Privacy and Public Records: Perils and possible solutions of releasing public safety records

December 6, 2023

NIST PUBLIC SAFETY COMMUNICATIONS RESEARCH

Disclaimer

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.

AGENDA

1 Introduction

2 Quick Poll

3 Presentation

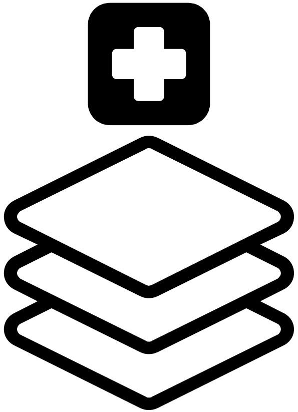
4 Live Q&A

Gary S. Howarth, II

- Physical Scientist and project manager for the Public Safety Communications Research Division (CTL) and the Privacy Engineering Program (ITL).

NIST PUBLIC SAFETY COMMUNICATIONS RESEARCH

Power in Data

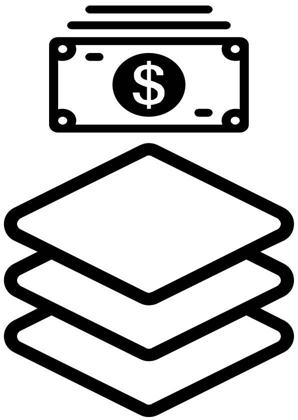


Medical Records

Analysis

- Model outcomes
- Identify risk factors
- Distinguish subpopulations

Power in Data

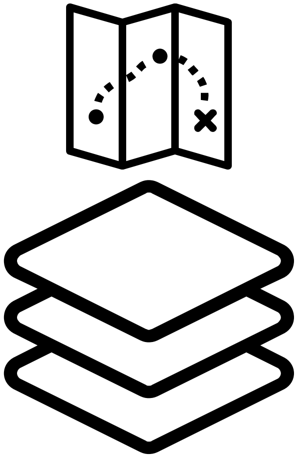


Financial Records

Analysis →

- Find waste and fraud
- Predict policy outcomes
- Identify opportunities

Power in Data



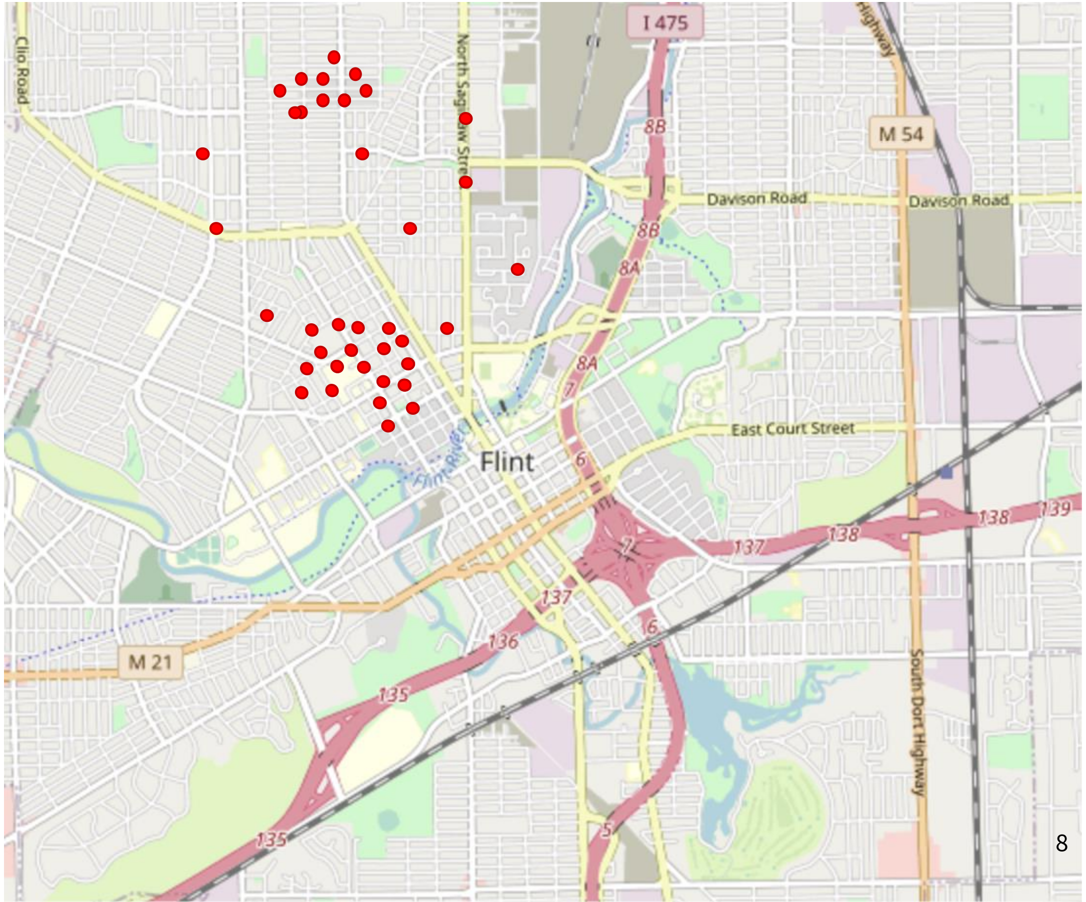
Location Data

Analysis →

- Infrastructure planning
- Optimize services
- Traffic routing

Power in Data

911 Calls for Service

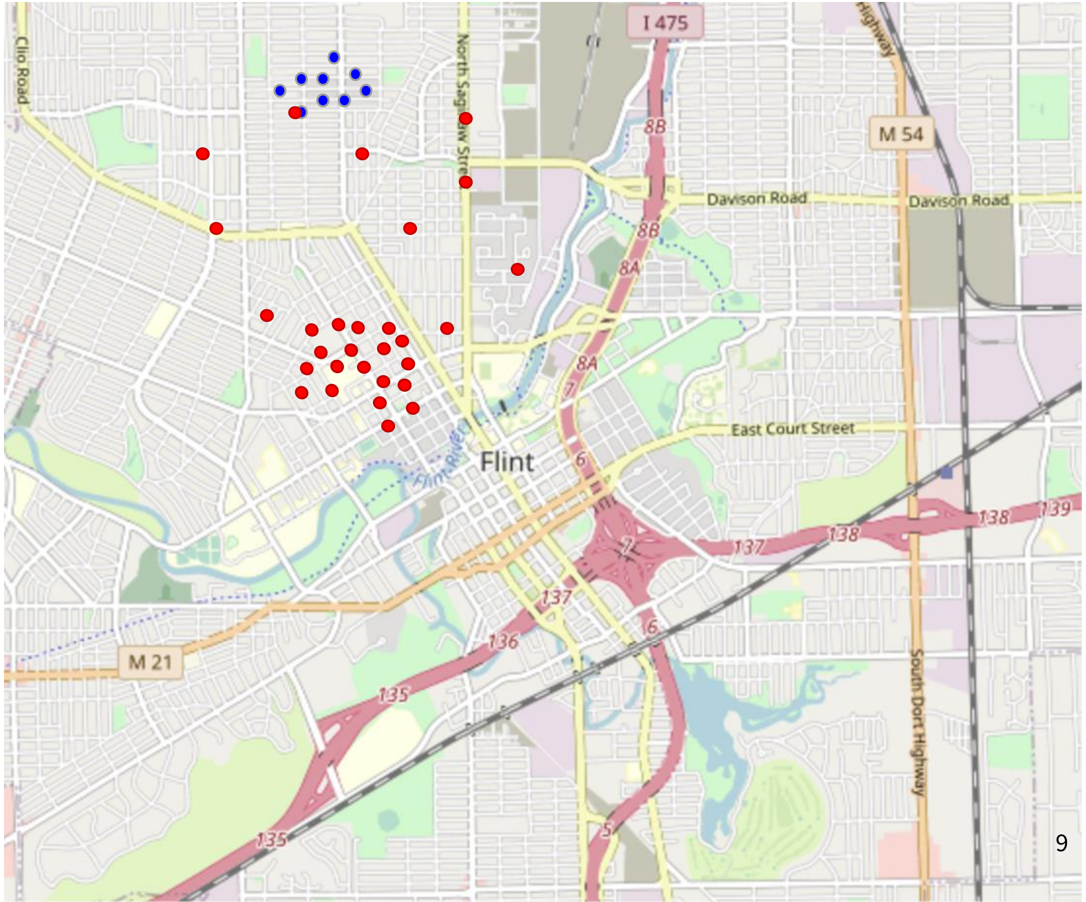


Power in Data

NIST

911 Calls for Service

- Parse by water quality-related calls



Protecting PII with Redaction?

Record Number	Name	DOB	Sex	Address	Date of Visit	Reason for Visit
132313	[redacted]	[redacted]/1979	Male	[redacted] Northampton, MA 01129	[redacted]/1997	Suicide attempt
318977	[redacted]	[redacted]/1992	Female	[redacted] Springfield, MA 01020	[redacted]/1997	Lead poisoning
218987	[redacted]z	[redacted]/1994	Female	[redacted] Springfield, MA 01020	[redacted]/1997	Lead poisoning
156465	[redacted]	[redacted]/1949	Male	[redacted] Cambridge, MA 03129	[redacted]/1997	Back pain

Can you protect PII with redaction?

