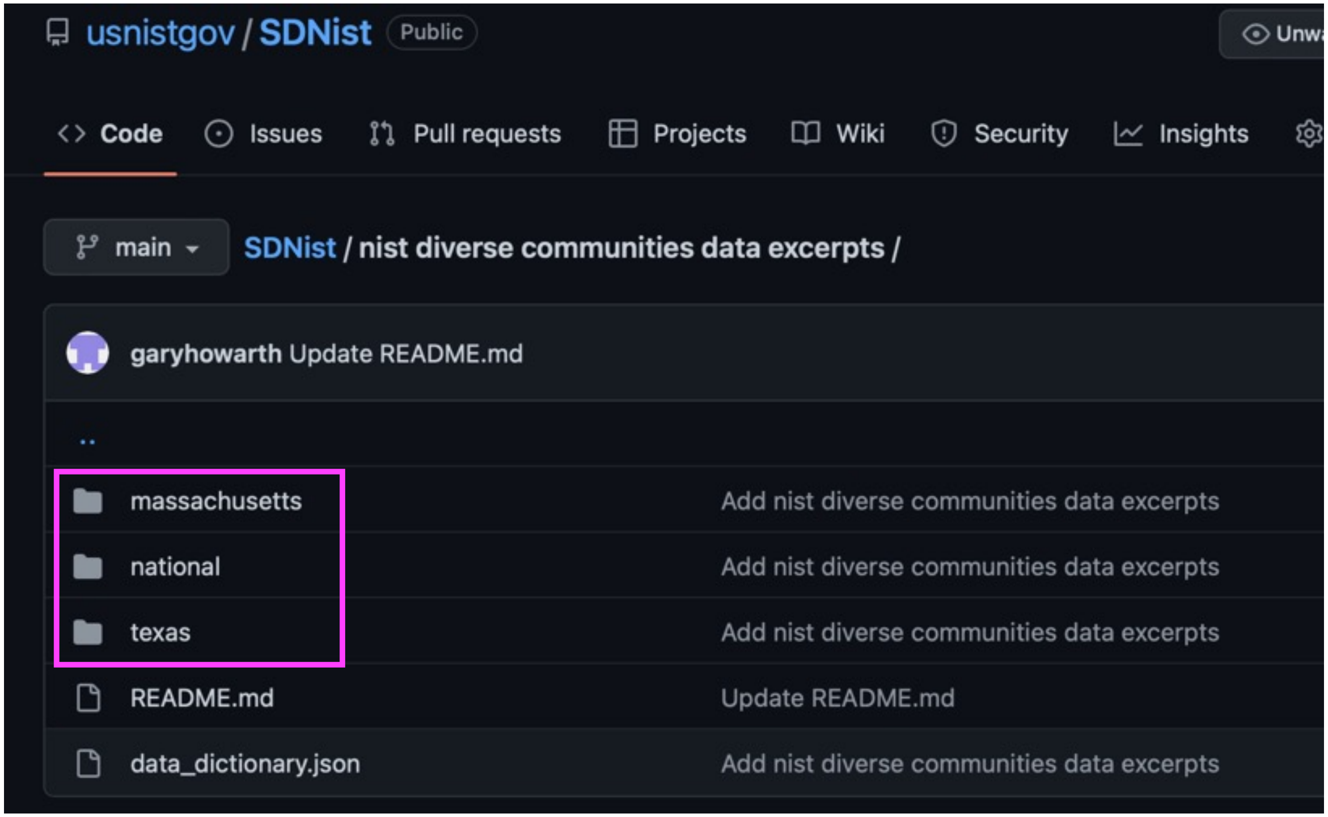
garyhowarth Update README.md 4f9a9d7 3 days ago 178 commits

challenge benchmark proble...	Add archived differential privacy challenge benchmark pr...	9 months ago
examples	Merge branch 'feature/evaluation-report'	7 months ago
leadersboard	Add leadersboard pdf for census and taxi challenge	last year
nist diverse communities dat...	Update README.md	3 days ago
sdnist	Update beta release version in readme files	4 months ago

Important Stuff: Data

- It’s easy to get at the data, to try it yourself.
- There’s three data sets. MA (North Boston) is less diverse, TX (West Dallas) is more diverse, and NATIONAL are 20 of the hardest geographies we could find from around the country – they were used to help select the final version of the Census DAS algorithm.

Google “SDNIST”



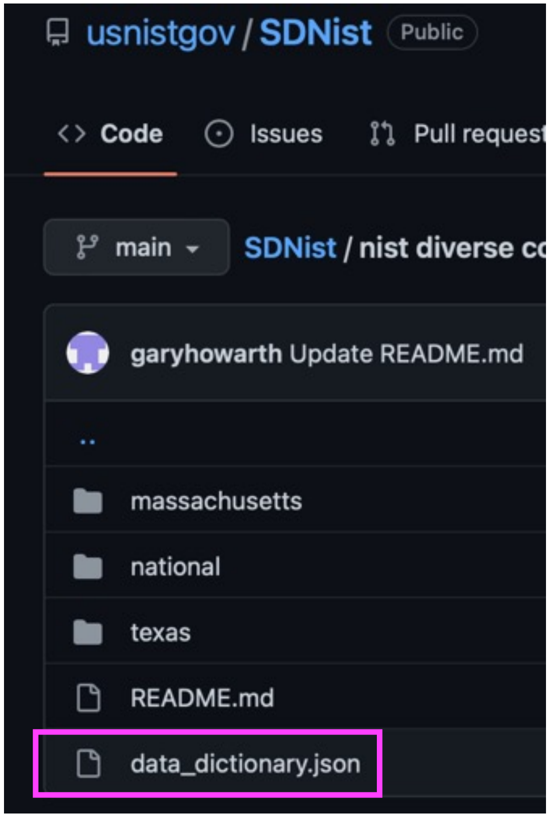
SDNist / nist diverse communities data excerpts /

garyhowarth Update README.md

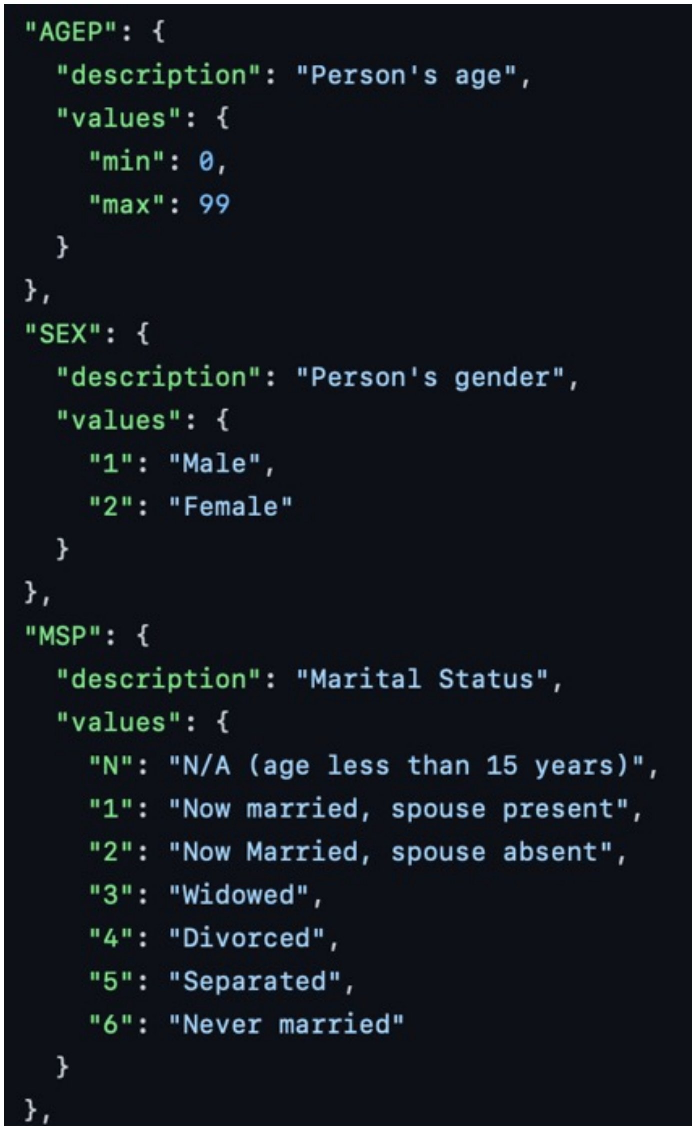
..	
massachusetts	Add nist diverse communities data excerpts
national	Add nist diverse communities data excerpts
texas	Add nist diverse communities data excerpts
README.md	Update README.md
data_dictionary.json	Add nist diverse communities data excerpts

