Differential Privacy Temporal Map Challenge :

A Better Meter Stick For Differential Privacy

Metric Contest Webinar #1

October 20, 2020

Gary Howarth (NIST), Christine Task (Knexus Research), Isaac Slavitt (DrivenData)

Agenda

- Background
- Challenge overview
- How to participate
- Q&A

PSCR OPEN INNOVATION

Solving the Public Safety needs of tomorrow...today

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.

PSCR Overview

PSCR is the primary federal laboratory conducting research, development, testing, and evaluation for public safety communications technologies.

5 Key Research Areas

Cross Cutting Research Areas

Why the Challenge?

- The Public Safety Communications Research Division (PSCR) of the National Institute of Standards and Technology (NIST) is sponsoring this exciting data science competition to help advance research for public safety communications technologies for America's First Responders
- As first responders utilize more advanced communications technology, there are opportunities to use data analytics to gain insights from public safety data, inform decision-making and increase safety.
- **But… we must assure data privacy and data utility.**

PSCR OPEN INNOVATION

Solving the Public Safety needs of tomorrow...today

What's the Problem?

Public Safety As Data Generators

- As Public Safety entities make enormous gains in cyber and data infrastructure leading to the routine collection of many large datasets.
- Governments and the public are demanding greater protections on individual privacy and the privacy of individual records.
- Open data initiatives are pushing for the release of more information.

Public Safety Generates Sensitive Information

- Included in the data is personally identifiable information (PII) for police officers, victims, persons of interest, witnesses, suspects, etc.
- Studies have found that a combination of just 3 "quasi-identifiers" (date of birth, 5 digit postal code, and gender) uniquely identifies 87% of the population.

Differentially private methods guarantee that records cannot be re-linked, but do not make assurances of data quality.

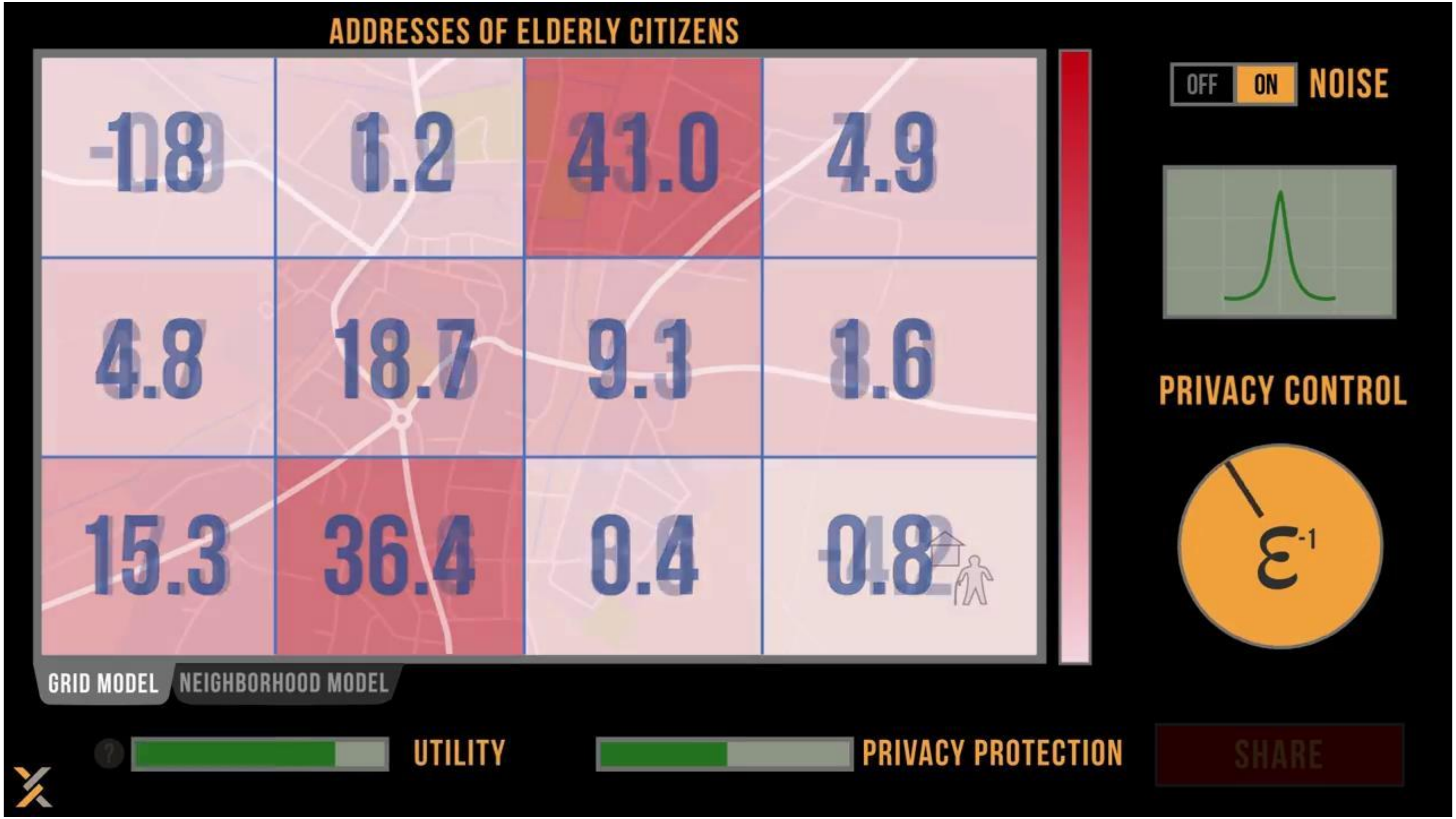
PSCR OPEN INNOVATION

Solving the Public Safety needs of tomorrow...today

Disclaimer

The following video is content created by a third-party. The contents of this video do not necessarily reflect the views or policies of the National Institute of Standards and Technology or the U.S. Government

What do we mean by Privacy?



Objective

In the Differential Privacy Temporal Map Challenge (DeID2) the objective is to **develop algorithms that preserve data utility as much as possible while guaranteeing individual privacy is protected**.

Submissions will be assessed based on

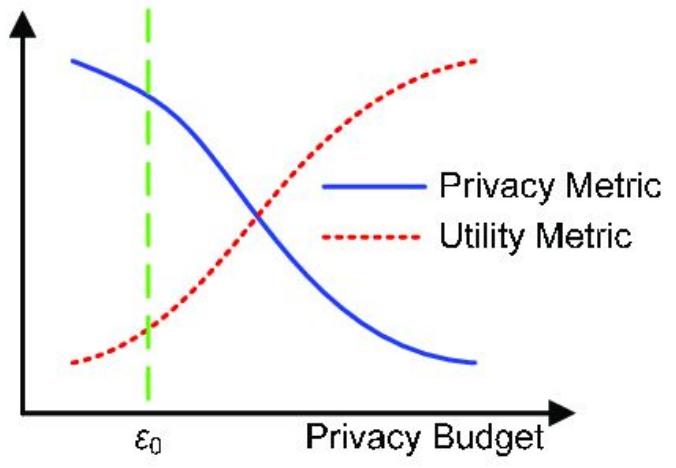
1. their ability to **prove they satisfy differential privacy**; and
2. the **accuracy of output data** as compared with ground truth.

1 Privacy write-ups

Confirmed by subject matter experts

2 Algorithm submissions

Evaluated by published performance metric



Sample illustration of the privacy-utility tradeoff. From Liu et al. "Privacy-Preserving Monotonicity of Differential Privacy Mechanisms." 2018.

PSCR OPEN INNOVATION
Solving the Public Safety needs of tomorrow...today

About Sprint 3 Scoring: The Metrics Challenge!

NIST PSCR invites solvers to develop metrics that best assess the accuracy of the data output by the algorithms that de-identify temporal map data. In particular, methods are sought that:

- Measure the quality of data with respect to temporal or geographic accuracy/utility, or both.
- Evaluate data quality in contexts beyond this challenge.
- Are clearly explained, and straightforward to correctly implement and use.

As you propose your evaluation metrics, be prepared to explain their relevance and how they would be used. These metrics may be your original content, based on existing work, or any combination thereof. If your proposed metrics are based on existing work or techniques, please provide citations. Participants will be required to submit both a broad overview of proposed approaches and specific details about the metric definition, properties and usage.

DeID2 - A Better Meter Stick for Differential Privacy

Help NIST PSCR by proposing metrics to better assess the accuracy and quality of differential privacy algorithm outputs.