Redacted data is vulnerable to *de-anonymization* attacks with auxiliary data sources

<u>Redacted Medical Record</u>

Diagnosis
Procedure
Medications
Year of visit
Ethnicity

Zip code
Birth year
Sex

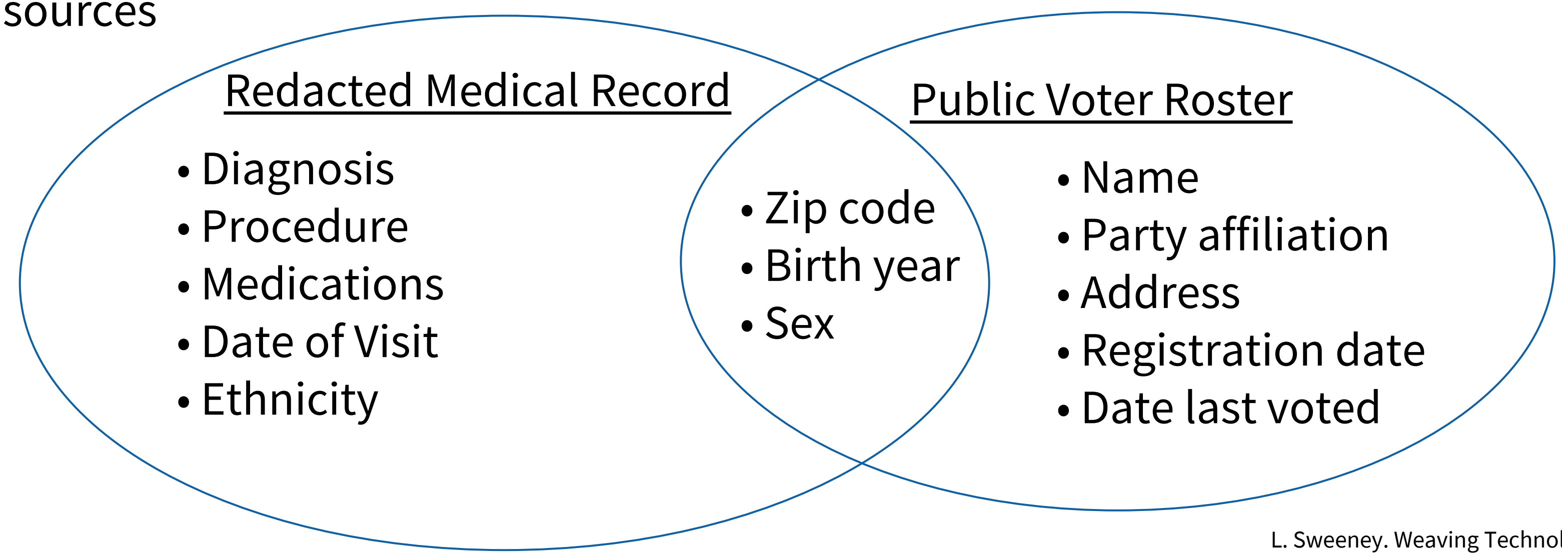
<u>Public Voter Roster</u>

Name
Party affiliation
Address
Registration date
Date last voted

Zip code
Birth year
Sex

Protecting PII with Redaction?

Redacted data is vulnerable to *de-anonymization* attacks with auxiliary data sources



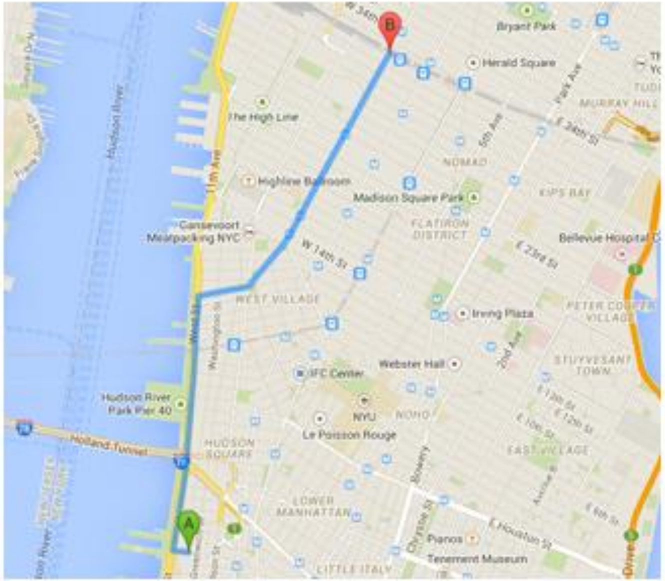
87% of people in U.S. can be re-identified using 3 quasi-identifiers.

L. Sweeney. Weaving Technology and Policy Together to Maintain Confidentiality. *Journal of Law, Medicine & Ethics*, 25, nos. 2&3 (1997): 98-110.

L. Sweeney. *Simple Demographics Often Identify People Uniquely*. Carnegie Mellon University, Data Privacy Working Paper 3. Pittsburgh 2000.

Know when and where what taxi was entered?

KATHERINE HEIGL



OCTOBER 4, 2013 • 1:21 PM - 1:40 PM
80 N. MOORE ST. TO 421 8TH AVE
$14.50 FARE • $3.62 TIP • ©WENN

J. Trotter. *Public NYC Taxicab Database Lets You See How Celebrities Tip.* Gawker. 14 Oct 2014.

Know when and where what taxi was entered?

NIST

Taxi Database

- Pickup and Dropoff Address
- Pickup and Dropoff Time
- Intermediate stops
- Fare
- Tip (some of the time)

Taxi Database ∩ Paparazzo Photo

- Pickup OR Dropoff address
- Partial taxi number
- Time and date

Paparazzo Photo

- Rider(s)
- Taxi number

Traditional disclosure control

Donor list of the Llama Freedom Foundation

Name	Age	Town	Income	Ethnicity or race	Religion
~~Bertram Wooster~~	31	NYC	900k	White	Anglican
~~Francis Hu~~	39	NYC	40k	Asian	none
~~Ollie McOld~~	~~119~~	~~NYC~~	~~60k~~	~~Black~~	~~Baptist~~
~~Lela Fox~~	~~44~~	~~NYC~~	~~70k~~	~~Aust. Aboriginal + Uyghur~~	~~Islam~~
~~Mohammed Abas~~	~~55~~	~~Tiny (pop. 20)~~	~~250k~~	~~Arab~~	~~Sunni~~
~~Bill Kirkland~~	45	Tiny (pop. 20)	100k	White	Baptist

Aggregate Metric	Original	Redacted
Mean Age	**55.5**	38.3
Mean Income	**237k**	346k

Anonymization is not enough

Practically all information is identifying.

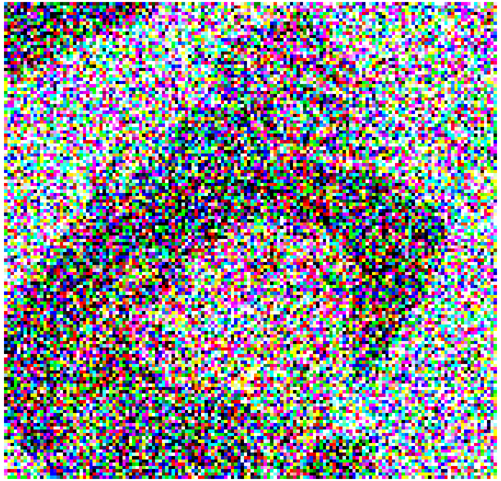
Field suppression, redaction, and *anonymization* techniques limit utility and may be highly vulnerable to attack.

Reidentification is bad

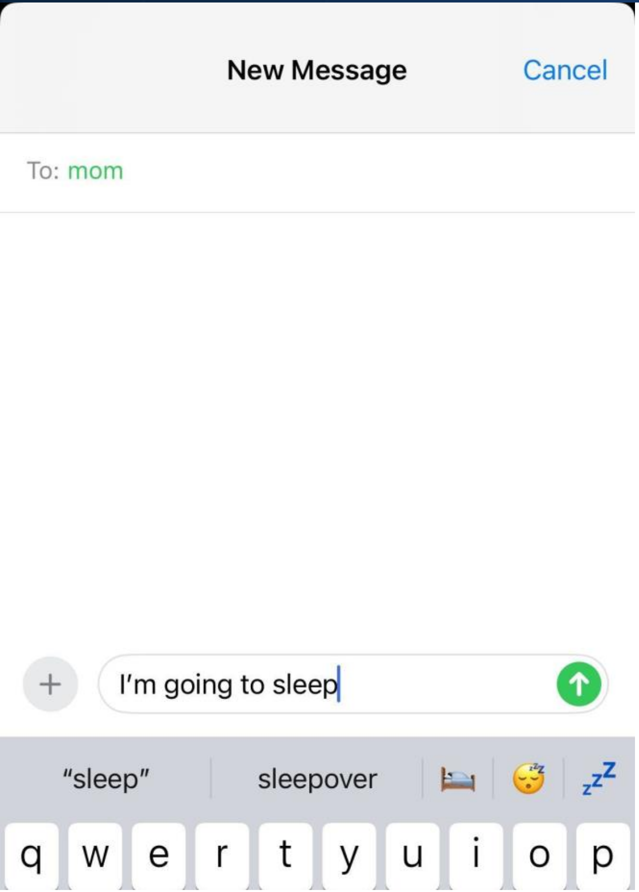
Reidentification attacks fuel:

- Discrimination, abuse, violence against minorities
- SWATing
- Predatory marketing, phishing, and cons
- Distrust of information collection programs

Adding noise to protect privacy



Adding noise to protect privacy



- Phones makes suggestions.
- Tech companies collect feedback.
- Some collections involve privacy noise.