Important Stuff: Data

- It’s easy to get at the data, to try it yourself.
- There’s three data sets. MA (North Boston) is less diverse, TX (West Dallas) is more diverse, and NATIONAL are 20 of the hardest geographies we could find from around the country – they were used to help select the final version of the Census DAS algorithm.
- There’s a .json data dictionary and detailed usage guidance. We make configuring easy.



Google “SDNIST”

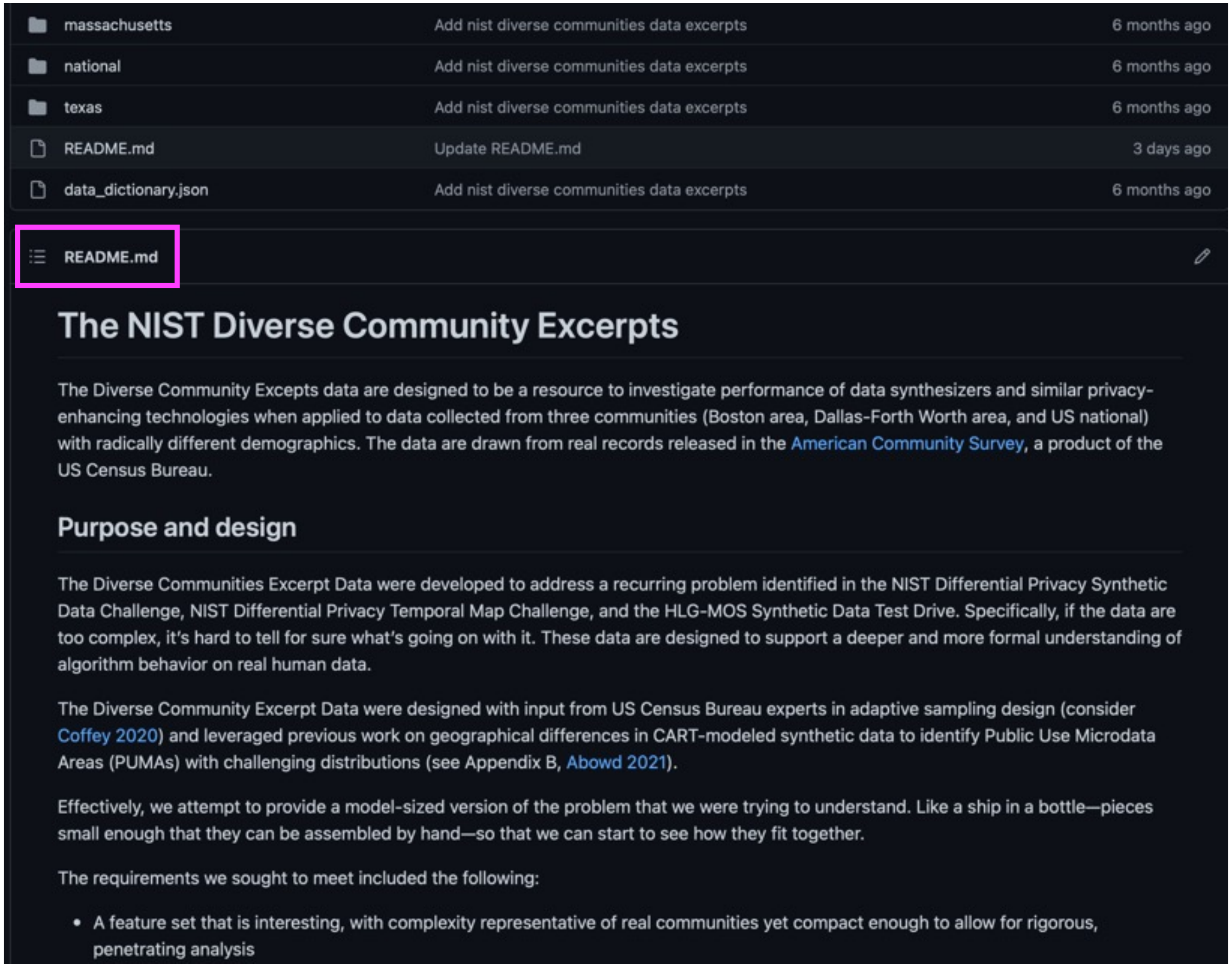


```
"AGEP": {
  "description": "Person's age",
  "values": {
    "min": 0,
    "max": 99
  }
},
"SEX": {
  "description": "Person's gender",
  "values": {
    "1": "Male",
    "2": "Female"
  }
},
"MSP": {
  "description": "Marital Status",
  "values": {
    "N": "N/A (age less than 15 years)",
    "1": "Now married, spouse present",
    "2": "Now Married, spouse absent",
    "3": "Widowed",
    "4": "Divorced",
    "5": "Separated",
    "6": "Never married"
  }
},
```

Important Stuff: Data

- It’s easy to get at the data, to try it yourself.
- There’s three data sets. MA (North Boston) is less diverse, TX (West Dallas) is more diverse, and NATIONAL are 20 of the hardest geographies we could find from around the country – they were used to help select the final version of the Census DAS algorithm.
- There’s a .json data dictionary and detailed usage guidance. We make configuring easy.

Google “SDNIST”



massachusetts	Add nist diverse communities data excerpts	6 months ago
national	Add nist diverse communities data excerpts	6 months ago
texas	Add nist diverse communities data excerpts	6 months ago
README.md	Update README.md	3 days ago
data_dictionary.json	Add nist diverse communities data excerpts	6 months ago

The NIST Diverse Community Excerpts

The Diverse Community Excepts data are designed to be a resource to investigate performance of data synthesizers and similar privacy-enhancing technologies when applied to data collected from three communities (Boston area, Dallas-Forth Worth area, and US national) with radically different demographics. The data are drawn from real records released in the American Community Survey, a product of the US Census Bureau.

Purpose and design

The Diverse Communities Excerpt Data were developed to address a recurring problem identified in the NIST Differential Privacy Synthetic Data Challenge, NIST Differential Privacy Temporal Map Challenge, and the HLG-MOS Synthetic Data Test Drive. Specifically, if the data are too complex, it's hard to tell for sure what's going on with it. These data are designed to support a deeper and more formal understanding of algorithm behavior on real human data.