Data Science | Government | Technology

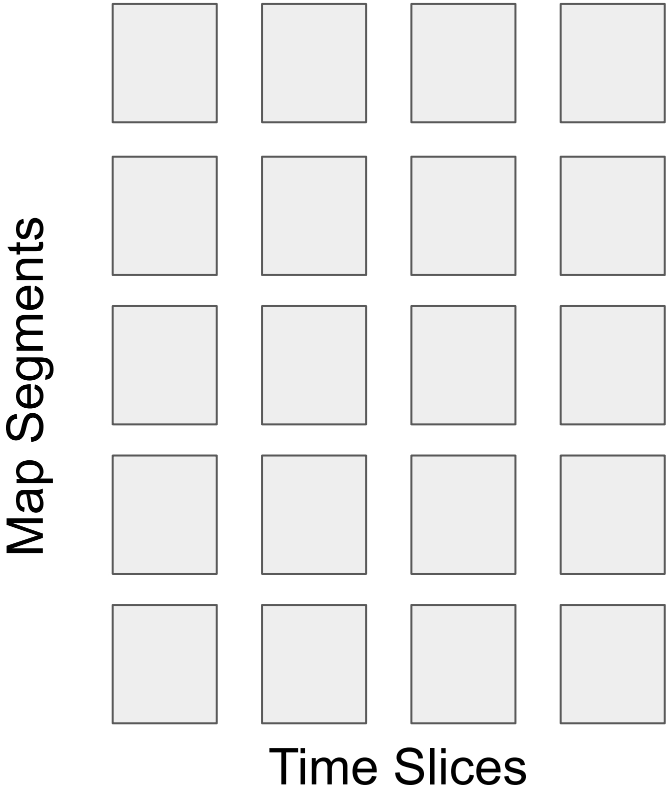
Stage: Enter

Prize: $29,000 Prize Purse

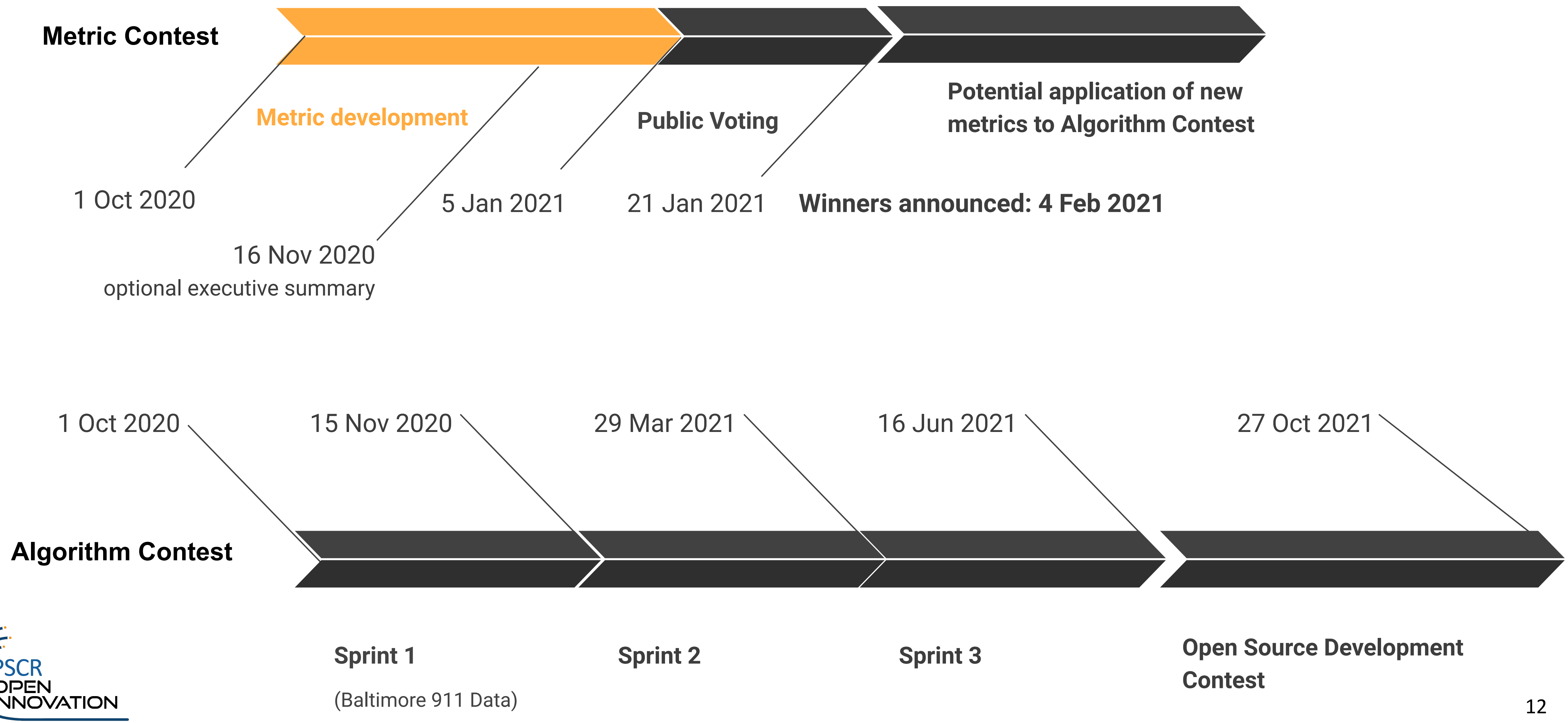
BEGIN ENTRY

Evaluation Space:

Aggregation of Event Types by Time Slice and Map Segment



Challenge Timeline



PSCR OPEN INNOVATION
Solving the Public Safety needs of tomorrow...today

Prize Awards

Metric Paper Prizes (prize purse of $29,000)
Technical Merit Winners are selected by the Judges, based on evaluation of submissions against the Judging Criteria. Up to $25,000 will be awarded. Submissions that have similar scores may be given the same prize award with up to 10 winners total. 1st Prize: Up to 2 winners of $5,000 each 2nd Prize: Up to 2 winners of $3,000 each 3rd Prize: Up to 3 winners of $2,000 each 4th Prize: Up to 3 winners of $1,000 each
People's Choice Prize Winners are selected by public voting on submitted metrics that have been pre-vetted by NIST PSCR for compliance with minimum performance criteria. Up to a total of $4,000 will be awarded to up to four winners. People's Choice: 4 @ $1,000

Challenge Structure

Timeline

Preregistration	August 24, 2020
Open to submissions	October 1, 2020
Executive Summaries due for optional preliminary review	**November 30th, 2020 10:00pm EST**
Complete submissions due	January 5, 2021 10:00pm EST
NIST PSCR Compliance check (for public voting)	January 5-6, 2021
Public voting	January 8, 2021 9:00am EST - January 21, 2021 10:00pm EST
Judging and Evaluation	January 5 - February 2, 2021
Winners Announced	February 4, 2021

Submission Template and Judging Criteria

TEMPLATE:

Executive Summary (1-2 pages)

Please provide a 1-2 page, easily readable review of the main ideas. This is likely to be especially useful for people reading multiple submissions during the public voting phase. The executive summary should be readily understood by a technical layperson and include: The high-level explanation of the proposed metric, reasoning and rationale for why it works, and an example use case.

Metric Definition

- Any technical background information needed to understand the metric.
- A written definition of the metric, including English explanation and pseudocode that has been clearly written and annotated with comments.
- Explanation of parameters and configurations.
- Walk-through examples of metric use in snapshot mode (quickly computable summary score) and/or deep dive mode (generates reports locating significant points of disparity between the real and synthetic data distributions) as applicable to the metric.

Metric Defense

- Describe the metric's tuning properties that control the focus, breadth, and rigor of evaluation
- Describe the discriminative power of the proposed metric: how well it identifies points of disparity
- Describe the coverage properties of the proposed metric: how well it abstracts/covers a breadth of uses for the data
- Address computing time constraints.
- Provide 2-3 very different data applications where the metric can be used.

CRITERIA

Clarity (30/100 points)

- Metric explanation is clear and well written, defines jargon and does not assume any specific area of technical expertise. Pseudocode is clearly defined and easily understood.
- Participants clearly address whether the proposed metric provides snapshot evaluation (quickly computable summary score) and/or deep dive evaluation (generates reports locating significant points of disparity between the real and synthetic data distributions), and explain how to apply it.
- Participants thoroughly answer the questions, and provide clear guidance on metric limitations.

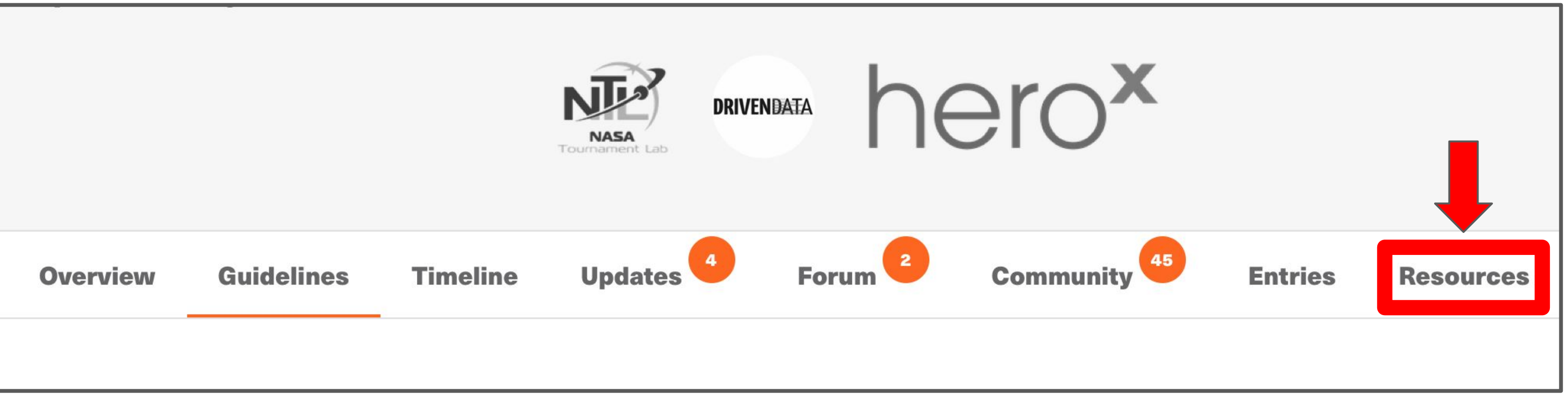
Utility (40/100 points)

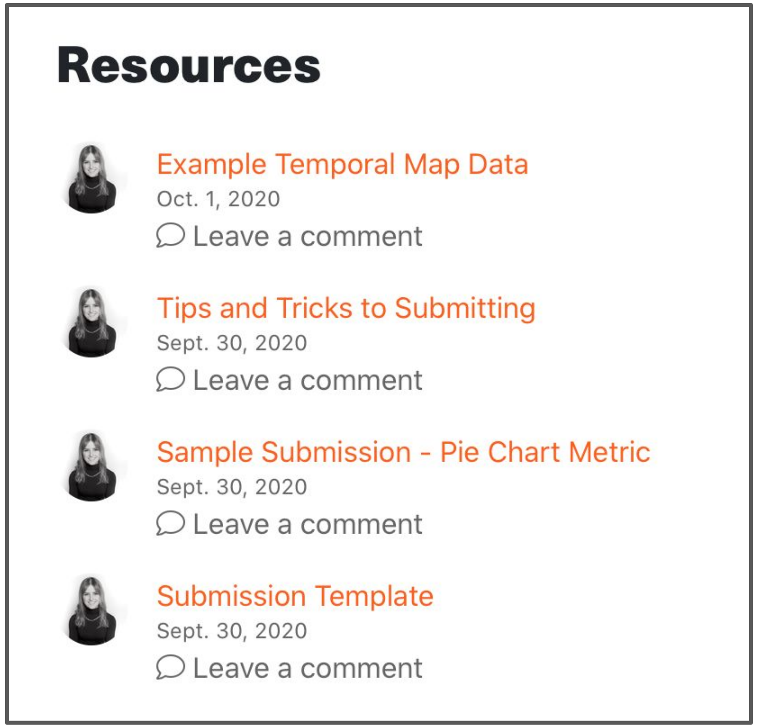
- The metric effectively distinguishes between real and synthetic data.
- The metric represents a breadth of use cases for the data.
- Motivating examples are clearly explained and fit the abstract problem definition.
- Metric is innovative, unique, and likely to lead to greater, future improvements compared with other proposed metrics.

Robustness (30/100 points)

- Metric is feasible to use for large volume use cases.
- The metric has flexible parameters that control the focus, breadth, and rigor of evaluation.
- The proposed metric is relevant in many different data applications that fit the abstract problem definition.

Resources Available





Resources Available

NASA Tournament Lab · DRIVENDATA · herox

Overview | Guidelines | Timeline | Updates 4 | Forum 2 | Community 45 | Entries | **Resources**

Resources

Example Temporal Map Data
Oct. 1, 2020
Leave a comment

Tips and Tricks to Submitting
Sept. 30, 2020
Leave a comment

Sample Submission - Pie Chart Metric
Sept. 30, 2020
Leave a comment

Submission Template
Sept. 30, 2020
Leave a comment

Example Temporal Map Data

Attachment
Example_Temporal_Map_Data.zip

Brief description
Each submission should include a demonstration of the metric on at least one data set. We provide two example data-sets here, but competitors are welcome to use or create any test data they like, as long as it includes both time-segment and map-segment information and is publicly available.