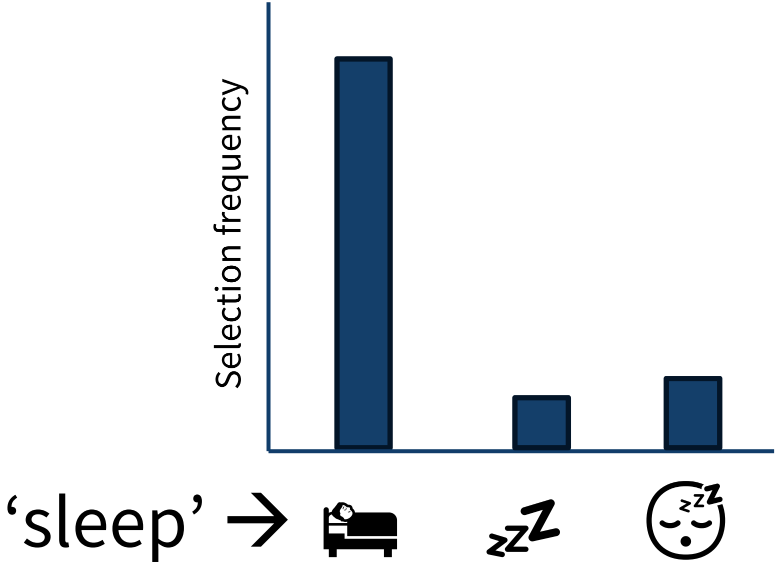
<u>Gary always chooses 🛌 to represent 'sleep.'</u>

Gary Selects	Phone transmits feedback
🛌	🛌
🛌	💤 (noise)
🛌	🛌
🛌	🛌
🛌	😴 (noise)

Sometimes the phone adds noise creating privacy (plausible deniability).

Adding noise to protect privacy

NIST

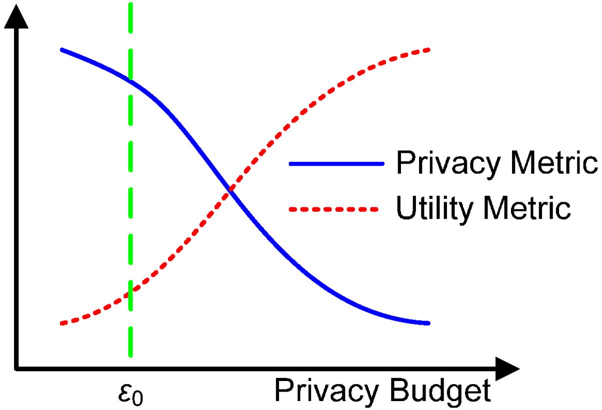


Phone provider can still analyze the noisy data for meaning

Sensitive survey examples:

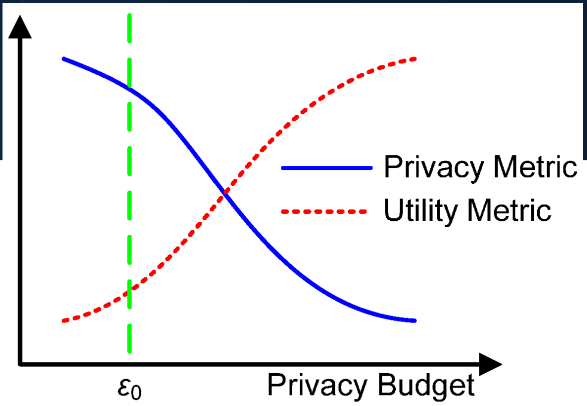
- Have you ever under-reported income on your taxes?
- What’s your HIV status?

Privacy-utility trade off

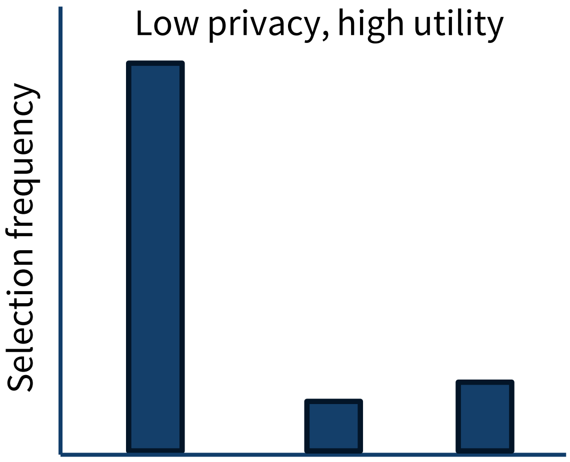


From Liu et al. “Privacy-Preserving Monotonicity of Differential Privacy Mechanisms.” 2018.

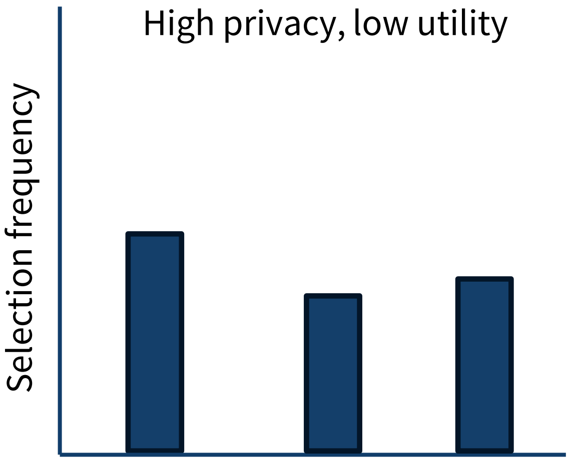
Privacy-utility trade off



Remember, Gary <u>always</u> chooses 🛌



‘sleep’ → 🛌 💤 😴



‘sleep’ → 🛌 💤 😴

Differential privacy

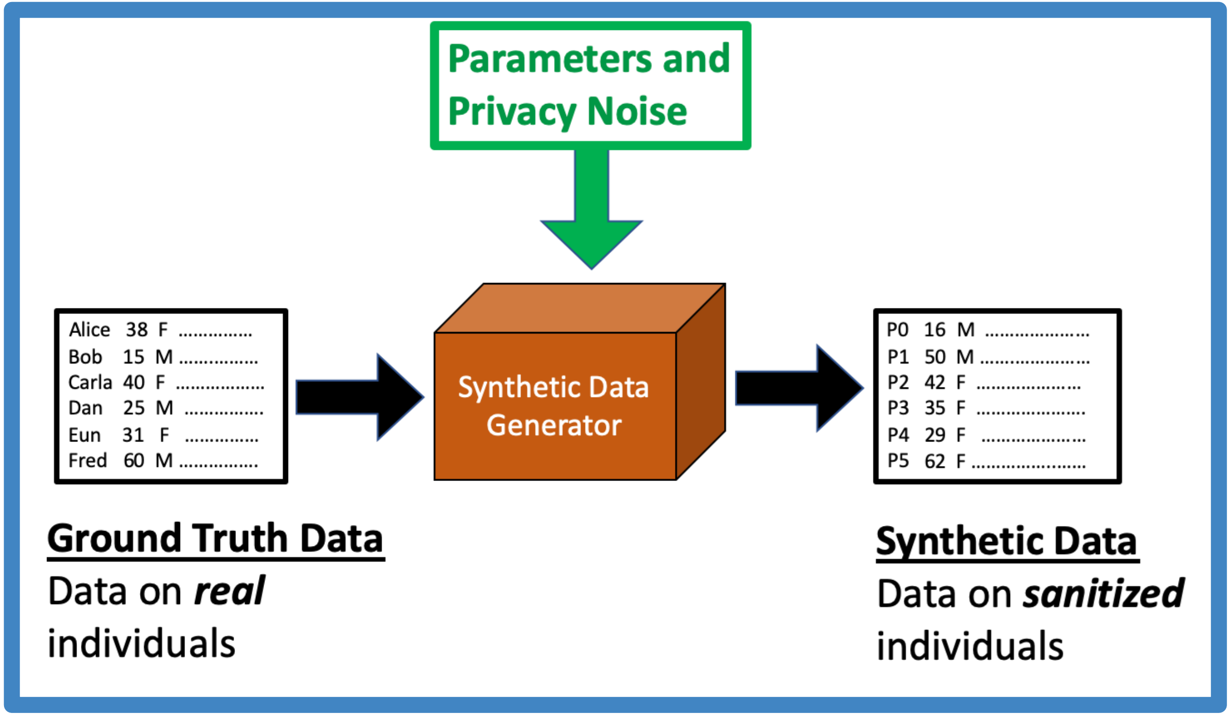
Differential privacy is:

- Rigorous mathematical definition of privacy
- A framework to add privacy noise

Differential privacy is not:

- A specific algorithm
- Silver bullet
- Bogie man

DP can be used to make synthetic records



DP can be used to make synthetic records

ORIGINAL DATA

Person	Age	Income	State
O1	24	31,000	CO
O2	88	45,000	NM
…	…	…	
O450	11	0	CO

→

DP algorithm

SYNTHETIC DATA

Person	Age	Income	State
S1	44	51,151	CO
S2	22	33,232	CO
…	…	…	
S450	35	12,223	NM

Aggregate Metric	Original	Synthetic
Mean age	44	44
Mean Income	51,231	51,244
People in CO	250	249

DP can be used to make synthetic records

ORIGINAL DATA

Person	Age	Income	State
O1	24	31,000	CO
O2	88	45,000	NM
…	…	…	
~~O450~~	~~11~~	~~0~~	~~CO~~

→

DP algorithm

SYNTHETIC DATA

Person	Age	Income	State
S1	43	51,845	NM
S2	22	31,412	NM
…	…	…	
S499	19	21,121	CO

Differential privacy limits how much can be learned about an individual in the data.

DP can be used to make synthetic records

ORIGINAL DATA

Person	Age	Income	State
O1	24	31,000	CO
O2	88	45,000	NM
…	…	…	
~~O450~~	~~11~~	~~0~~	~~CO~~

→ DP algorithm

SYNTHETIC DATA

Person	Age	Income	State
S1	43	51,845	NM
S2	22	31,412	NM
…	…	…	
S499	19	21,121	CO

	Original		Synthetic	
Metric	All data	~~O450~~	All data	~~O450~~
Mean age	44	45	44	44
Mean Income	51,231	51,345	51,244	51,243
People in CO	252	251	249	249

DP is tunable for privacy

Smaller ε
More noise
More privacy
Less accuracy

Larger ε
Less noise
Less privacy
More accuracy

Data Publishing mandates

<u>*Case study:*</u>

U.S. Census Bureau is mandated to make accurate counts of people

(U.S. Constitution Article I, Section 2)

U.S. Census Bureau is required by law to protect respondent confidentiality at every stage of the data lifecycle with *criminal penalties* for violations

(U.S. Code 13 § 8-9 / 141)

“Differential privacy is the best science available to protect 2020 Census respondent confidentiality while minimizing the impact on statistical validity.”[1]

1. *Disclosure Avoidance and the 2020 Census Redistricting Data, U.S. Census Bureau*

Big Questions of Differential Privacy?

- What types of data can we successfully de-identify?
- How much noise must we add?
- Are the noisy data still useful / accurate?
- Are the output data actually private?
- Are the noisy data accurate for all subgroups in the data?

NIST Innovates: 2019 Synthetic Data Challenge

NIST

Goal: De-identify records from San Francisco Calls for Service portal.

NIST gave Competitors:

- training data
- basic, ‘baseline’ algorithm
- scoring methodology
- public leaderboard

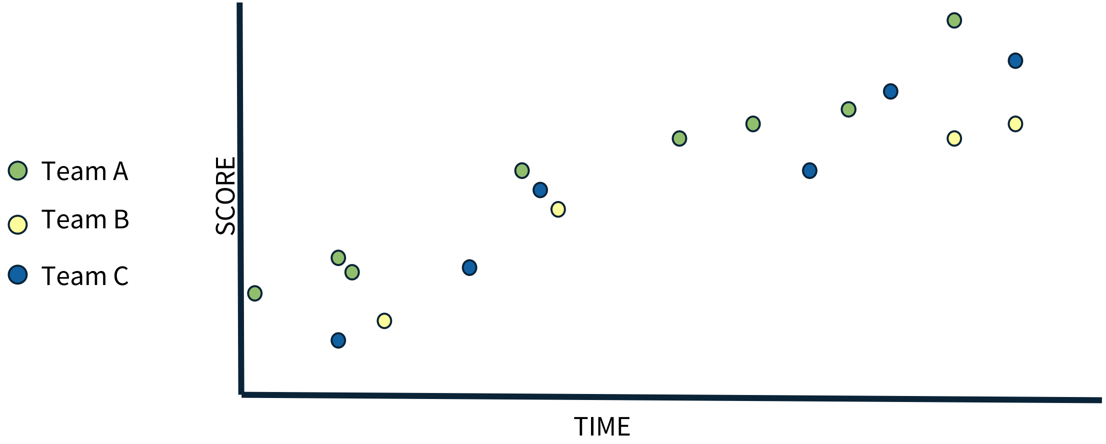
Competitors gave NIST:

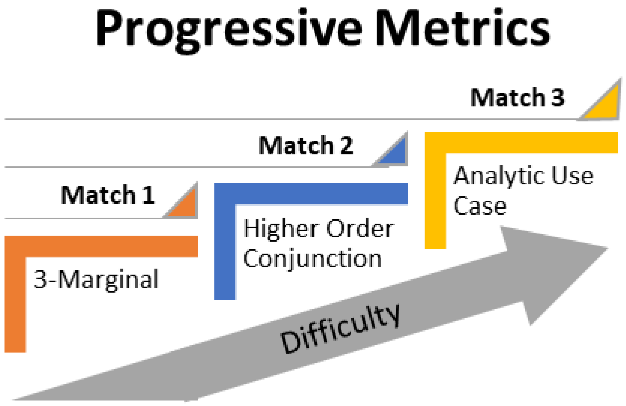
- deidentified data
- new, innovative algorithms
- mathematical proofs their algorithms were DP

https://data.sfgov.org/Public-Safety/Fire-Department-Calls-for-Service/

NIST Innovates: 2019 Synthetic Data Challenge

Public leaderboard within a match
(simulated example)





NIST Innovates: 2019 Synthetic Data Challenge

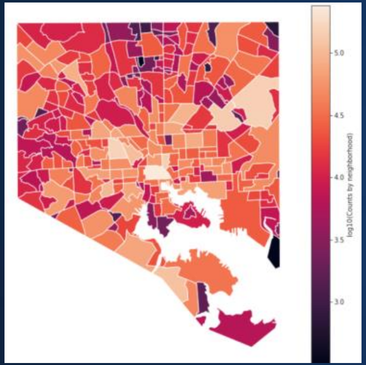
NIST

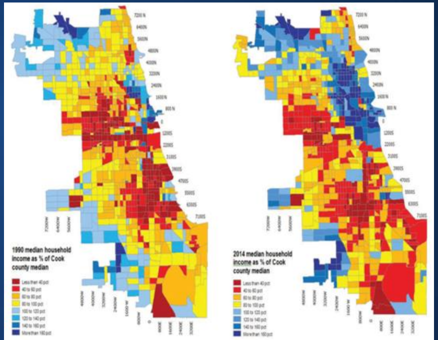
Concept Paper	Match #1	Match #2	Match #3	Open Source
1st place $15k 2nd place $10k 3rd place $5k People's Choice 2 x $5k	1st place $10k 2nd place $7k 3rd place $5k 4th place $2k 5th place $1k Progressive 4 x $1k	1st place $15k 2nd place $10k 3rd place $5k 4th place $3k 5th place $2k Progressive 4 x $1k	1st place $25k 2nd place $15k 3rd place $10k 4th place $5k 5th place $3k Progressive 4 x $1k	Additional $4k/team
$40k	**$29k**	**$39k**	**$62k**	**$20k**

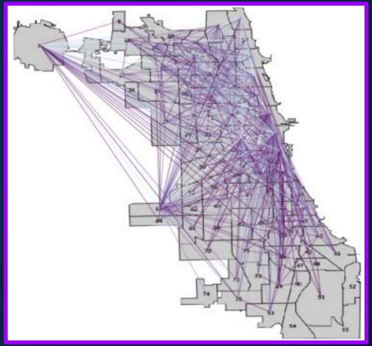
Acknowledgements:
- Terese Manley, NIST PSCR, Prize Manager
- Christine Task, Knexus Research, Technical Lead

https://doi.org/10.6028/NIST.TN.2151

NIST Innovates: 2020 Temporal Map Challenge







Sprint 1

Baltimore 911 Incidents

- Highly variable PS data
- Training data: 2019
- Evaluation data: 2016 & 2020

Sprint 2

American Community Survey (US Census)