The Diverse Community Excerpt Data were designed with input from US Census Bureau experts in adaptive sampling design (consider Coffey 2020) and leveraged previous work on geographical differences in CART-modeled synthetic data to identify Public Use Microdata Areas (PUMAs) with challenging distributions (see Appendix B, Abowd 2021).

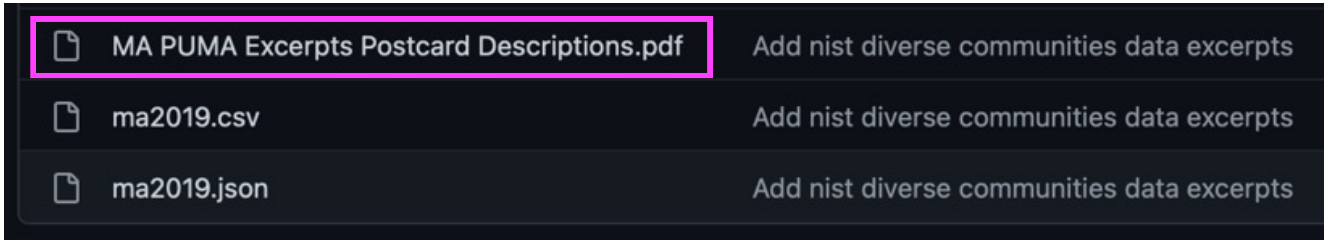
Effectively, we attempt to provide a model-sized version of the problem that we were trying to understand. Like a ship in a bottle—pieces small enough that they can be assembled by hand—so that we can start to see how they fit together.

The requirements we sought to meet included the following:

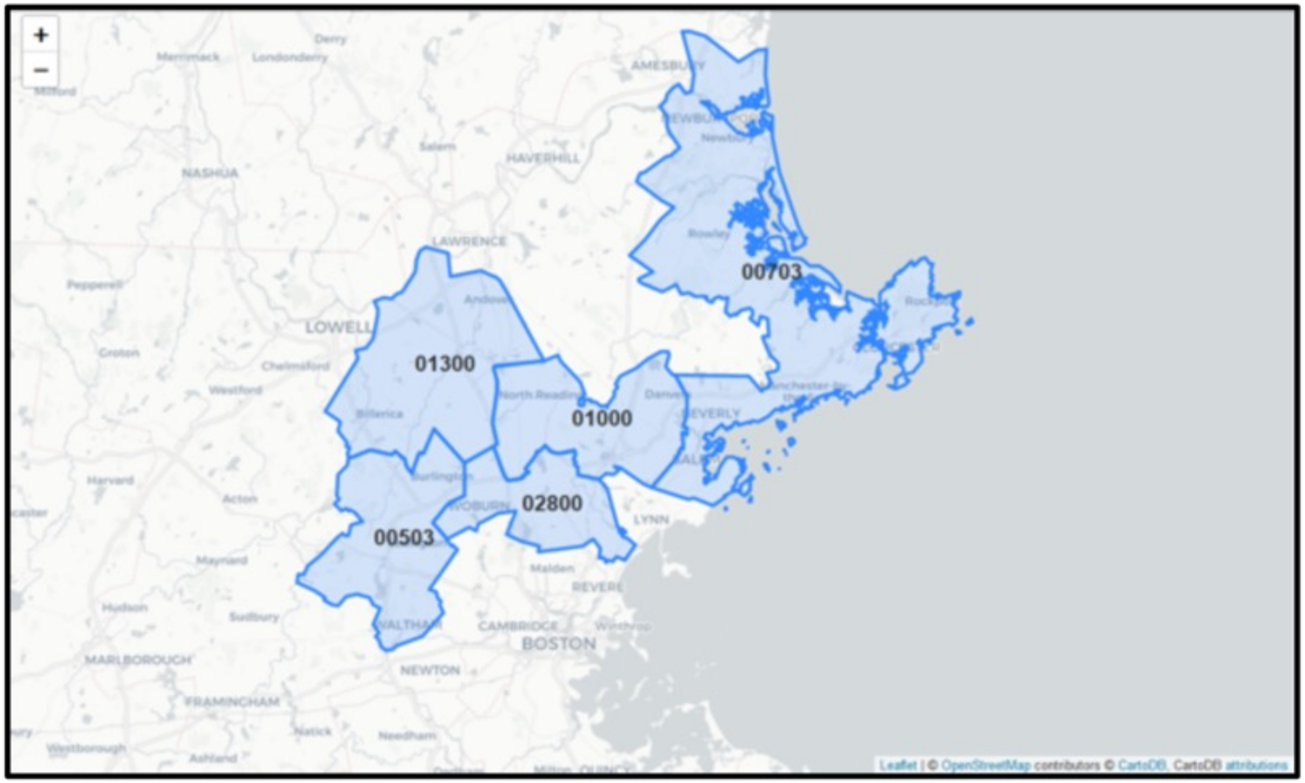
- A feature set that is interesting, with complexity representative of real communities yet compact enough to allow for rigorous, penetrating analysis

Important Stuff: Data

- It’s easy to get at the data, to try it yourself.
- There’s three data sets. MA (North Boston) is less diverse, TX (West Dallas) is more diverse, and NATIONAL are 20 of the hardest geographies we could find from around the country – they were used to help select the final version of the Census DAS algorithm.
- There’s a .json data dictionary and detailed usage guidance. We make configuring easy.
- There’s some really pretty documentation that I and DJ Streat spent a long time making, and you should go appreciate it.

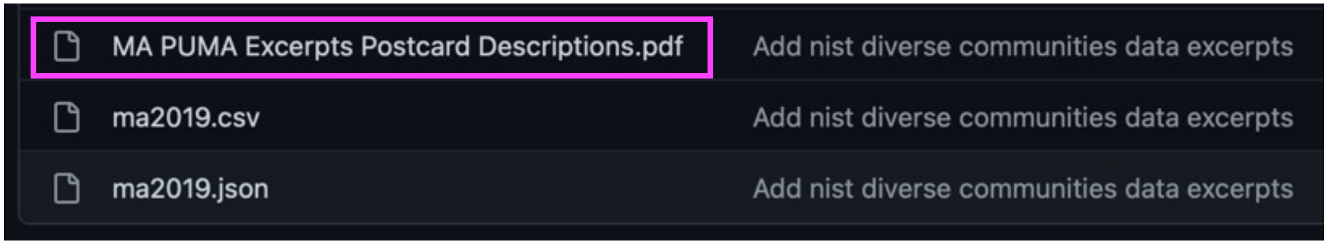


MA PUMA Excerpts Postcard Descriptions.pdf	Add nist diverse communities data excerpts
ma2019.csv	Add nist diverse communities data excerpts
ma2019.json	Add nist diverse communities data excerpts



Important Stuff: Data

- It’s easy to get at the data, to try it yourself.
- There’s three data sets. MA (North Boston) is less diverse, TX (West Dallas) is more diverse, and NATIONAL are 20 of the hardest geographies we could find from around the country – they were used to help select the final version of the Census DAS algorithm.
- There’s a .json data dictionary and detailed usage guidance. We make configuring easy.
- There’s some really pretty documentation that I and DJ Streat spent a long time making, and you should go appreciate it.