Additional information
The first data set we've provided contains event data, the 2019 Baltimore, MD 911-Call and Police Incident data, which is being used as the development phase data set for the first Sprint of the Differential Privacy Temporal Map Challenge (https://deid.drivendata.org/). The second data set is survey data, including demographic and financial features-- A small subset of IPUMS American Community Survey data for Maryland, from 2010-2018. Both examples include ground truth data, as well as privatized data at varying levels of quality, and data dictionaries.

Resources Available

NASA Tournament Lab DRIVENDATA herox

Overview Guidelines Timeline Updates 4 Forum 2 Community 45 Entries **Resources**

Resources

Example Temporal Map Data
Oct. 1, 2020
Leave a comment

Tips and Tricks to Submitting
Sept. 30, 2020
Leave a comment

Sample Submission - Pie Chart Metric
Sept. 30, 2020
Leave a comment

Submission Template
Sept. 30, 2020
Leave a comment

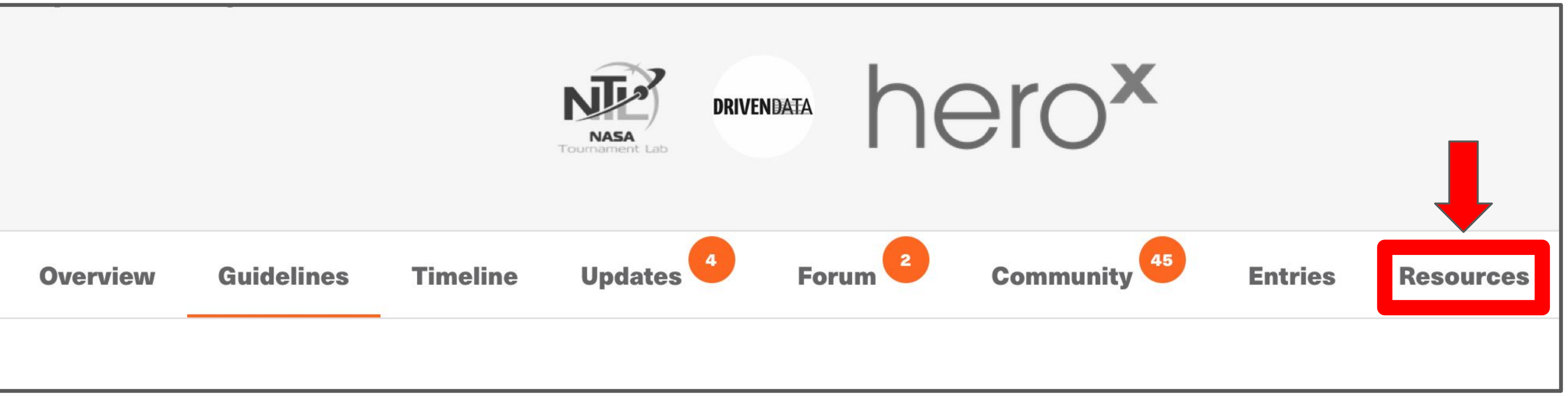
Sample Submission - Pie Chart Metric

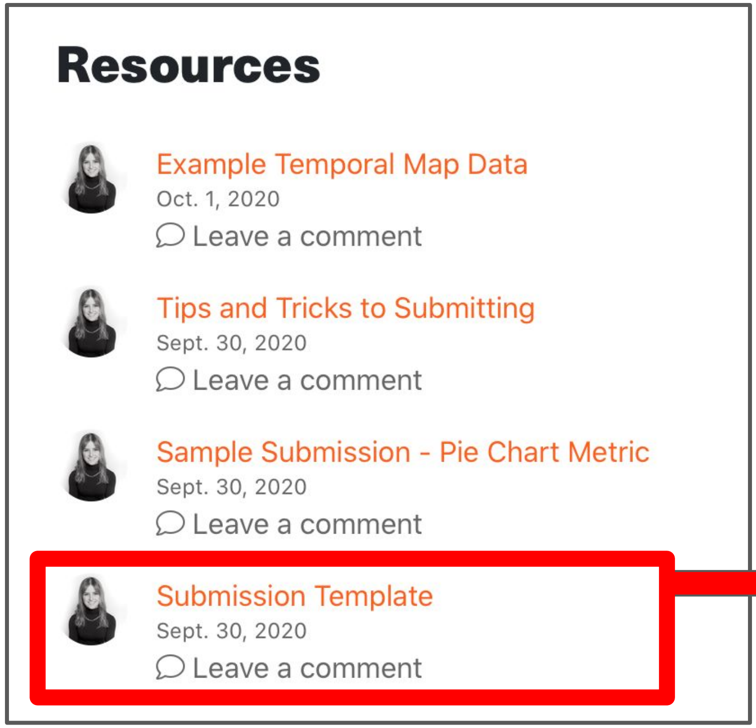
Attachment
PieChartMetric.pdf

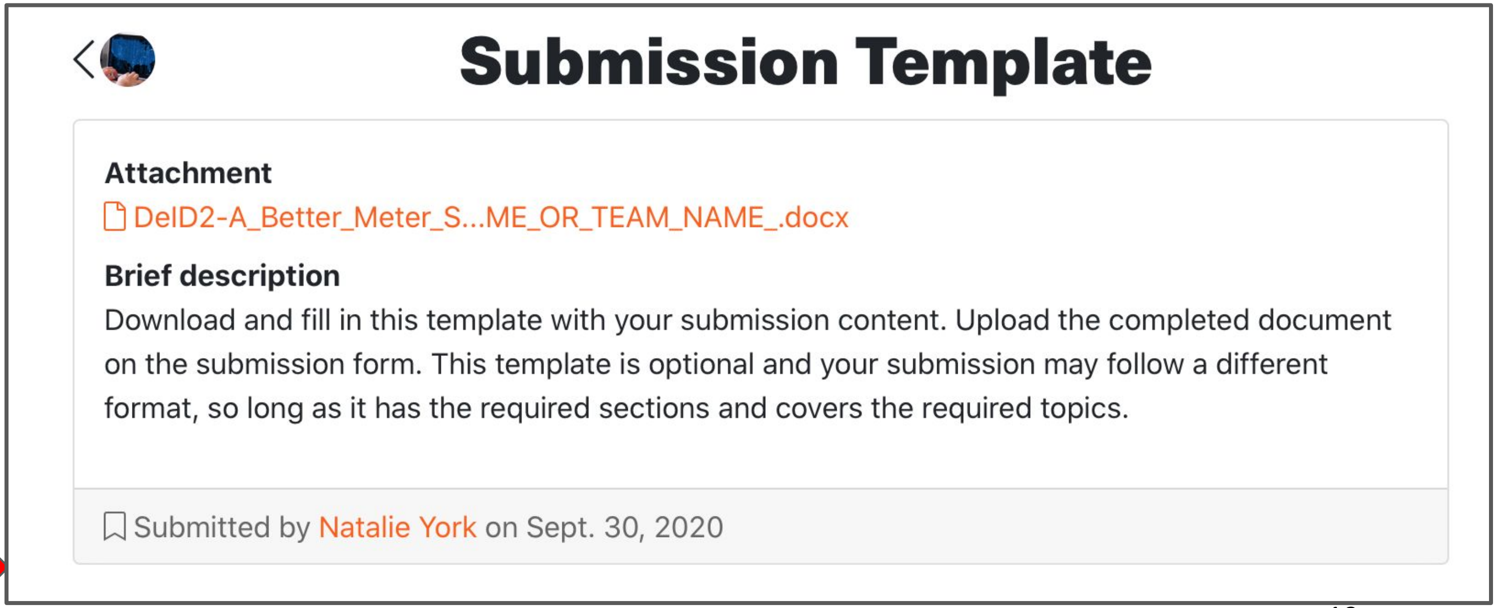
Brief description
This is an example of a high quality submission to the 'A Better Meter Stick for Differential Privacy Challenge.' Please use this example as a guideline only. You are encouraged to be creative with your submission and how you present it, so long as it fits within the template provided. There are notes, and tips and tricks, from NIST throughout the document to assist you with your submission.