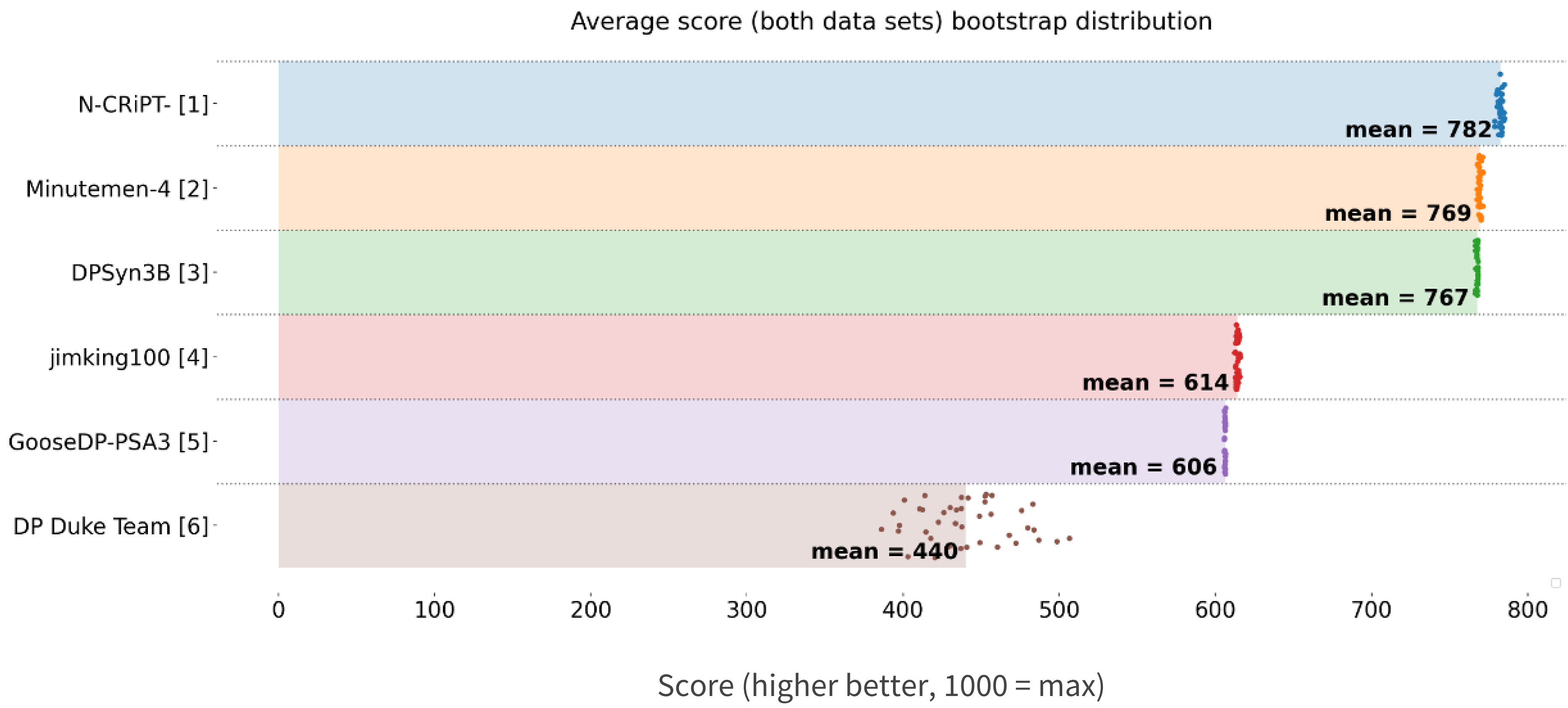
- Complex demographic information
- Training data: IL + OH
- Evaluation data: NY + PA & NC+SC+GA

Sprint 3

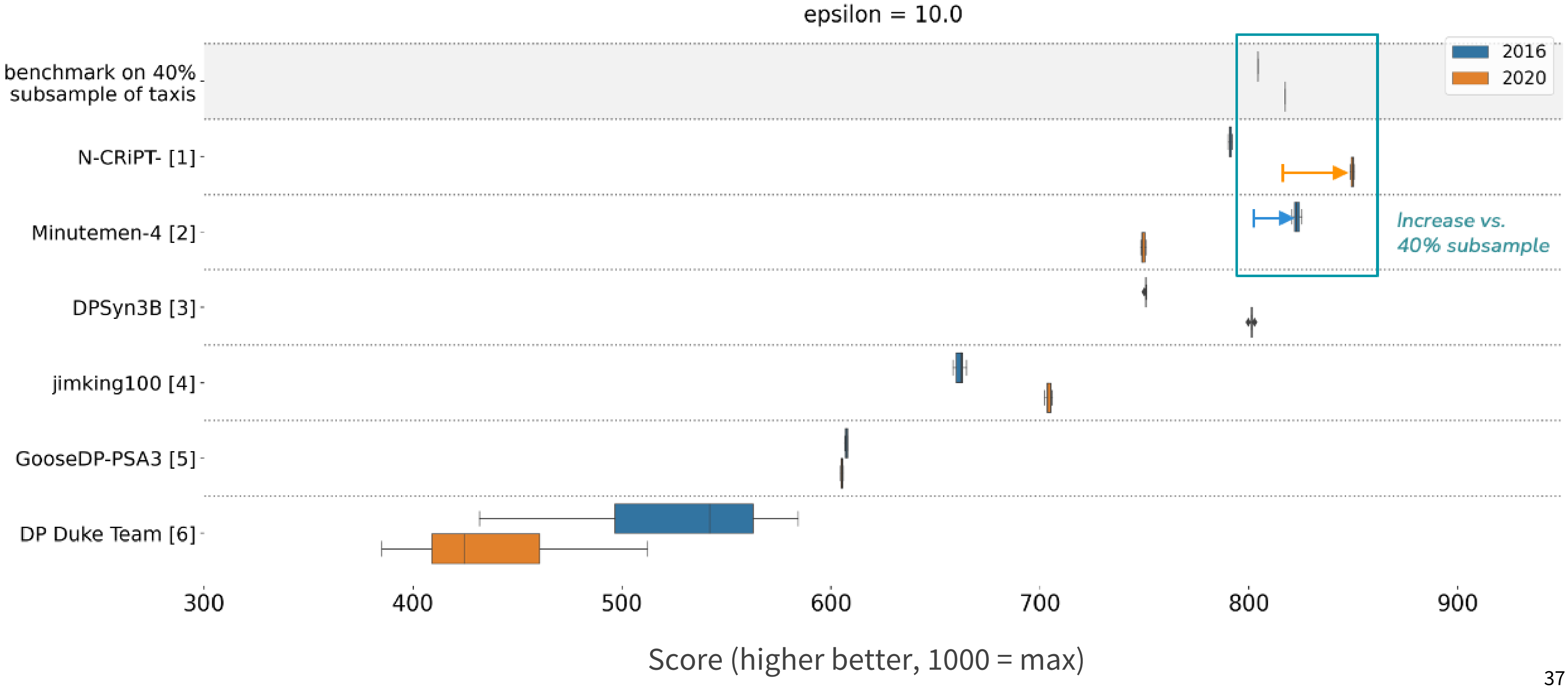
Chicago Taxi Rides

- Linked trip information
- Training data: 2019
- Evaluation data: 2016 & 2020

Temporal Map Challenge Outcomes



Temporal Map Challenge Outcomes



Temporal Map Challenge Outcomes

NIST

- About +
- Research Portfolios +
- Funding Opportunities +
- **Open Innovation Prize Challenges** −
  - Current and Upcoming Prize Challenges +
  - **Past Prize Challenges** −
    - 2021 Mobile Fingerprinting Innovation Technology Challenge
    - 2021 First Responder UAS Triple Challenge +
    - 2020 CHARIoT Challenge
    - 2020 First Responder UAS Endurance Challenge
    - 2020 Enhancing Computer Vision for Public Safety Challenge
    - 2020 Automated Stream Analysis for Public Safety Challenge
    - **2020 Differential Privacy Temporal Map Challenge**

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.

Differential Privacy Temporal Map Challenge

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

The challenge was highly successful with more than 70 unique algorithms submissions across all three sprints of the challenge. Four of those algorithms have been open sourced (links in winners table below). Three solutions participated in the Development Contest, where teams were coached by NIST experts to improve the robustness and documentation of their code, creating easy-to-use implementations of sophisticated differential privacy algorithms.

The challenge was implemented by DrivenData with assistance from HeroX . Christine Task from Knexus Research Corporation served as the program's technical lead. Gary Howarth served as the prize manager.