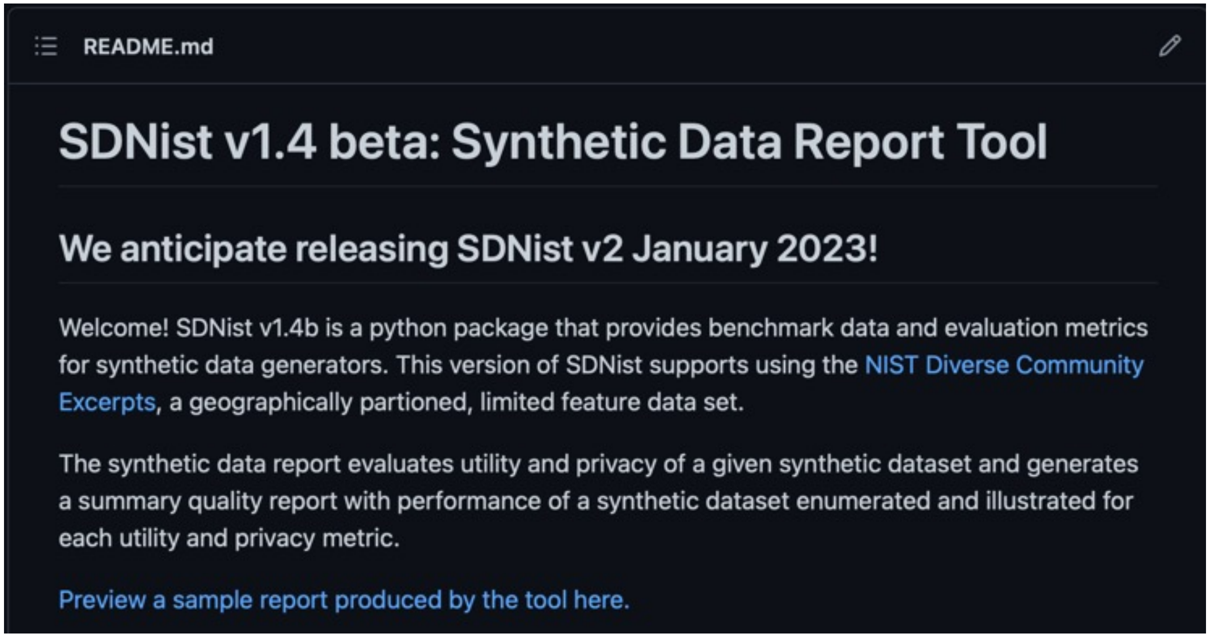
MA PUMA Excerpts Postcard Descriptions.pdf	Add nist diverse communities data excerpts
ma2019.csv	Add nist diverse communities data excerpts
ma2019.json	Add nist diverse communities data excerpts

25-2800: Woburn, Melrose Cities, Saugus, Wakefield & Stoneham Towns

Main Street at the Stoneham Theatre - [1]

Important Stuff: Report

- It’s easy to run the report tool yourself if you want. The readme gives step-by-step installation instructions.



README.md

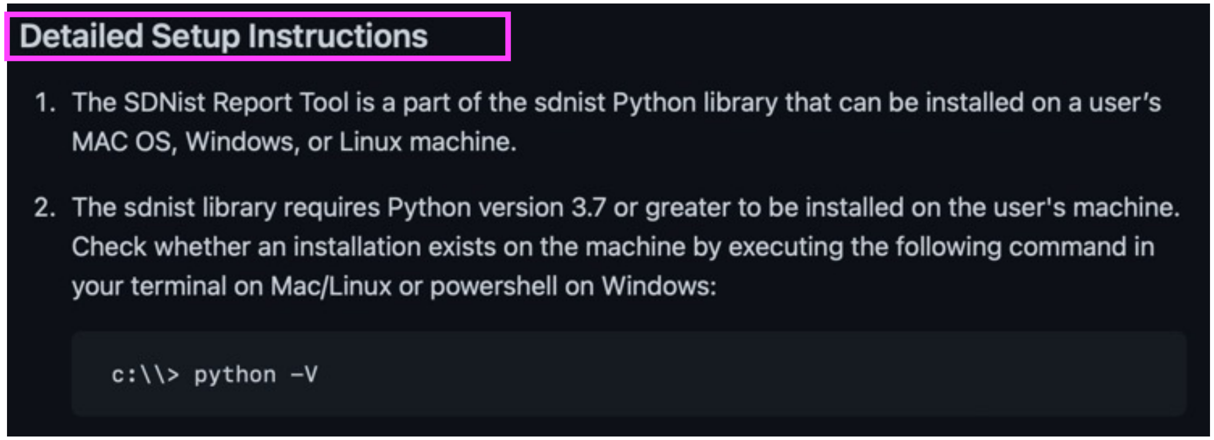
SDNist v1.4 beta: Synthetic Data Report Tool

We anticipate releasing SDNist v2 January 2023!

Welcome! SDNist v1.4b is a python package that provides benchmark data and evaluation metrics for synthetic data generators. This version of SDNist supports using the NIST Diverse Community Excerpts, a geographically partioned, limited feature data set.

The synthetic data report evaluates utility and privacy of a given synthetic dataset and generates a summary quality report with performance of a synthetic dataset enumerated and illustrated for each utility and privacy metric.

Preview a sample report produced by the tool here.



Detailed Setup Instructions

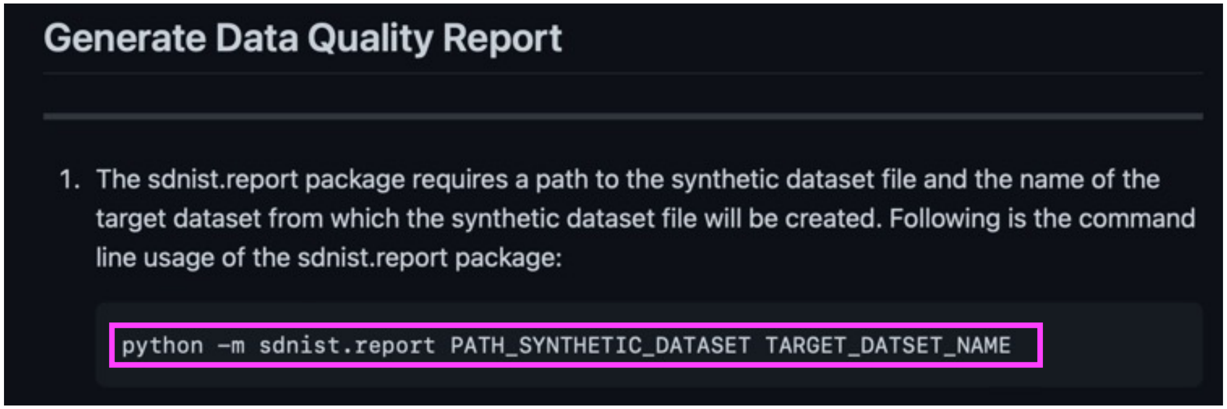
1. The SDNist Report Tool is a part of the sdnist Python library that can be installed on a user’s MAC OS, Windows, or Linux machine.
2. The sdnist library requires Python version 3.7 or greater to be installed on the user's machine. Check whether an installation exists on the machine by executing the following command in your terminal on Mac/Linux or powershell on Windows:

```
c:\\> python -V
```

Important Stuff: Report

- It’s easy to run the report tool yourself if you want. The readme gives step-by-step installation instructions.
- To run you just tell it the path to your deidentified data and whether you want to use “MA”, “TX” or “NATIONAL” as the ground truth target data.
- It yells at you (politely) if your data isn’t in the correct schema–that’s easy to correct.