Submitted by Natalie York on Sept. 30, 2020

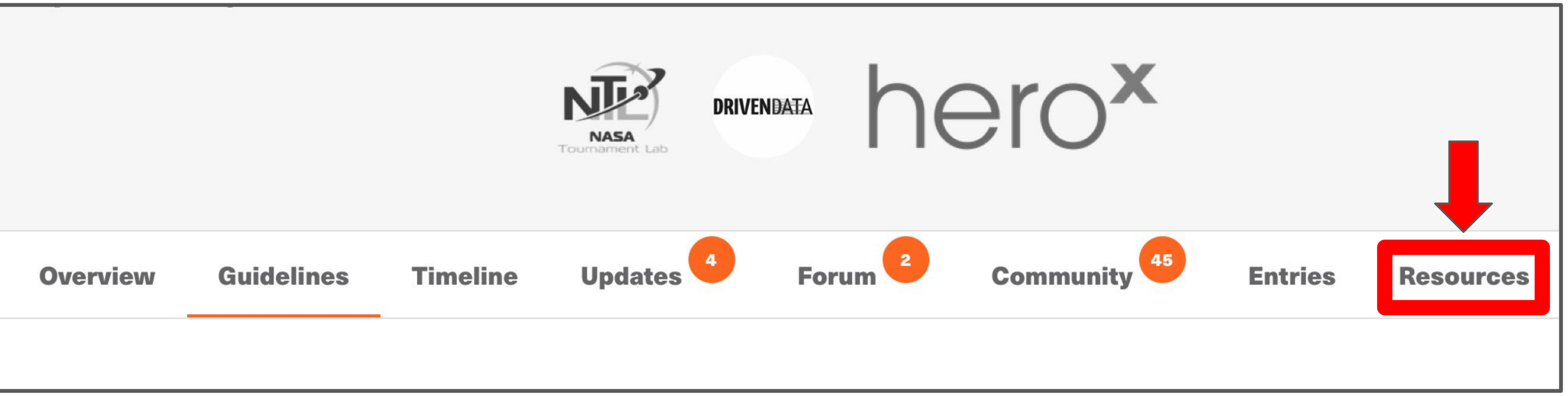
Resources Available

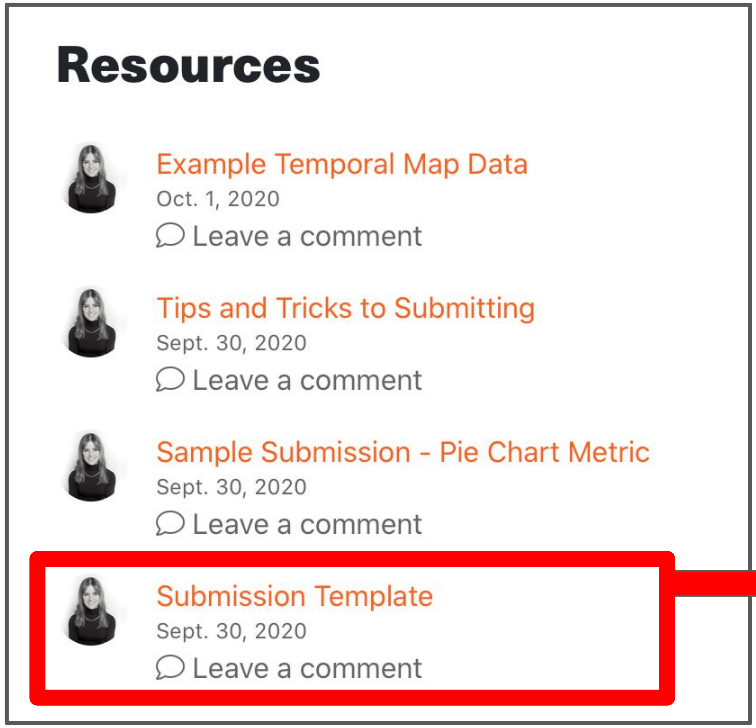


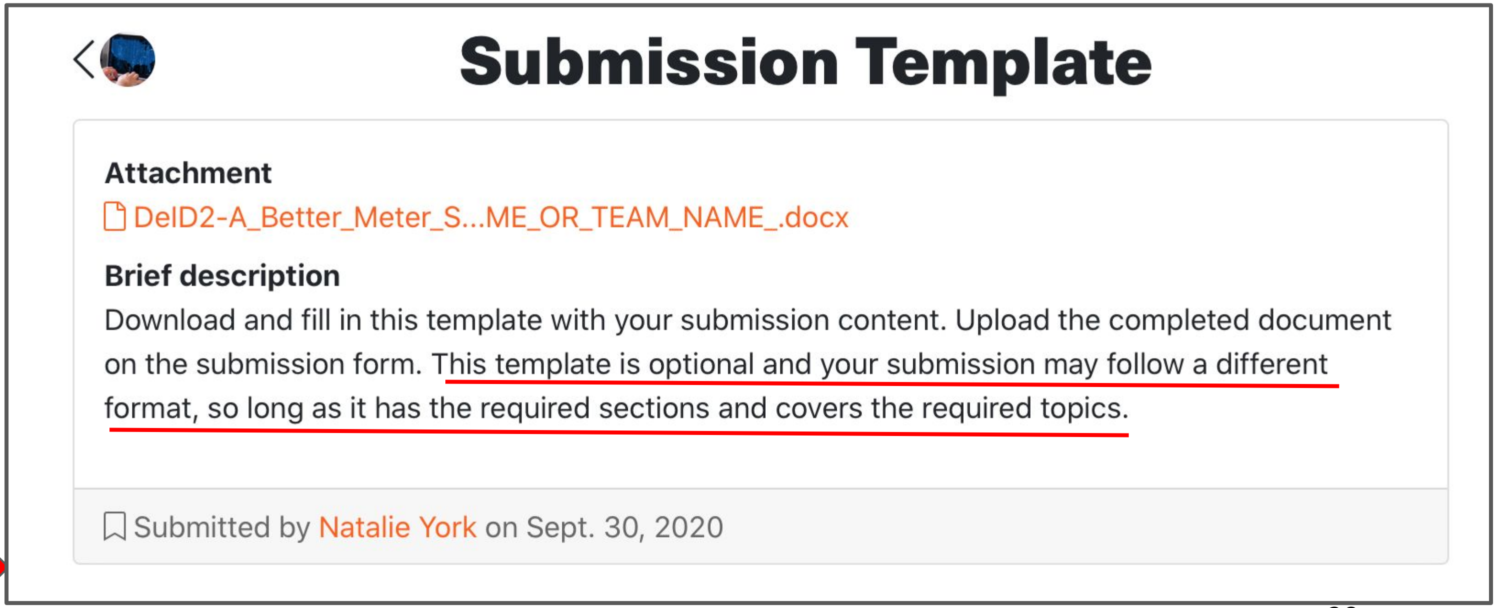




Resources Available







Also: Algorithms Competitor's Pack Contents

Differential Privacy Temporal Map Challenge: Sprint 1 (Open Arena)

HOSTED BY NIST PSCR

2 WEEKS LEFT

HOME PROBLEM DESCRIPTION ABOUT

Glory!

LEADERBOARD

DATA DOWNLOAD

SUBMISSIONS

PRE-SCREEN SUBMISSION

TEAM

External Data

Link
Competitor pack

The competition runtime repository also includes several helpful materials for developing your solution:

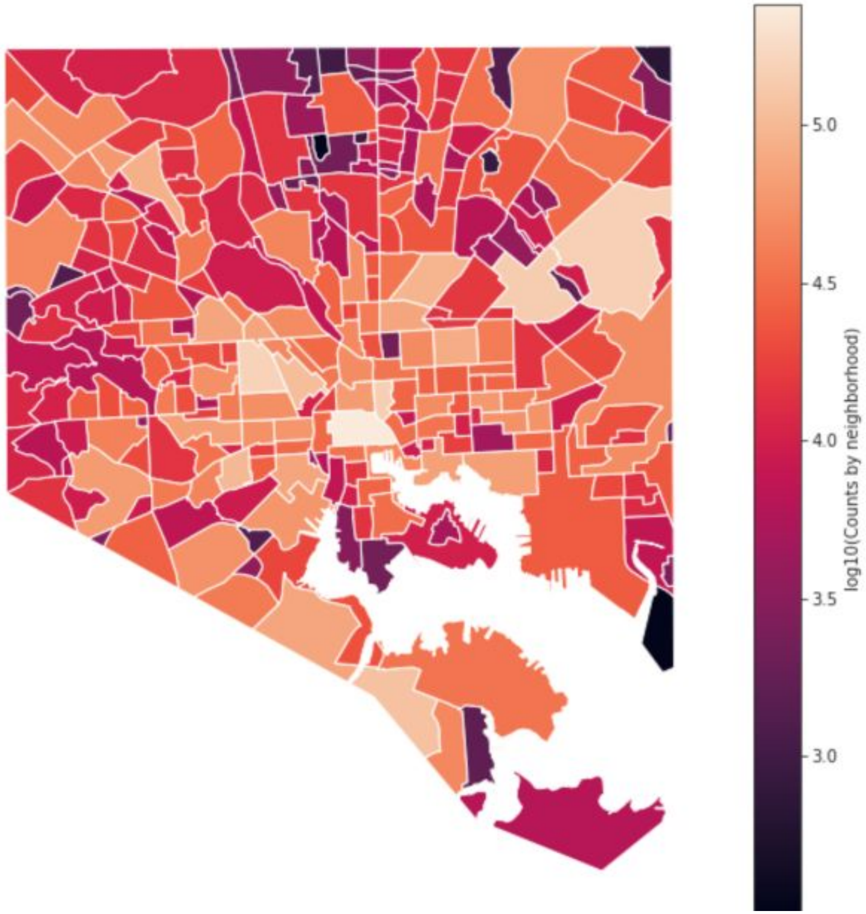
- A copy of the competition data
- A baseline solution implemented in python
- A sample privacy write-up for the baseline
- An implementation of the scoring metric for local testing
- A score visualization script that generates an HTML file displaying score outputs by neighborhood and month

PSCR OPEN INNOVATION

Solving the Public Safety needs of tomorrow...today

About the Example Data: Baltimore 911 Incidents

The first example data set is the data used in Sprint 1 of the algorithms challenge, the Baltimore 911 Incidents data. We've privatized it using the naive baseline differential privacy algorithm that we provide the Algorithms competitors as a starting point. To help you test your algorithms, we've provided the ground truth data and privatized data at three different levels of quality (although, because it's the baseline, none of them are 'great' quality).

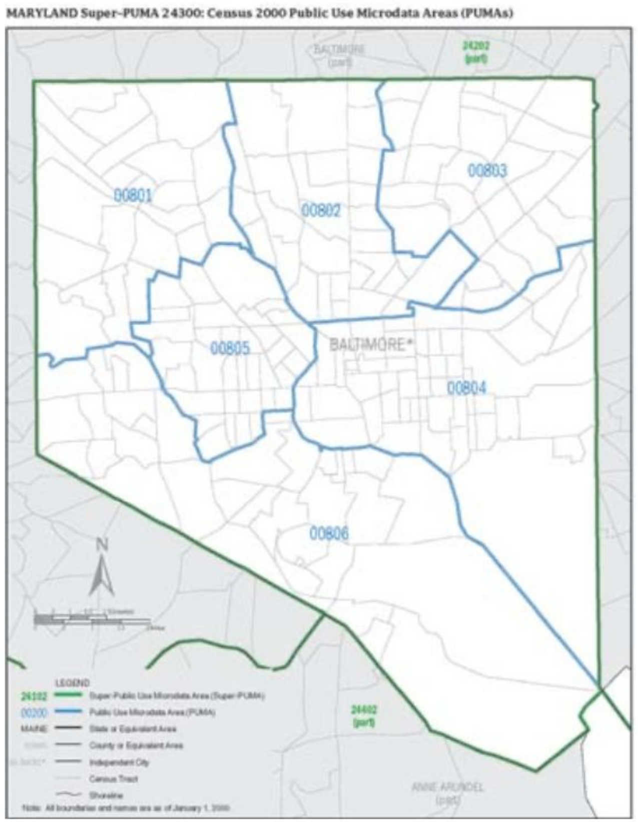


event_id	year	month	day	hour	minute	neighborhood	incident_type	sim_resident
140203235110672	2019	1	1	0	0	29	167	4081
140203737381840	2019	1	1	0	0	166	168	6115
140202952922576	2019	1	1	0	0	147	163	17498
140203118608848	2019	1	1	0	0	251	166	30987
140203196663184	2019	1	1	0	0	166	163	35984
...	...	...	...	...	...	...	...	...

- `event_id` (int) — Unique ID for each row of the data.
- `year`, `month`, `day`, `hour`, `minute` (int) — Time when the call took place.
- `neighborhood` (categorical [int]) — Code for the neighborhood in which the incident took place. See the codebook for the human-readable name corresponding to this code.
- `incident_type` (categorical [int]) — Code for which type of incident took place. See the codebook for the human-readable name corresponding to this code.
- `sim_resident` (int) — Unique, synthetic ID for the notional person to which this event was attributed. The largest number of incidents attributed to a single simulated resident is provided in the parameters.json file as `max_records_per_individual`.