"NIST temporal map challenge"

Acknowledgements

- Dr. Christine Task, Knexus Research, Technical Lead
- John Garofolo, NIST ITL, Portfolio Lead
- DrivenData and HeroX

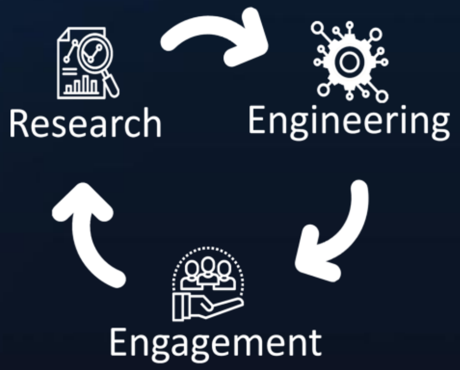
Collaborative Research Cycle (CRC)

NIST privacy prize challenges have:

- Provided essential proof-of-concept experiments
- Accelerated practical synthetic data generating techniques
- Expanded the audience for and consumers of differential privacy

NIST CRC seeks to:

- Expand the scope and breadth of synthetic data evaluations
- Compare different algorithms on the same underlying data
- Provide a venue for cooperation



The Diverse Communities Excerpt Data

NIST

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

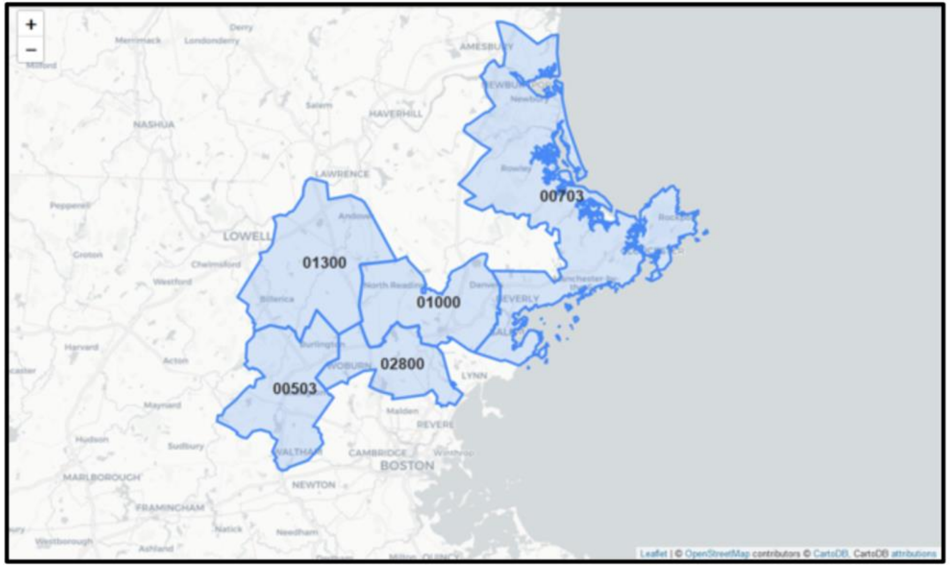
Feature Name	Feature Description
INDP	Industry codes
INDP_CAT	Industry categories
EDU	Educational attainment
PINCP	Person's total income in dollars
PINCP_DECILE	Person's total income in 10-percentile bins
POVPIP	Income-to-poverty ratio (ex: 250 = 2.5 x poverty line)
DVET	Veteran service connected disability rating (percentage)
DREM	Cognitive difficulty
DPHY	Ambulatory (walking) difficulty
DEYE	Vision difficulty
DEAR	Hearing difficulty

The Diverse Communities Excerpt Data

Data PUMA and Postcard Descriptions:

Massachusetts Dataset

Postcard Descriptions



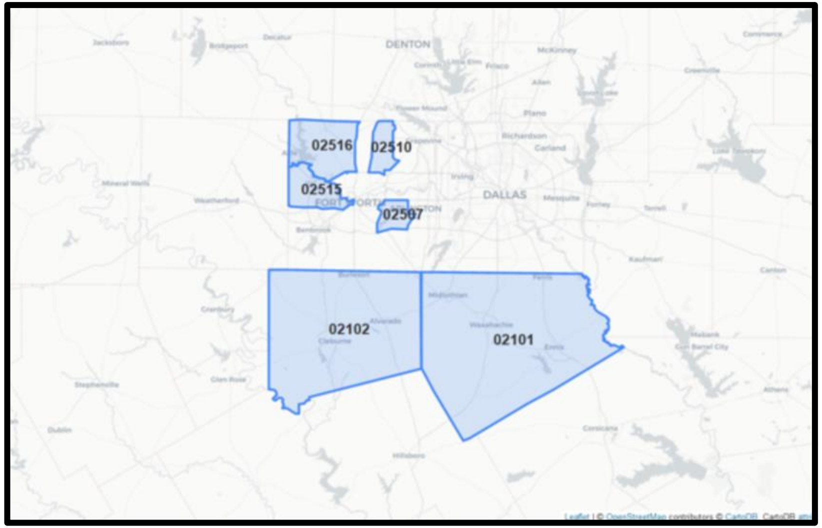
These PUMA from North and East of Boston, Massachusetts include suburbs that began as small towns in the 17th century, historically working-class neighborhoods, historically wealthy neighborhoods, and rapidly growing newer communities connected to the tech industry.

The Diverse Communities Excerpt Data

Data PUMA and Postcard Descriptions:

Texas Dataset

Postcard Descriptions



These PUMA from South and West of Fort Worth Texas include a selection of urban, suburban and rural communities—some communities predate Texas joining the United States. Their economies draw from a wide variety of sectors including agriculture, industry, military, business, and entertainment (museums, theme parks, sports). Railroads, and then highways, have played a major role in how these communities have grown.

The Diverse Communities Excerpt Data

Data PUMA and Postcard Descriptions:

National Dataset Postcard Descriptions

PUMA
36-03710: NYC-Bronx Community District 1 & 2--Hunts Point, Longwood & Melrose
06-07502: San Francisco County (North & East)--North Beach & Chinatown
26-02702: Washtenaw County (East Central)--Ann Arbor City Area
06-08507: Santa Clara County (Southwest)--Cupertino, Saratoga Cities & Los Gatos Town
32-00405: Las Vegas City (Southeast)
51-01301: Arlington County (North)
01-01301: Birmingham City (West)
30-00600: East Montana (Outside Billings City)
24-01004: Montgomery County (South)--Bethesda, Potomac & North Bethesda

40-00200: Cherokee, Sequoyah & Adair Counties
13-04600: Atlanta Regional Commission--Fulton County (Central)--Atlanta City (Central)
29-01901: St. Louis City (North)
08-00803: Boulder County (Central)--Boulder City
17-03529: Chicago City (South)--South Shore, Hyde Park, Woodlawn, Grand Boulevard & Douglas
38-00100: West North Dakota--Minot City
19-01700: Des Moines City
51-51255: Alexandria City
17-03531: Chicago City (South)--Auburn Gresham, Roseland, Chatham, Avalon Park & Burnside
36-04010: NYC-Brooklyn Community District 17--East Flatbush, Farragut & Rugby
28-01100: Central Region--Jackson City (East & Central)

The SDNist Evaluator

Data Evaluation Report

Data Description

Synthetic Data:

Property	Value
Filename	na_syn_b101_e4
Total Records	27188
Total Features	22

Target Data:

Property	Value
Filename	national2019
Total Records	27253
Total Features	22

The SDNist Evaluator (sdnist v2.3)

pip install sdnist

Data Evaluation Report

Data Description

Synthetic Data:

Property	Value
Filename	na_syn_b101_e4
Total Records	27188
Total Features	22

Target Data:

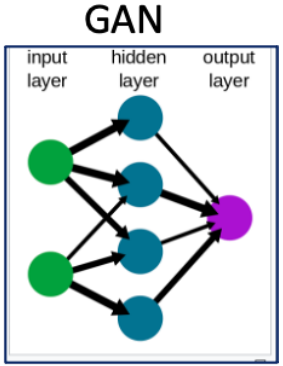
Property	Value
Filename	national2019
Total Records	27253
Total Features	22

Algorithms: A Sample of Four Deidentification Approaches

DP Histogram: Add randomized noise to counts

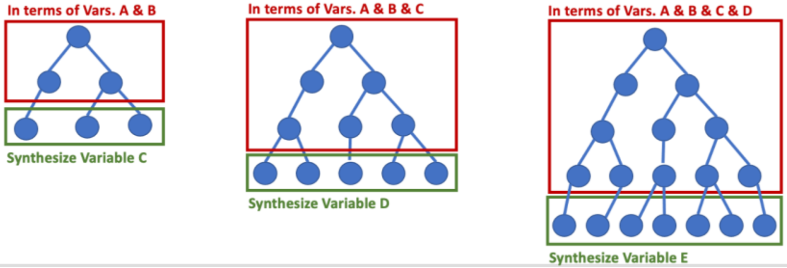
Differential Private Histogram (ε = 10)

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



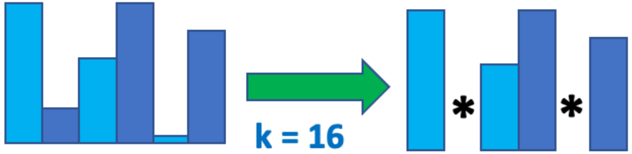
PATECTGAN Differential Private GAN (ε = 10)