Google “SDNIST”



Generate Data Quality Report

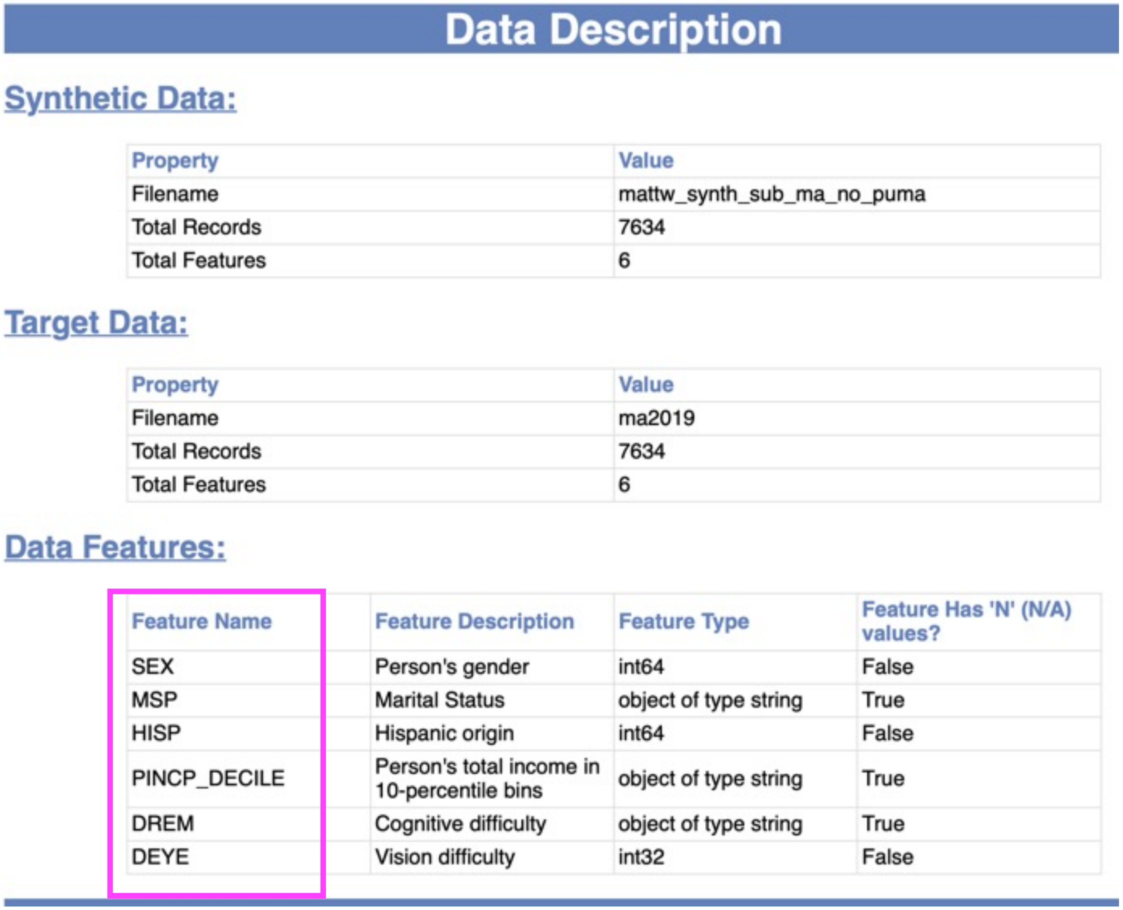
1. The sdnist.report package requires a path to the synthetic dataset file and the name of the target dataset from which the synthetic dataset file will be created. Following is the command line usage of the sdnist.report package:

```
python -m sdnist.report PATH_SYNTHETIC_DATASET TARGET_DATSET_NAME
```

Important Stuff: Report

- It's easy to run the report tool yourself if you want. The readme gives step-by-step installation instructions.
- To run you just tell it the path to your deidentified data and whether you want to use "MA", "TX" or "NATIONAL" as the ground truth target data.
- It yells at you (politely) if your data isn't in the correct schema–that's easy to correct.
- You can run on any subset of the features. Really! *(slightly more true for version 2.0)

Google "SDNIST"



Data Description

Synthetic Data:

Property	Value
Filename	mattw_synth_sub_ma_no_puma
Total Records	7634
Total Features	6

Target Data:

Property	Value
Filename	ma2019
Total Records	7634
Total Features	6

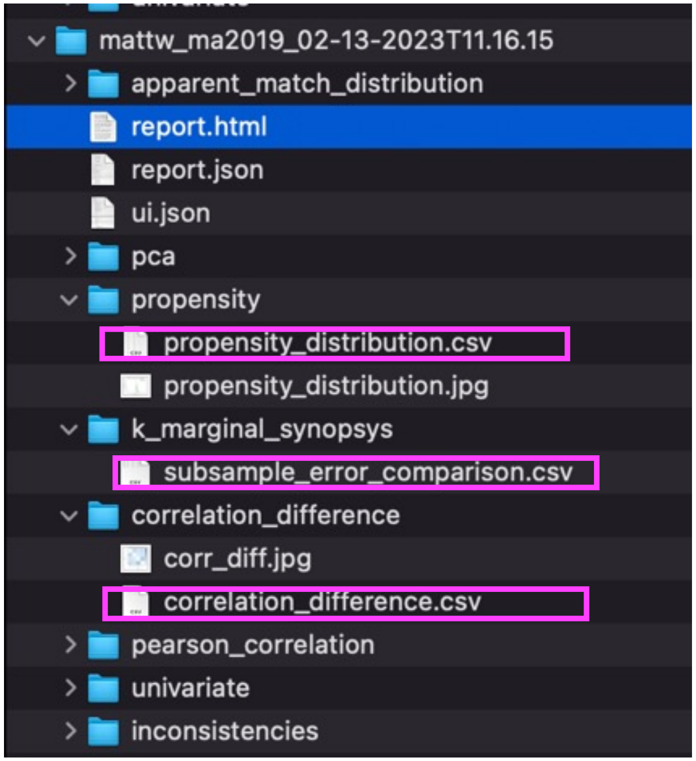
Data Features:

Feature Name	Feature Description	Feature Type	Feature Has 'N' (N/A) values?
SEX	Person's gender	int64	False
MSP	Marital Status	object of type string	True
HISP	Hispanic origin	int64	False
PINCP_DECILE	Person's total income in 10-percentile bins	object of type string	True
DREM	Cognitive difficulty	object of type string	True
DEYE	Vision difficulty	int32	False

Important Stuff: Report

- It's easy to run the report tool yourself if you want. The readme gives step-by-step installation instructions.
- To run you just tell it the path to your deidentified data and whether you want to use "MA", "TX" or "NATIONAL" as the ground truth target data.
- It yells at you (politely) if your data isn't in the correct schema–that's easy to correct.
- You can run on any subset of the features. Really!   *(slightly more true for version 2.0)
- The tool produces both machine-readable csv results and a human-readable html report.