About the Example Data: Maryland ACS Data

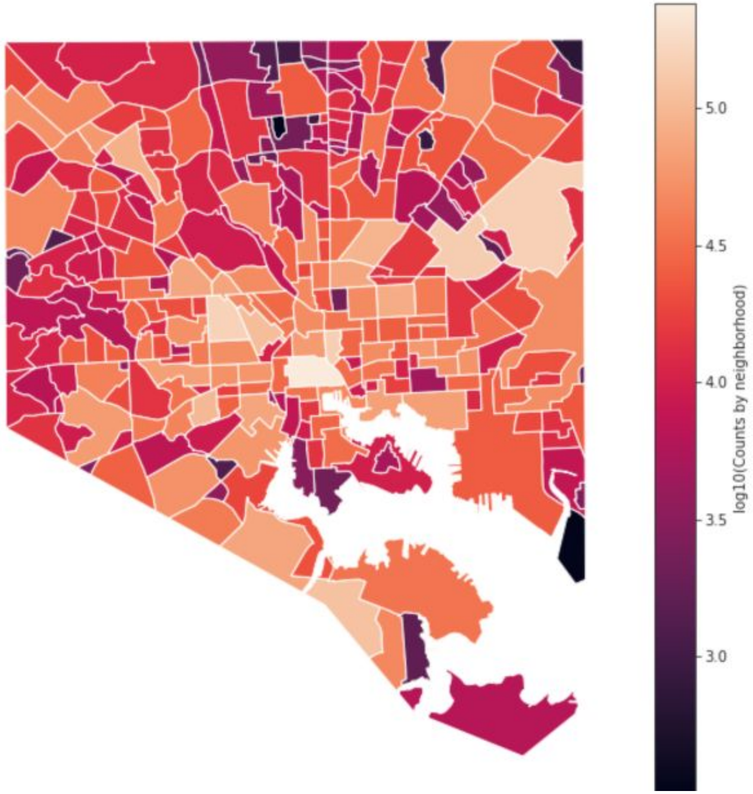
To provide a data set with demographic and financial variables, along with map segments and time segments, our second example data is excerpted from 8 years of Maryland American Community Survey data (IPUMS archive). The map segments are Public Use Microdata Area (shown for Baltimore below), and the time segments are years. We’ve privatized it using the non-differentially private Knexthetic Synthesizer developed by Knexus Research. To help you test your algorithms, we’ve provided the ground truth data and privatized data at three different levels of quality.



Time Seg.	**Map Seg.**	**Demographics**	**Financials**	**Other**
YEAR	PUMA	AGE	INCWAGE	ARRIVAL
		SEX	INCEARN	DEPARTURE
		RACE	INCTOT	
		HISPAN	POVERTY	
		EDUC		

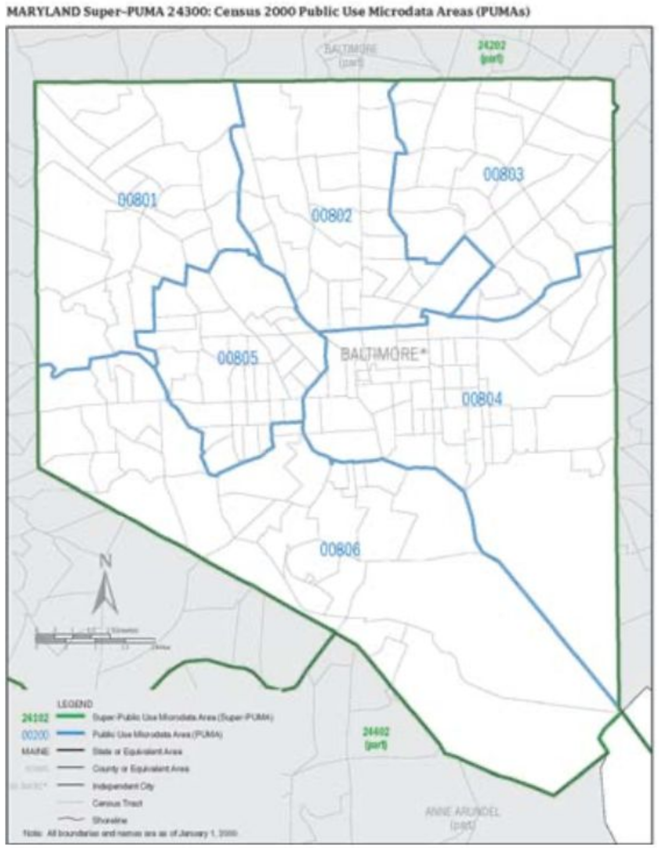
About the Event Data:

Important note: Competitors aren’t required to use *either* of the provided example data-sets.

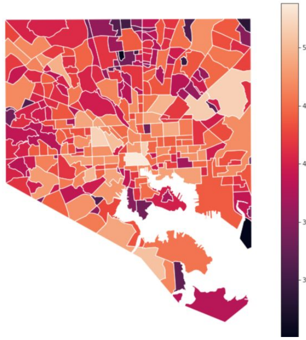


These are simply provided as a convenience; you are welcome and encouraged to find other temporal map data sets (any data set with both timestamp and map segment information) to demonstrate your proposed evaluation metric.

It doesn’t even have to involve Maryland!



About the Problem: Pile of Records in Space Time



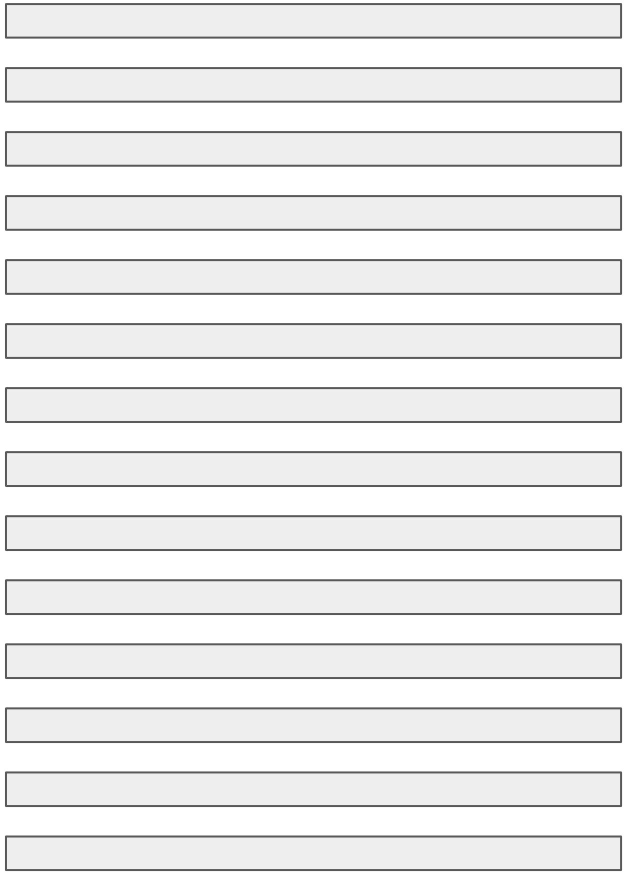
The **Temporal Map Data** problem is fundamentally a problem of counts of record types spread across space time.

It’s challenging for privacy because when one individual may contribute data at multiple points in time, they leave a larger and more unique footprint in the data, and that can be hard to cover up.

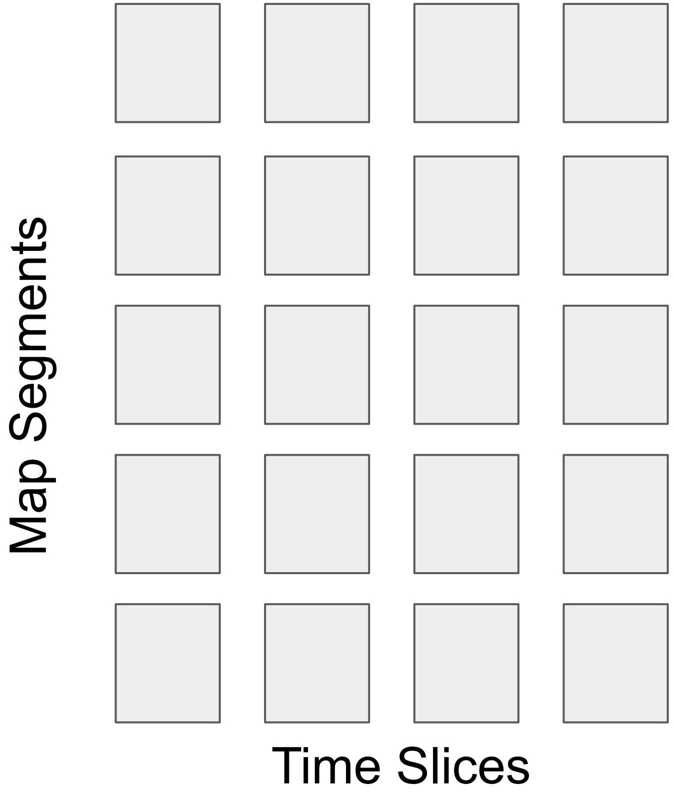
It’s challenging for Metrics, because we need to make sure data correlations are maintained across both geography and over time. In complex, higher dimensional spaces it can be easy for metrics to have ***unknown*** blindspots.

Known blindspots are fine, all merics will have blindspots because they’re simplified measures of something much more complex. But having a robust understanding of your metric’s behavior is important.

<u>Raw Records:</u>
Timestamp, Map Segment, Record Data



<u>Evaluation Space:</u>
Aggregation of Event Types by Time Slice and Map Segment



An Example! Sprint 1 Scoring: The Pie Chart Metric

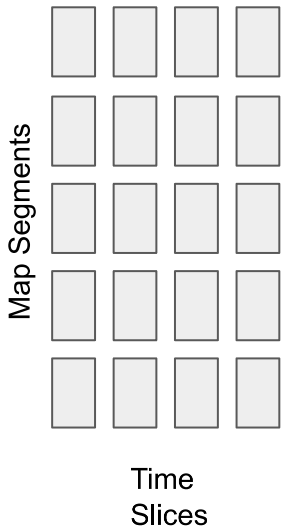
Baseline Piechart Score:

The objective of the pie chart is to measure how faithfully the privatization algorithm preserves the most significant patterns in the data, within each map/time segment. It does this by only considering the record types that make up at least k% of the total records (the ‘sufficiently thick pie slices’).

gt	0	2	28	20
	0%	4%	56%	40%

dp	26	0	2	22
	52%	0%	4%	44%

Evaluation Space:



An Example! Sprint 1 Scoring: The Pie Chart Metric

Baseline Piechart Score:

The objective of the pie chart is to measure how faithfully the privatization algorithm preserves the most significant patterns in the data, within each map/time segment. It does this by only considering the record types that make up at least k% of the total records (the 'sufficiently thick pie slices').