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



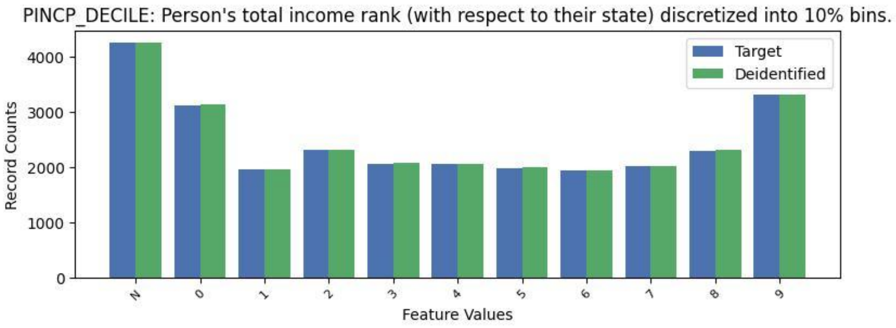
CART-model Synthesis (non-DP synthetic)

Cell Suppression: Redact small counts

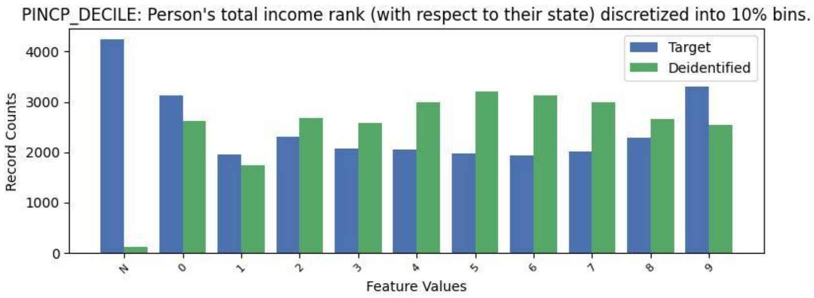


Cell Suppression (k = 6)

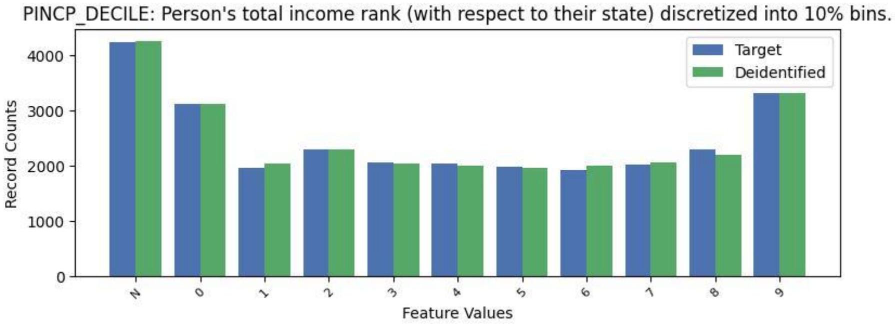
Metrics: Univariate



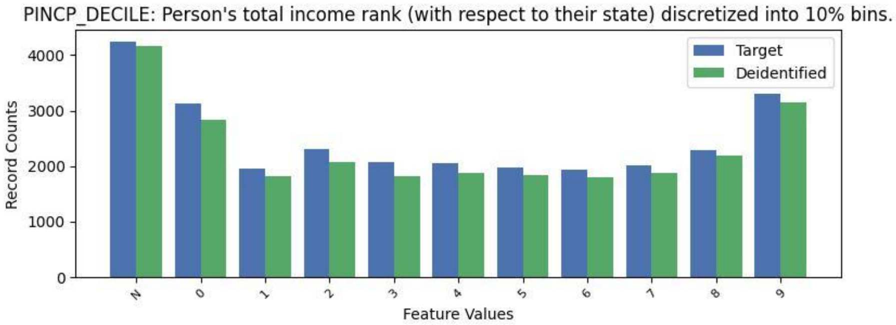
Differential Private Histogram (ε = 10)



PATECTGAN Differential Private GAN (ε = 10)

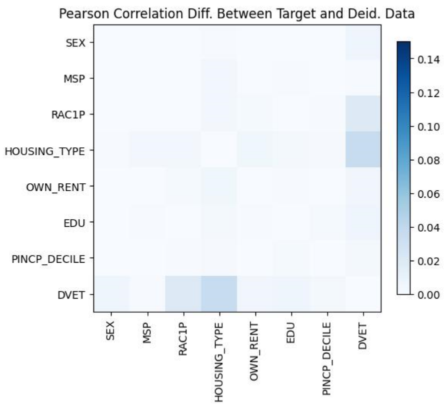


CART-model Synthesis (non-DP synthetic)

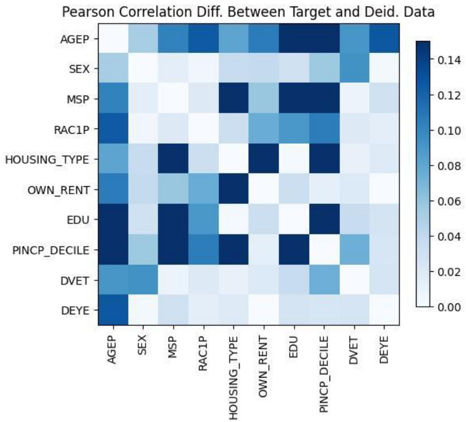


Cell Suppression (k = 6)

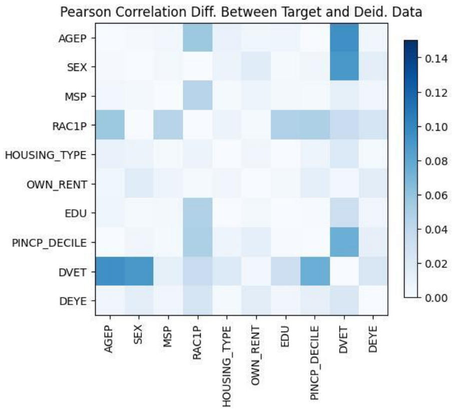
Metrics: Pairwise Correlations



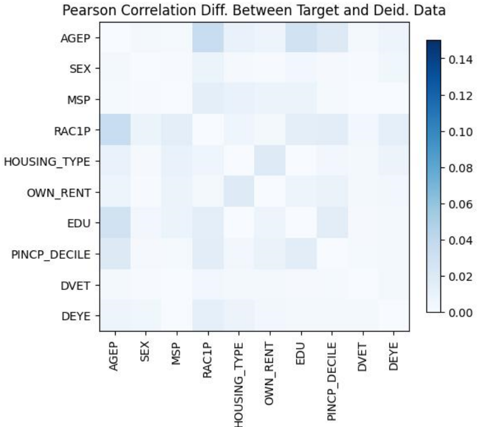
Differential Private Histogram (ε = 10)



PATECTGAN Differential Private GAN (ε = 10)



CART-model Synthesis (non-DP synthetic)

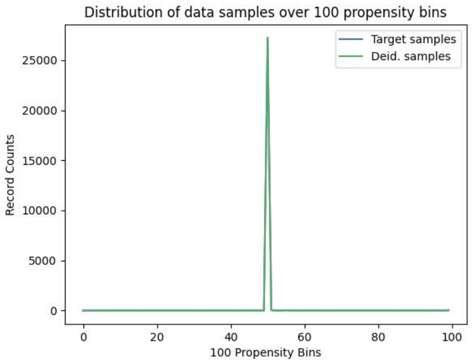


Cell Suppression (k = 6)

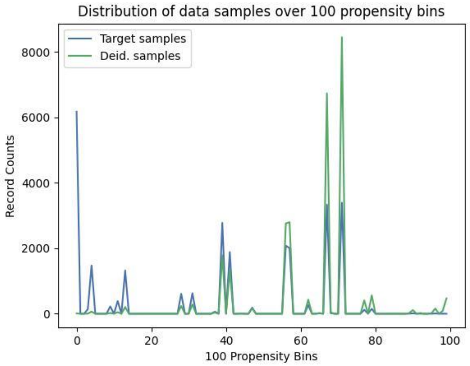
Pairwise Correlations: A key goal of deidentified data is to preserve the feature correlations from the target data, so that analyses performed on the deidentified data provide meaningful insight about the target population. Which correlations are the deidentified data preserving, and which are being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

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

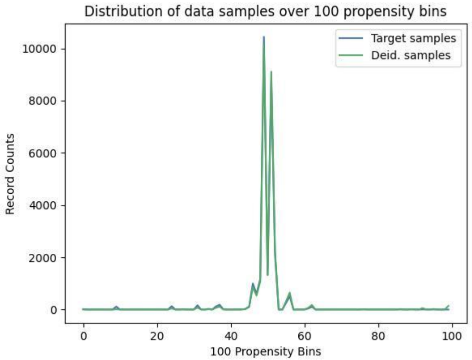
Metrics: Propensity



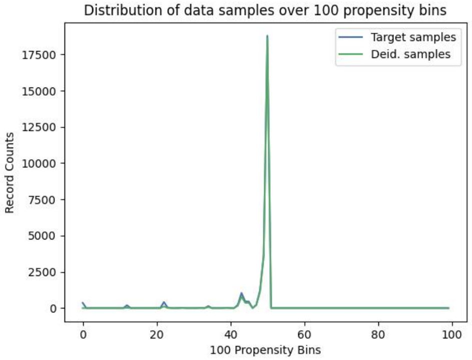
Differential Private Histogram (ε = 10)



PATECTGAN Differential Private GAN (ε = 10)