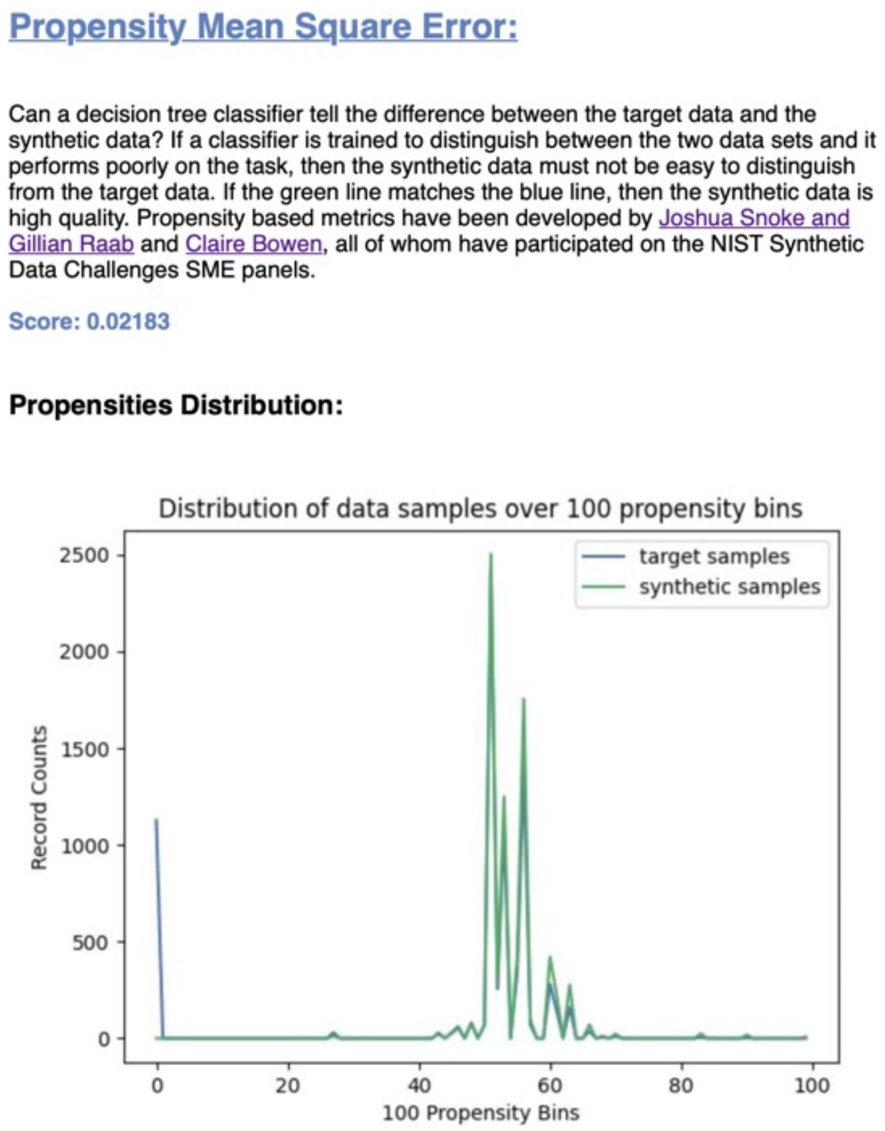
Google "SDNIST"



```
⌄ mattw_ma2019_02-13-2023T11.16.15
    › apparent_match_distribution
      report.html
      report.json
      ui.json
    › pca
    ⌄ propensity
        propensity_distribution.csv
        propensity_distribution.jpg
    ⌄ k_marginal_synopsys
        subsample_error_comparison.csv
    ⌄ correlation_difference
        corr_diff.jpg
        correlation_difference.csv
    › pearson_correlation
    › univariate
    › inconsistencies
```

Important Stuff: Report

- It’s easy to run the report tool yourself if you want. The readme gives step-by-step installation instructions.
- To run you just tell it the path to your deidentified data and whether you want to use “MA”, “TX” or “NATIONAL” as the ground truth target data.
- It yells at you (politely) if your data isn’t in the correct schema–that’s easy to correct.
- You can run on any subset of the features. Really! *(slightly more true for version 2.0)
- The tool produces both machine-readable csv results and a human-readable html report. The report includes citations, explanations for the metrics



Propensity Mean Square Error:

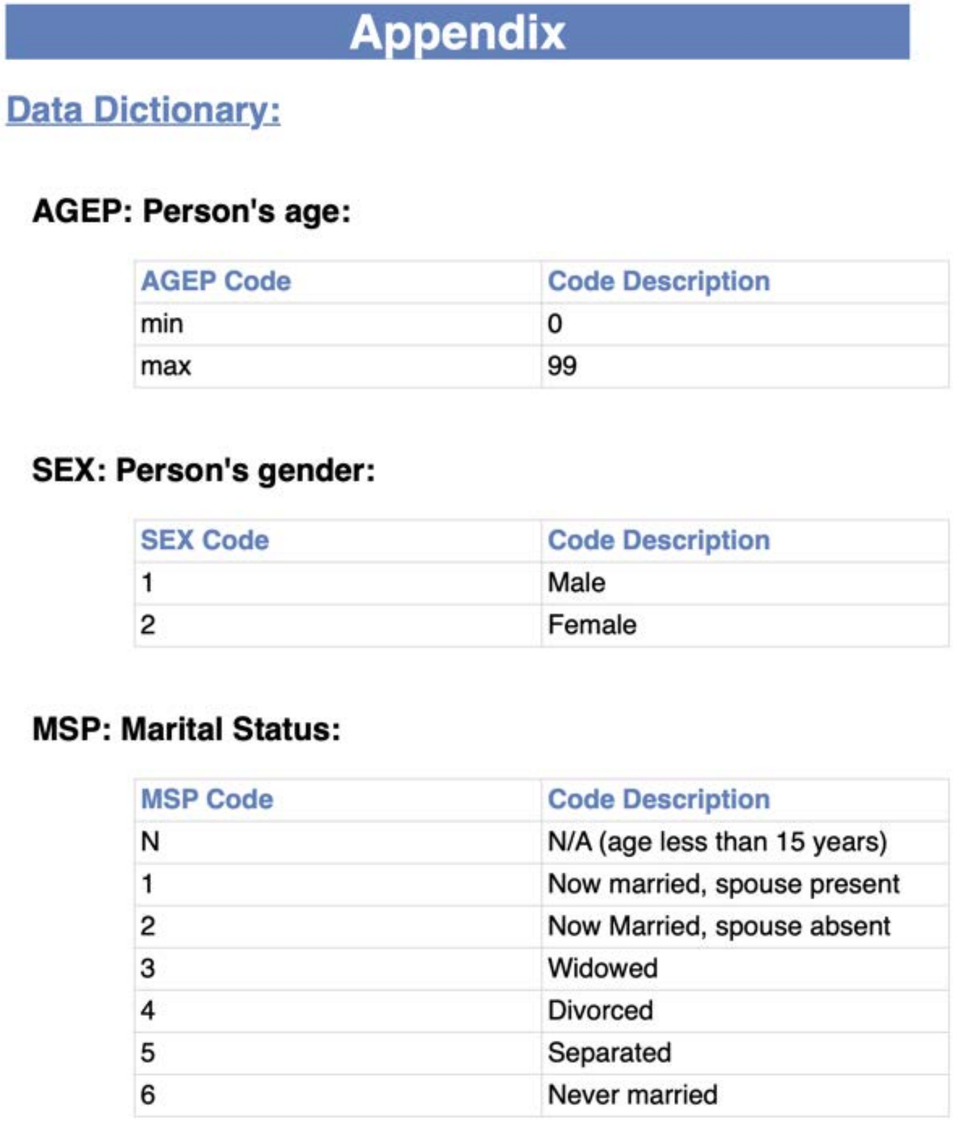
Can a decision tree classifier tell the difference between the target data and the synthetic data? If a classifier is trained to distinguish between the two data sets and it performs poorly on the task, then the synthetic data must not be easy to distinguish from the target data. If the green line matches the blue line, then the synthetic data is high quality. Propensity based metrics have been developed by Joshua Snoke and Gillian Raab and Claire Bowen, all of whom have participated on the NIST Synthetic Data Challenges SME panels.