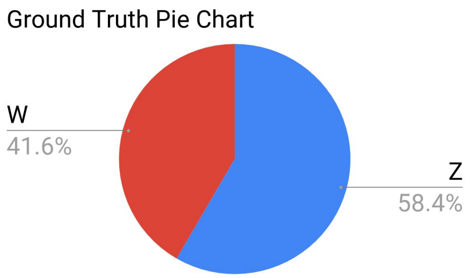
gt	0	2	28	20
	0%	4%	56%	40%

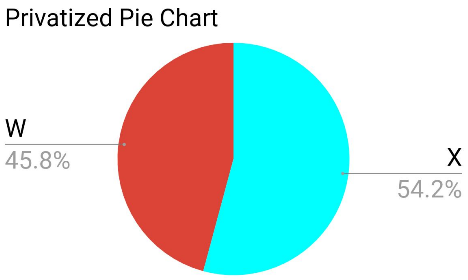
dp	26	0	2	22
	52%	0%	4%	44%

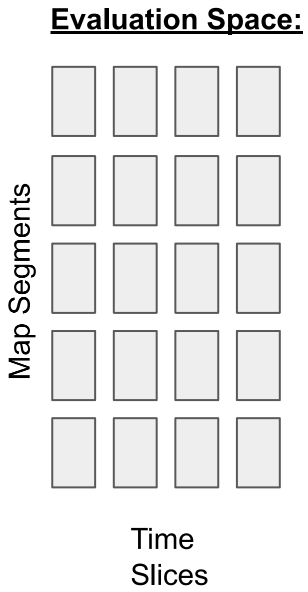
Zero out non-significant counts in each vector, re-normalize, and compute the Jensen-Shannon Distance to get the baseline piechart score (0.7505).

gt	0	0	28	20
	0%	0%	58%	42%

dp	26	0	0	22
	54%	0%	0%	46%







An Example! Sprint 1 Scoring: The Pie Chart Metric

Baseline Piechart Score:

The objective of the pie chart is to measure how faithfully the privatization algorithm preserves the most significant patterns in the data, within each map/time segment. It does this by only considering the record types that make up at least k% of the total records (the 'sufficiently thick pie slices').

gt	0	2	28	20
	0%	4%	56%	40%

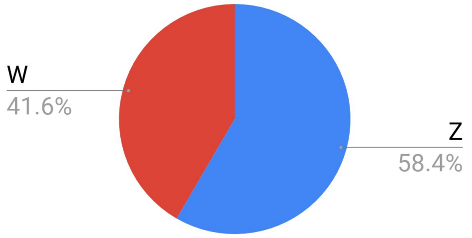
dp	26	0	2	22
	52%	0%	4%	44%

Zero out non-significant counts in each vector, re-normalize, and compute the Jensen-Shannon Distance to get the baseline piechart score (0.7505).

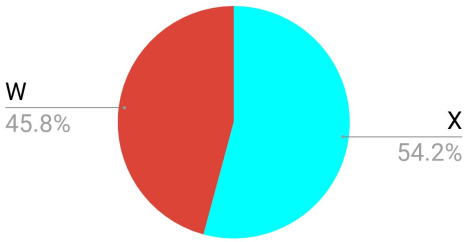
gt	0	0	28	20
	0%	0%	58%	42%

dp	26	0	0	22
	54%	0%	0%	46%

Ground Truth Pie Chart



Privatized Pie Chart



Technical Background

Jensen Shannon Distance:

The Jensen-Shannon distance (metric) is the square root of the Jensen-Shannon divergence. Given two probability vectors ***p*** and ***q***, the Jensen-Shannon distance is defined as,

$$\sqrt{\frac{D(p \parallel m)+D(q \parallel m)}{2}}$$

where ***m*** is the pointwise mean of ***p*** and ***q*** and ***D*** is the Kullback-Leibler divergence.

The Jensen Shannon Distance is based on the more commonly used KL divergence; however, unlike KL divergence it measures the symmetric distance between two distributions.   A symmetric metric is important in our use case because due to positive privatization noise, the differentially private pie chart may include labels that do not appear in the ground truth data. KL divergence measures difference from a specified baseline distribution (ie, distance from the ground truth in our case) and is undefined at points where the baseline distribution is 0.   This effectively means that the KL divergence is infinite whenever the privatized pie chart includes an extra label.   Our metric penalizes spurious labels with the Misleading Presence Penalty… but an infinite penalty might be too harsh.

An Example! Sprint 1 Scoring: The Pie Chart Metric

Baseline Piechart Score:

The objective of the pie chart is to measure how faithfully the privatization algorithm preserves the most significant patterns in the data, within each map/time segment. It does this by only considering the record types that make up at least k% of the total records (the ‘sufficiently thick pie slices’).

gt	0	2	28	20
	0%	4%	56%	40%

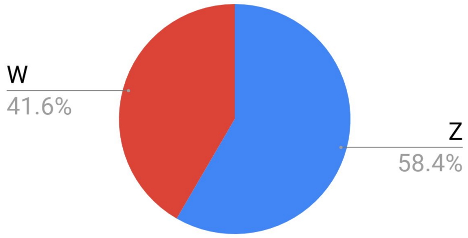
dp	26	0	2	22
	52%	0%	4%	44%

Zero out non-significant counts in each vector, re-normalize, and compute the Jensen-Shannon Distance to get the baseline piechart score (0.7505).

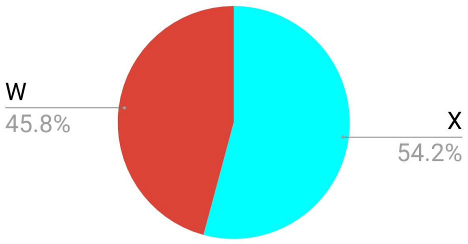
gt	0	0	28	20
	0%	0%	58%	42%

dp	26	0	0	22
	54%	0%	0%	46%

Ground Truth Pie Chart



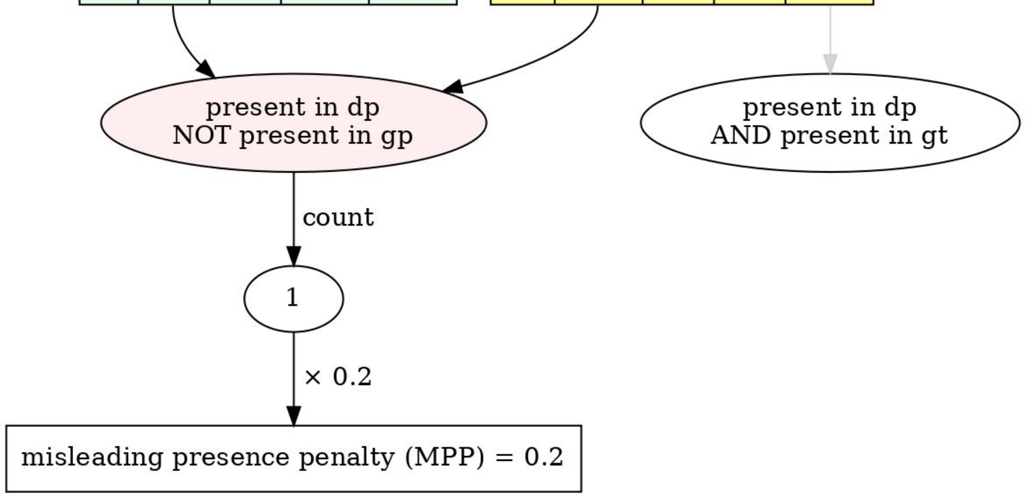
Privatized Pie Chart



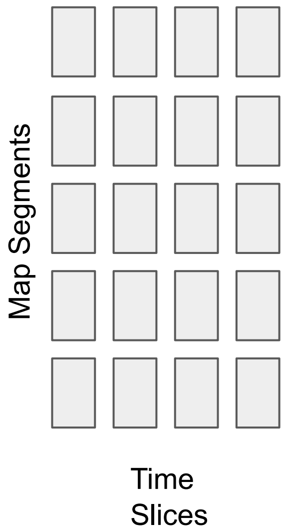
Penalties:

gt	0	0	28	20
	0%	0%	58%	42%

dp	26	0	0	22
	54%	0%	0%	46%



Evaluation Space:



An Example! Sprint 1 Scoring: The Pie Chart Metric

Baseline Piechart Score:

The objective of the pie chart is to measure how faithfully the privatization algorithm preserves the most significant patterns in the data, within each map/time segment. It does this by only considering the record types that make up at least k% of the total records (the ‘sufficiently thick pie slices’).

gt	0	2	28	20
	0%	4%	56%	40%

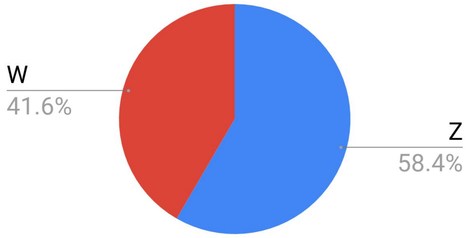
dp	26	0	2	22
	52%	0%	4%	44%

Zero out non-significant counts in each vector, re-normalize, and compute the Jensen-Shannon Distance to get the baseline piechart score (0.7505).