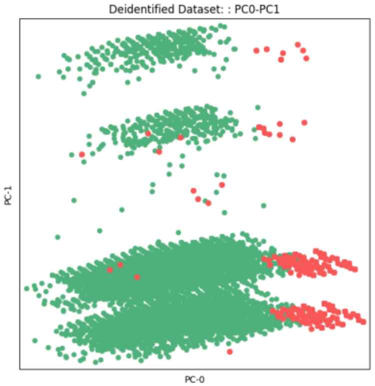
CART-model Synthesis (non-DP synthetic)



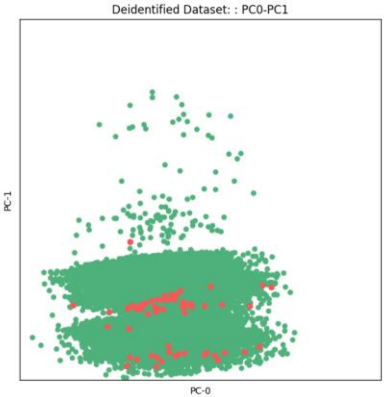
Cell Suppression (k = 6)

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

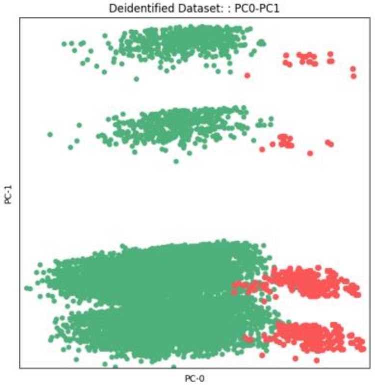
Metrics: Pairwise PCA



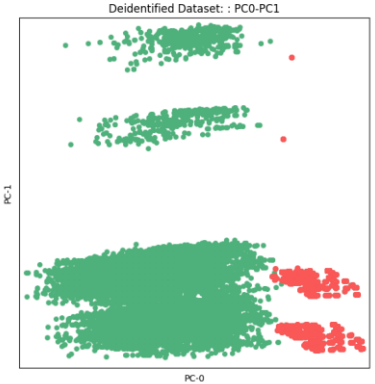
Differential Private Histogram (ε = 10)



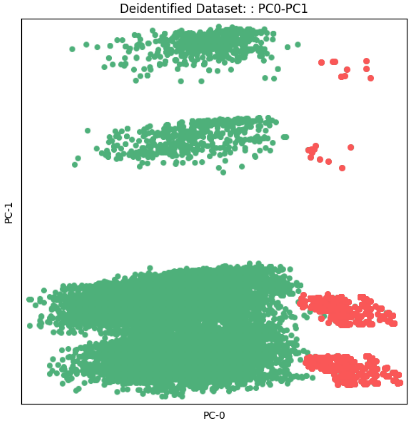
PATECTGAN Differential Private GAN (ε = 10)



CART-model Synthesis (non-DP synthetic)



Cell Suppression (k = 6)



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

Metrics: Consistency Checks

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

Differential Private Histogram (ε = 10)

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

PATECTGAN Differential Private GAN (ε = 10)

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

CART-model Synthesis (non-DP synthetic)

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

Cell Suppression (k = 6)

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

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

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

Metrics: Unique Exact Match Rate

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

Differential Private Histogram (ε = 10)

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

PATECTGAN Differential Private GAN (ε = 10)

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

CART-model Synthesis (non-DP synthetic)

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

Cell Suppression (k = 6)

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

NIST Collaborative Research Cycle: Far more than four algorithms

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

Collaborative Research Cycle

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

Resources include:

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

Contents:

Open Source:

Commercial Products:

SmartNoise MST

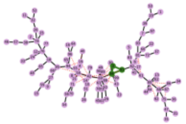
SmartNoise library implementation of MST, winner of the 2018 NIST Differential Privacy Synthetic Data Challenge. Data is generated from a differentially private PGM instantiated with noisy marginals. The structure of the PGM is a Maximum Spanning Tree (MST) capturing the most significant pair-wise feature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

Library: smartnoise-synth (Python)

Privacy: Differential Privacy

References:

- SmartNoise MST Documentation



[Mckenna 2019]

SmartNoise MWEM

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

Library: smartnoise-synth (Python)

Privacy: Differential Privacy

References:

- SmartNoise MWEM Documentation

Inputs: Data set B over a universe D,
Number of iterations $T \in \mathbb{N}$,
Privacy parameter $\varepsilon > 0$.

Let n denote $\|B\|$, the number of records in B.
Let A_0 denote n times the uniform distribution over D.
For iteration $i = 1, \ldots, T$:

1. *Exponential Mechanism:* Sample a query $q_i \in Q$ using the Exponential Mechanism parametrized with epsilon value $\varepsilon/2T$ and the score function $$s_i(B, q) = |q(A_{i-1}) - q(B)|.$$
2. *Laplace Mechanism:* Let measurement $m_i = q_i(B) + Lap(2T/\varepsilon)$.
3. *Multiplicative Weights:* Let A_i be n times the distribution whose entries satisfy $$A_i(x) \propto A_{i-1}(x) \times \exp(q_i(x) \times (m_i - q_i(A_{i-1}))/2n).$$

Output: $A = avg_{i<T} A_i$.

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

RSynthpop CART

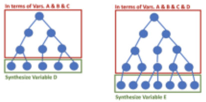
R Synthpop library implementation of fully conditional CART model-based synthesis (default syn() function). New records are generated one feature at a time, using a sequence of decision trees that select plausible new values for each feature, based on the values synthesized for previous features. Data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

Library: synthpop (R)

Privacy: Synthetic Data (Non-differentially Private)

References:

- R Synthpop Documentation



RSynthpop Catall

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

Library: synthpop (R)

Privacy: Differential Privacy

References:

- R Synthpop Catall Documentation

CRC Workshop

NIST

December 18: 10:30 AM – 2:00PM ET

- Results of CRC submissions
- Practical lessons on DP, reidentification, and other topics
- Register and see the full agenda here:

Collaborative Research Cycle

Acknowledgements

- Christine Task (Knexus)
- Karan Bhagat (Knexus)
- Aniruddha Sen (U. Mass.)
- Dhruv Kapur (U. Mich.)
- Ashley Simpson (Knexus)

NIST Differential Privacy Guidelines

NIST Special Publication
NIST SP 800-226 ipd

Guidelines for Evaluating Differential Privacy Guarantees

Authors

Joseph P. Near
University of Vermont

David Darais
Galois, Inc.

Editors
Naomi Lefkovitz
Gary Howarth
Applied Cybersecurity Division, Information Technology Laboratory, NIST

Have some data? Have some ideas?

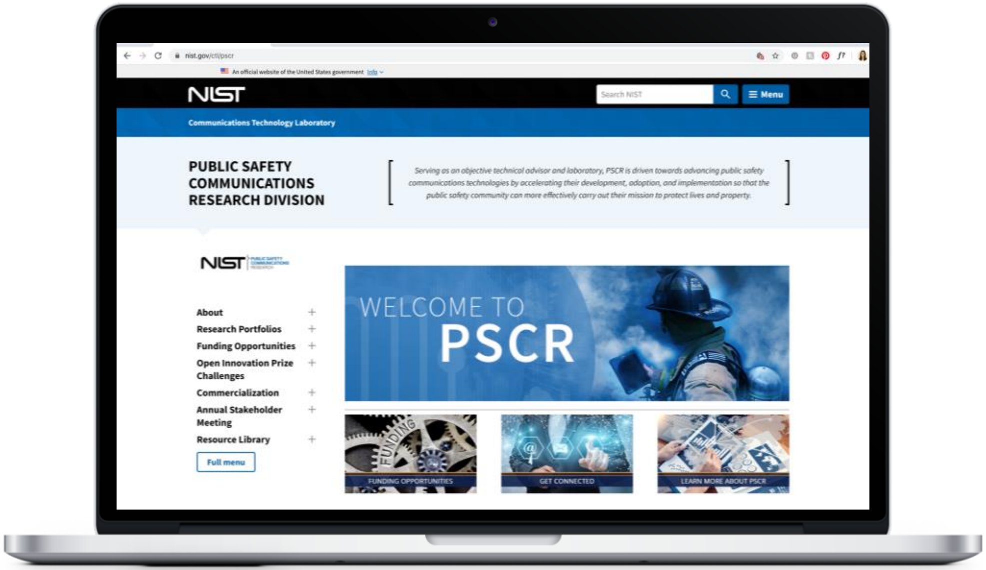
NIST

Contact me to talk about a potential pilot!

- Guidance on how to try it at home
- Internal-only sandbox to try out ideas
- Help with potential public releases

Gary Howarth
gary.howarth@nist.gov
(720)-360-9158

Resources



Contact

Gary Howarth
gary.howarth@nist.gov
(720)-360-9158

Get Connected

Subscribe to the **NIST PSCR newsletter** at nist.gov/ctl/pscr/get-connected

Thank You!

PUBLIC SAFETY COMMUNICATIONS RESEARCH

NIST PUBLIC SAFETY COMMUNICATIONS RESEARCH