Score: 0.02183

Propensities Distribution:

Important Stuff: Report

- It's easy to run the report tool yourself if you want. The readme gives step-by-step installation instructions.
- To run you just tell it the path to your deidentified data and whether you want to use "MA", "TX" or "NATIONAL" as the ground truth target data.
- It yells at you (politely) if your data isn't in the correct schema–that's easy to correct.
- You can run on any subset of the features. Really! *(slightly more true for version 2.0)
- The tool produces both machine-readable csv results and a human-readable html report. The report includes citations, explanations for the metrics, and a data dictionary for easy reference
- Version 2.0 will be out 2/20, including the new regression and consistency metrics.



Appendix

Data Dictionary:

AGEP: Person's age:

AGEP Code	Code Description
min	0
max	99

SEX: Person's gender:

SEX Code	Code Description
1	Male
2	Female

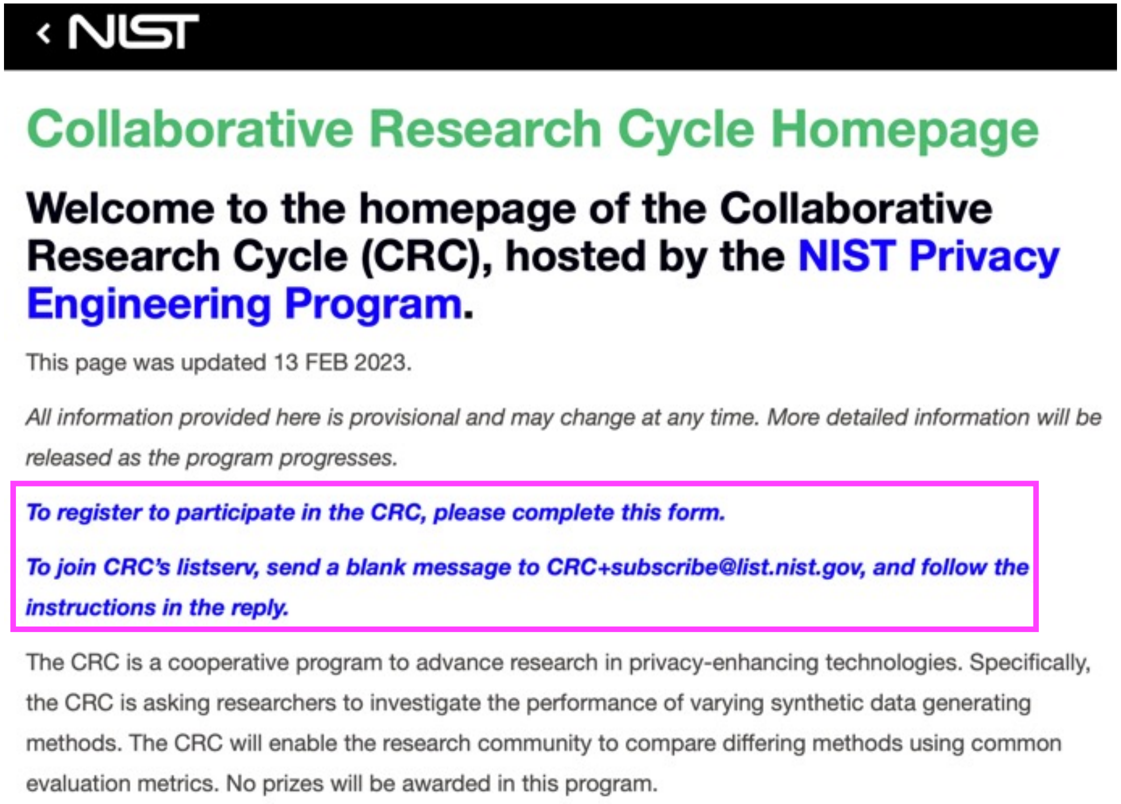
MSP: Marital Status:

MSP Code	Code Description
N	N/A (age less than 15 years)
1	Now married, spouse present
2	Now Married, spouse absent
3	Widowed
4	Divorced
5	Separated
6	Never married

Important Stuff: Register Your Team, Submit Data

- We're launching now!
- Google "Collaborative Research Cycle". I made that phrase up, so we've got great SEO at the moment.
- There's a link to sign up for our newsletter, we'll have blogposts and updates and you're welcome to follow along even if you're not participating.
- But there's also a link to register your team, and it would be great if you'd participate.
- The data submission site will be live soon (expected 2/20). Right now we're collecting data and algorithms. Submit early and often, as many times as you like.
- They don't have to be good algorithms! We want to study everything.

Google "Collaborative Research Cycle"



Collaborative Research Cycle Homepage

Welcome to the homepage of the Collaborative Research Cycle (CRC), hosted by the NIST Privacy Engineering Program.

This page was updated 13 FEB 2023.

All information provided here is provisional and may change at any time. More detailed information will be released as the program progresses.

To register to participate in the CRC, please complete this form.