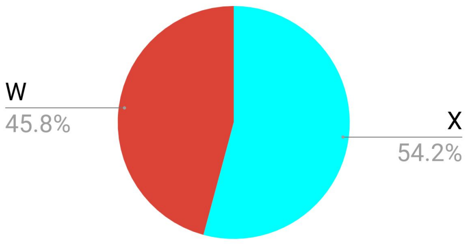
gt	0	0	28	20
	0%	0%	58%	42%

dp	26	0	0	22
	54%	0%	0%	46%

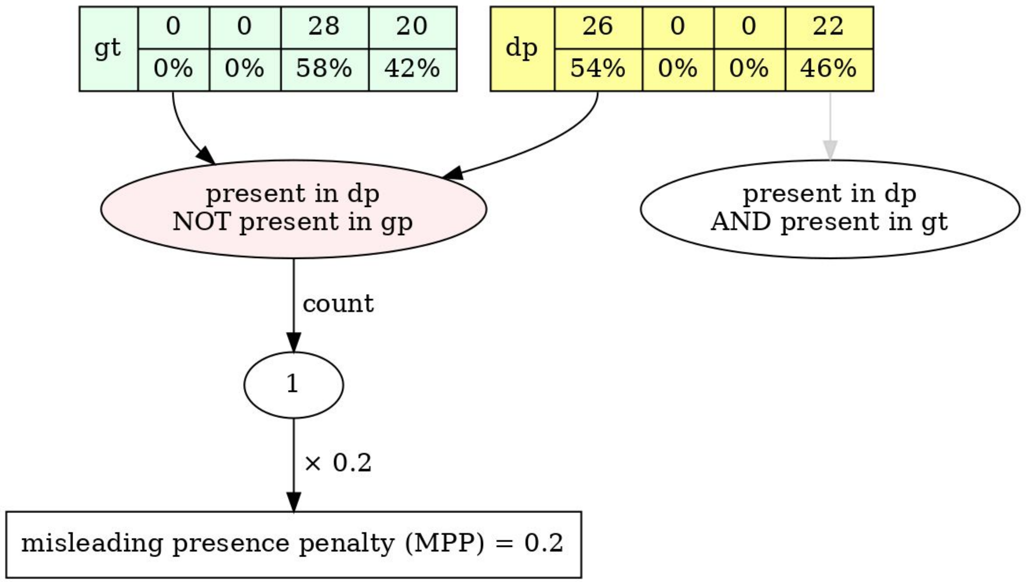
Ground Truth Pie Chart

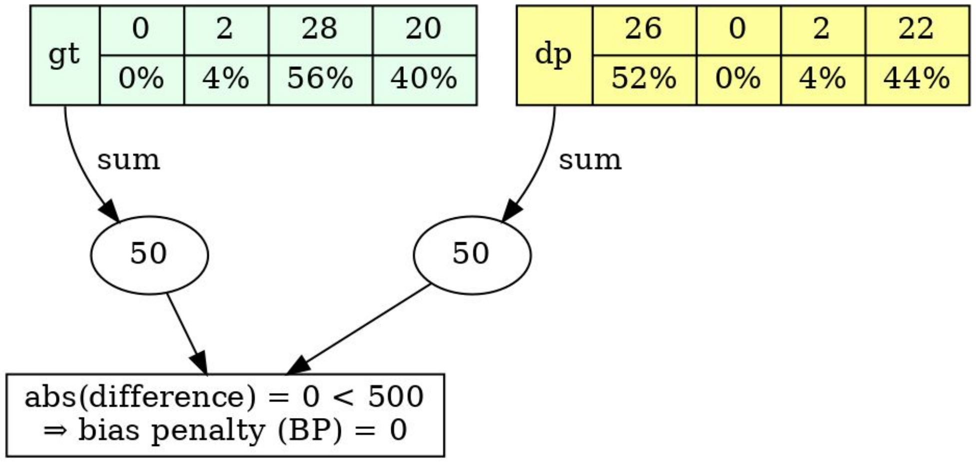


Privatized Pie Chart

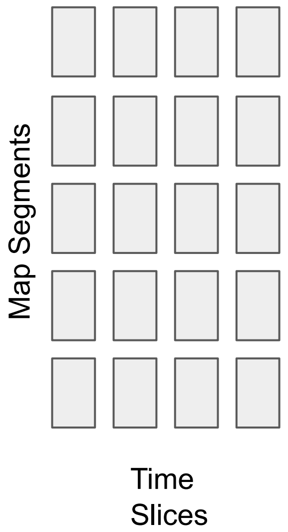


Penalties:





Evaluation Space:



An Example! Sprint 1 Scoring: The Pie Chart Metric

Baseline Piechart Score:

The objective of the pie chart is to measure how faithfully the privatization algorithm preserves the most significant patterns in the data, within each map/time segment. It does this by only considering the record types that make up at least k% of the total records (the ‘sufficiently thick pie slices’).

gt	0	2	28	20
	0%	4%	56%	40%

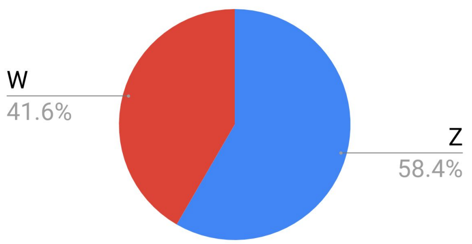
dp	26	0	2	22
	52%	0%	4%	44%

Zero out non-significant counts in each vector, re-normalize, and compute the Jensen-Shannon Distance to get the baseline piechart score (0.7505).

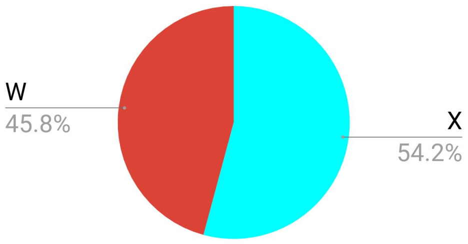
gt	0	0	28	20
	0%	0%	58%	42%

dp	26	0	0	22
	54%	0%	0%	46%

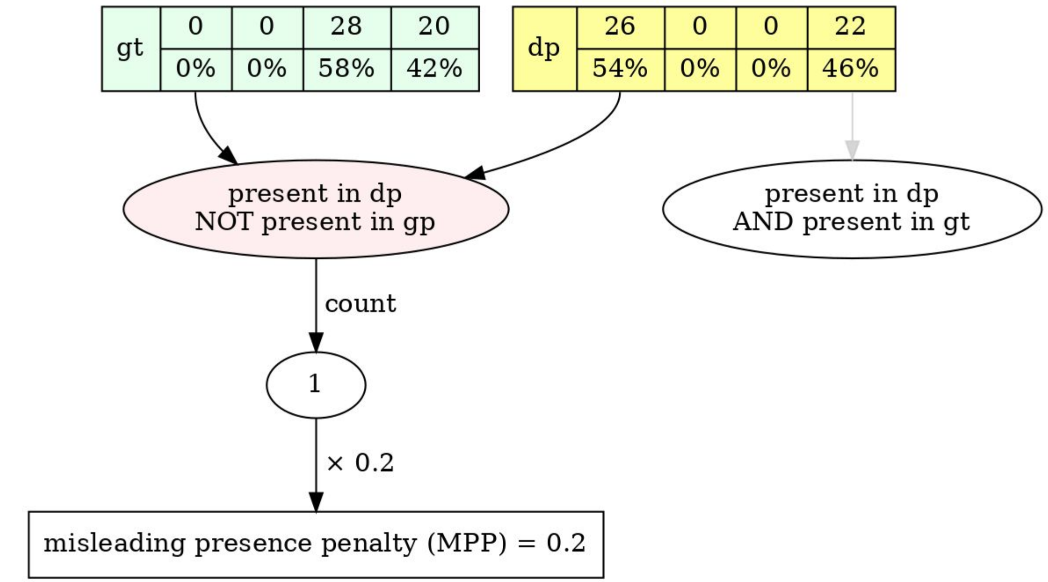
Ground Truth Pie Chart

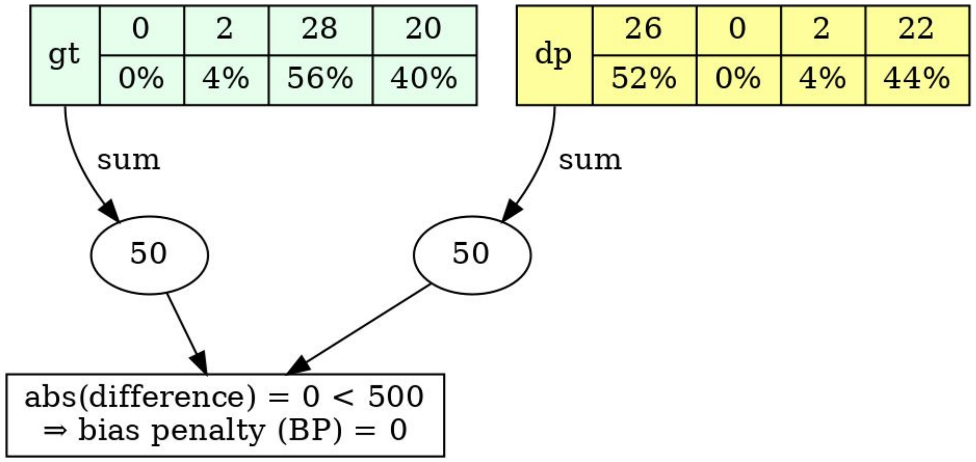


Privatized Pie Chart



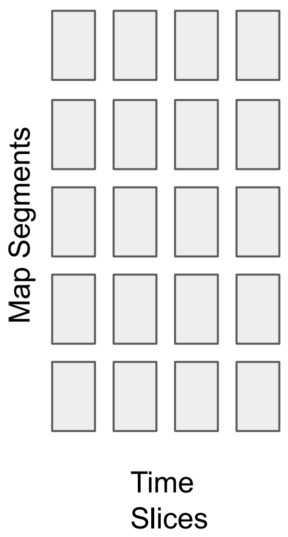
Penalties:



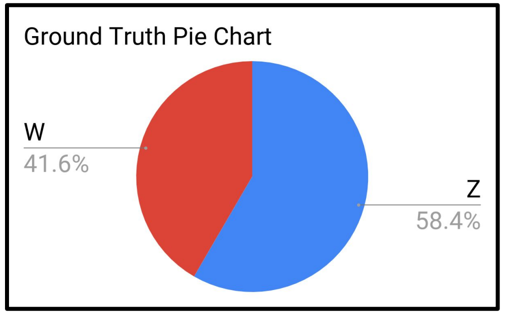


To get a total score for the whole temporal map, ***the pie chart scores of all individual map/time segments are averaged together.***

Evaluation Space:

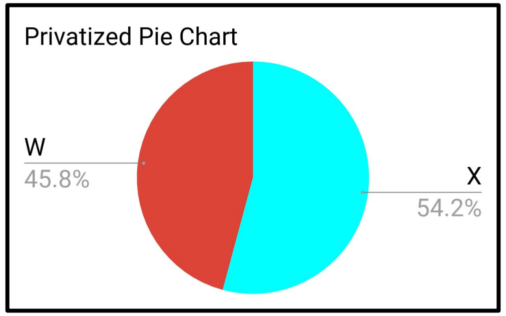


An Example! Sprint 1 Scoring: The Pie Chart Metric



gt	0	0	28	20
	0%	0%	58%	42%

dp	26	0	0	22
	54%	0%	0%	46%



Description of Discriminative Power:

In this section we briefly outline some of the capabilities and limitations of the pie chart metric with respect to its discriminative power-- how well it can distinguish between the ground truth and privatized data.

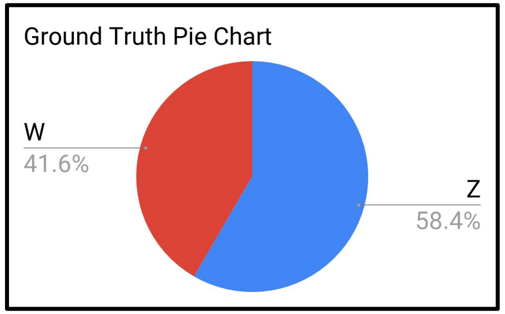
Capabilities

- The pie chart metric identifies disparities between the distribution of high frequency record types
- The pie chart metric specifically penalizes positive (or negative) bias in total record counts
- The pie chart metric specifically penalizes when rare large noise values are sampled from the laplace distribution, causing spurious record types to appear to be high frequency in the privatized data.
- We've demonstrated that the pie chart metric responds to small changes in the value of epsilon (or sampling error), and allows us to meaningfully understand the impact of those changes on the data.