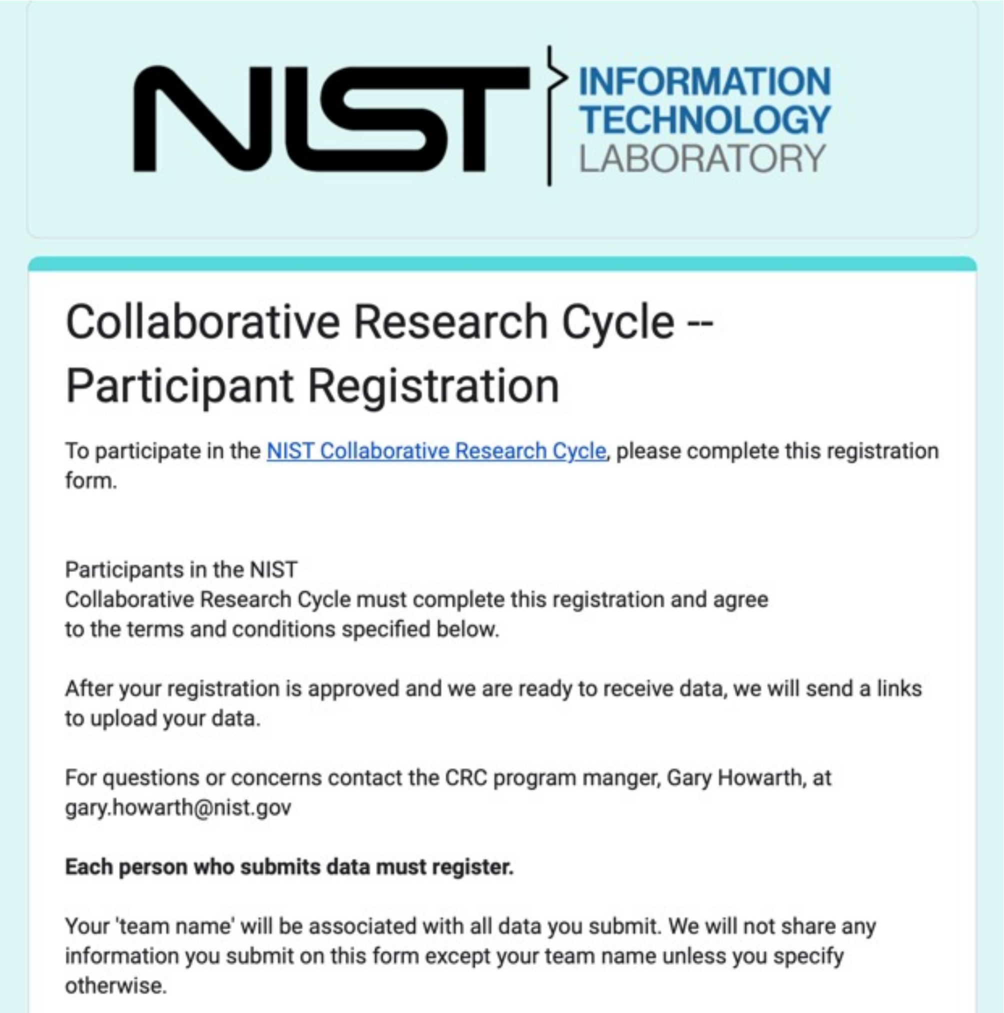
To join CRC's listserv, send a blank message to CRC+subscribe@list.nist.gov, and follow the instructions in the reply.

The CRC is a cooperative program to advance research in privacy-enhancing technologies. Specifically, the CRC is asking researchers to investigate the performance of varying synthetic data generating methods. The CRC will enable the research community to compare differing methods using common evaluation metrics. No prizes will be awarded in this program.

Important Stuff: Register Your Team, Submit Data

- We’re launching now!
- Google “**Collaborative Research Cycle**”. I made that phrase up, so we’ve got great SEO at the moment.
- There’s a link to sign up for our newsletter, we’ll have blogposts and updates and you’re welcome to follow along even if you’re not participating.
- But there’s also a link to register your team, and it would be great if you’d participate.
- The data submission site will be live soon (expected 2/20). Right now we’re collecting data and algorithms. Submit early and often, as many times as you like.
- They don’t have to be good algorithms! We want to study everything.

Google “Collaborative Research Cycle”



Collaborative Research Cycle -- Participant Registration

To participate in the NIST Collaborative Research Cycle, please complete this registration form.

Participants in the NIST
Collaborative Research Cycle must complete this registration and agree
to the terms and conditions specified below.

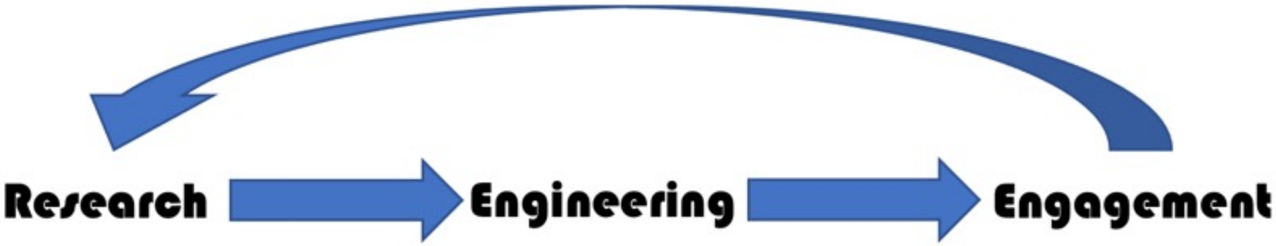
After your registration is approved and we are ready to receive data, we will send a links to upload your data.

For questions or concerns contact the CRC program manger, Gary Howarth, at gary.howarth@nist.gov

Each person who submits data must register.

Your 'team name' will be associated with all data you submit. We will not share any information you submit on this form except your team name unless you specify otherwise.

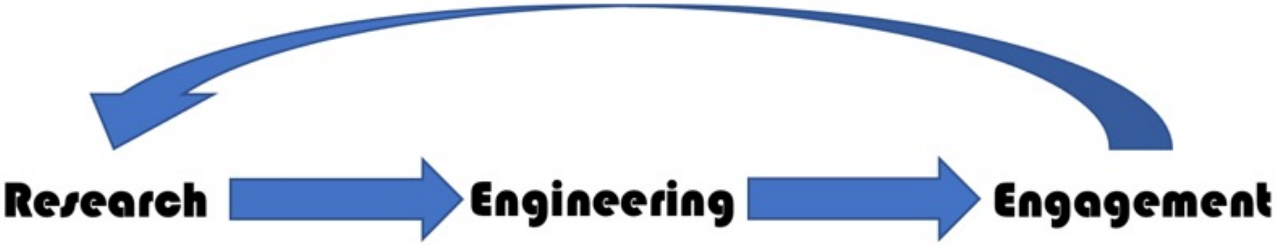
Important Stuff: Timeline



You Are Here

Important Stuff: Timeline

May - September:
Find a neat open problem and look into it.
We'll be sharing our own work all summer
with blogs, new tools and invited SME talks.



Research → Engineering → Engagement

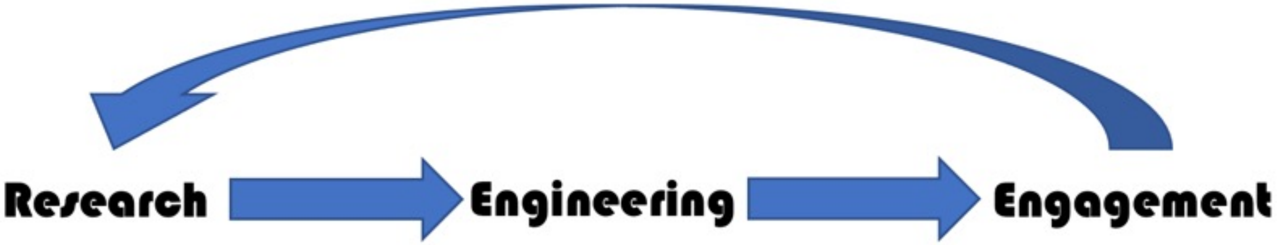
You Are
Here

Feb – April:
Configure your/any
approach for our
data and submit

Early May:
We release a Research
Acceleration Bundle of
everyone's approaches,
data and metric results.

Important Stuff: Timeline

May - September:
Find a neat open problem and look into it. We'll be sharing our own work all summer with blogs, new tools and invited SME talks.



Research → Engineering → Engagement

Late September:
Submit to our tiny paper (3pg + appendix) workshop in November. Incremental results are fine! You'll have time to extend it for submission to PPAI next year.

Feb – April:
Configure your/any approach for our data and submit

Early May:
We release a Research Acceleration Bundle of everyone's approaches, data and metric results.

Questions? Email me at Christine.task@knexusresearch.cm
Or Gary Howarth at Gary.Howarth@nist.gov