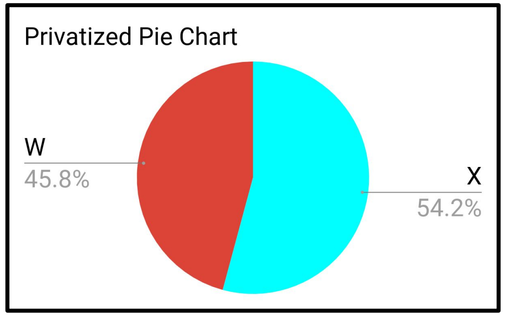
Limitations

- The pie chart metric does not measure the impact of privatization noise on patterns in less frequent record types; those record types are discarded during the frequency thresholding.
- The pie chart metric does not capture the relative ranking of high frequency record types. In general, Jensen-Shannon distance summarizes the total difference in distributions; two similarly-sized pie slices may change order without significantly impacting the JS score.
- The pie chart metric does not measure trends across time. Scores are summed across all time/map segments without attention to any broader patterns.
- The pie chart metric focuses on record *types*, effectively categorical information; if the data includes numerical features these must be partitioned into

An Example! Sprint 1 Scoring: The Pie Chart Metric



gt	0	0	28	20	dp	26	0	0	22
	0%	0%	58%	42%		54%	0%	0%	46%



Description of Coverage:

In this section we briefly outline some of the capabilities and limitations of the pie chart metric with respect to its coverage-- how well it represents a breadth of possible use cases.

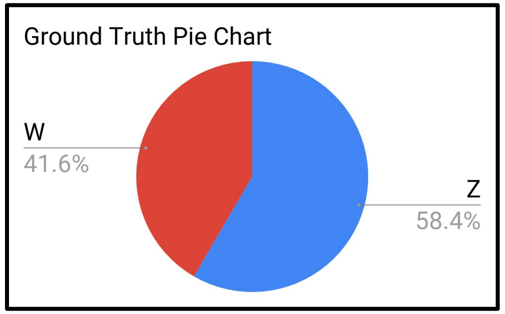
Capabilities

- Pie charts are a commonly used tool for communicating survey results in public policy and other decision making contexts. They provide the reader with a quick sense of the 'most significant' simple features of the data. The pie chart metric evaluates whether the privatized data will be suitable for these common, basic applications.
- Because the pie chart metric evaluates the relative proportion of frequent record types, a high pie chart score is an indicator that the privatized that will maintain utility for policy decisions on questions like funding, fairness, and staff/resource placement (although we have not explicitly studied the relationship between the pie chart score and funding disbursement).

Limitations

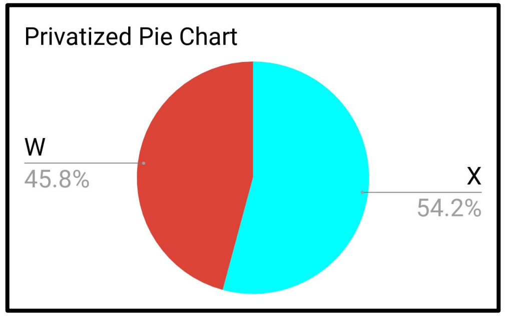
- The pie chart metric doesn't cover more complex analytics or machine learning tasks (ex: regression, classification) which may have different sensitivities to added noise.
- As mentioned above, the pie chart metric doesn't cover applications that study larger patterns/trends across time or geography
- The pie chart metric focuses on categorical information (record types), and does not evaluate for many criteria specific to numerical data, such as whether the privatized data maintains long tailed distributions or maintains accuracy across repeated numerical operations over privatized data ("differences of differences").
- The pie chart metric doesn't cover applications that reference infrequent record types

An Example! Sprint 1 Scoring: The Pie Chart Metric



gt	0	0	28	20
	0%	0%	58%	42%

dp	26	0	0	22
	54%	0%	0%	46%



Scalability/Feasibility:

The pie chart metric is a constant time operation in terms of the number of record types, map segments and time segments. The pie chart metric on the 196 record types and 10,008 map/time segments in the Sprint 1 Baltimore Police Data can be computed within seconds on a typical laptop.

Generalizability, Alternate Use Cases:

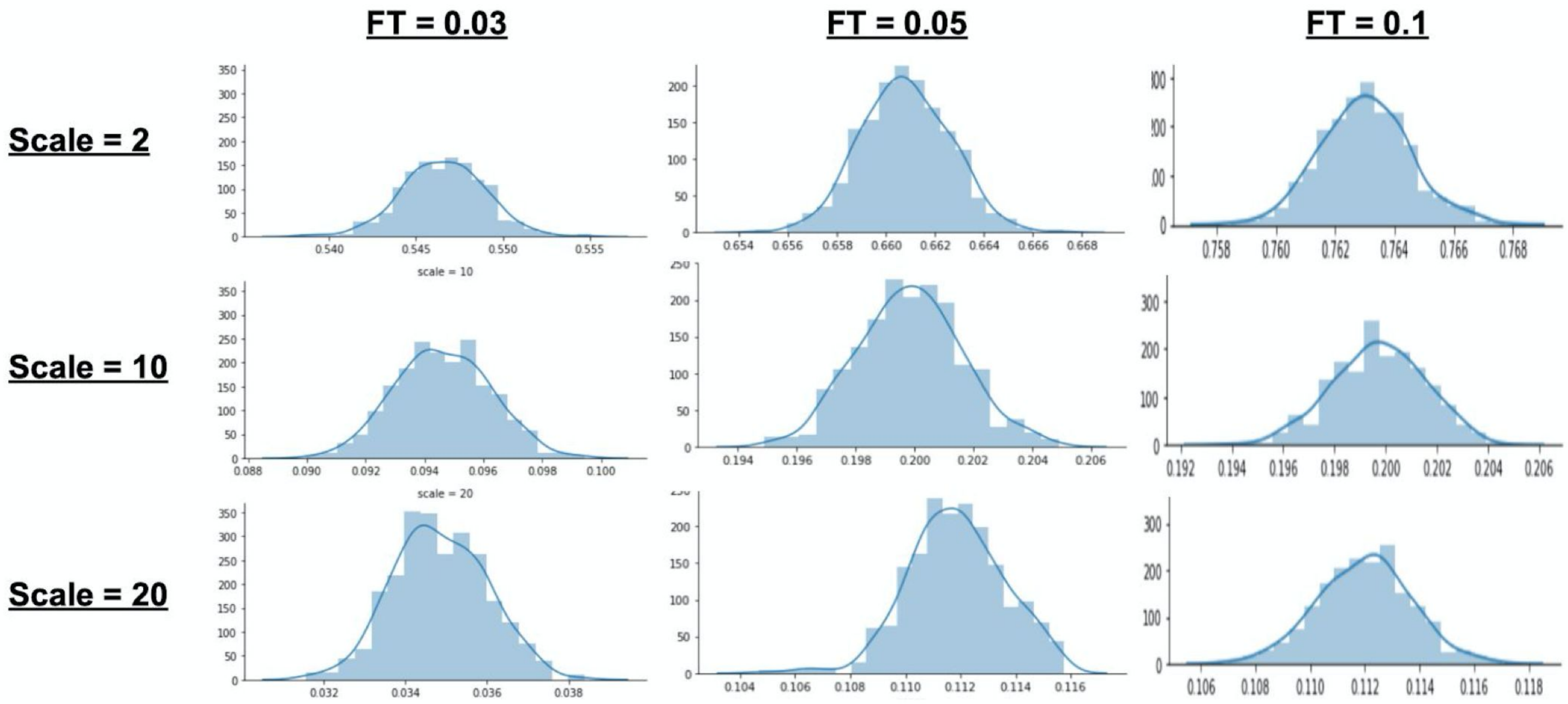
In this document, we’ve demonstrated how the pie chart metric can apply to event records. Here are a few additional examples how the pie chart metric could be applied:

- The pie chart metric can apply to demographic data, financial data, etc, by using marginals (and binning numerical features) when defining the “record type”, as outlined in the Parameters-Configuration, in the Metric Definition.
- Although by default the pie chart metric does not capture trends across time or geography, it could be applied to capture more temporal information. By setting the record types to be “Events that happened between timestamp x_1 and timestamp x_2”, a pie chart can be drawn up to measure clusters of records in time the same way we’ve used it above to measure distributions in feature space. Similarly, the pie chart metric could be used to capture frequent patterns geographically, by setting the record types to be “Events that happened in region X”. Combinations of time and geography are also possible.

An Example! Pie Chart Parameter Exploration

Effect of Frequency Threshold:

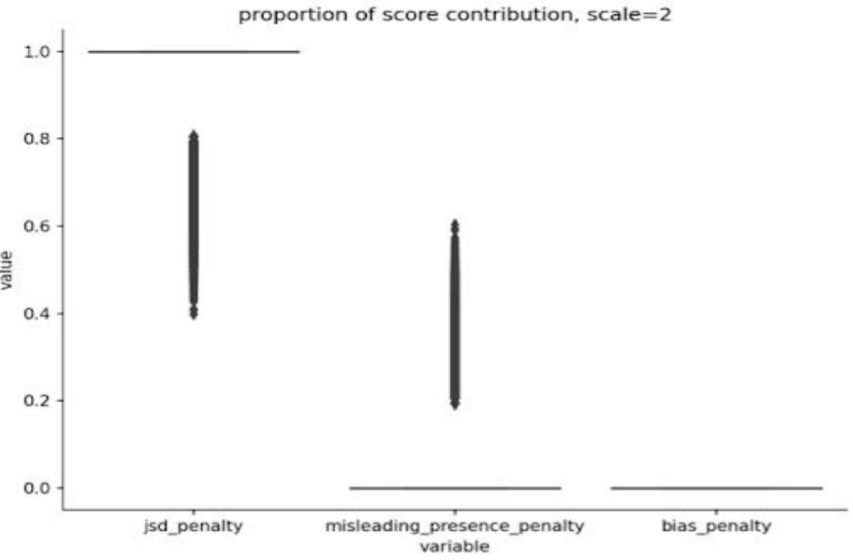
We now briefly explore the effect of increasing and decreasing the frequency threshold from the default value 5% (FT = 0.05).

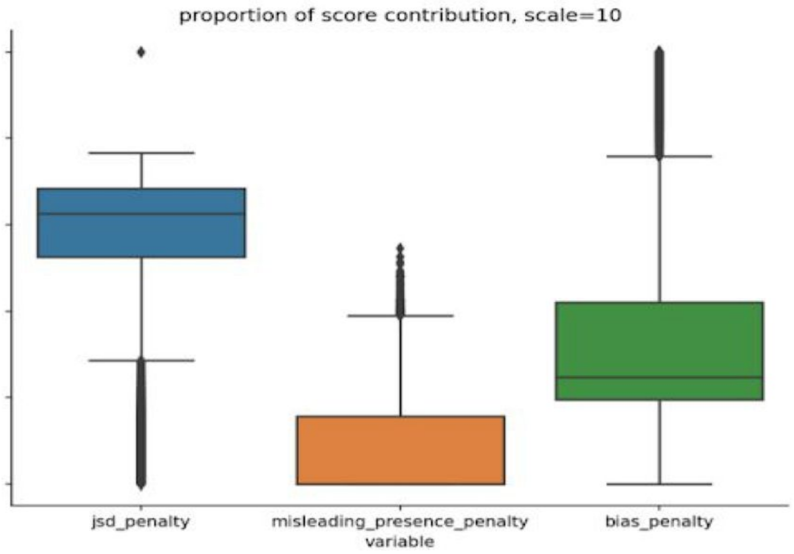


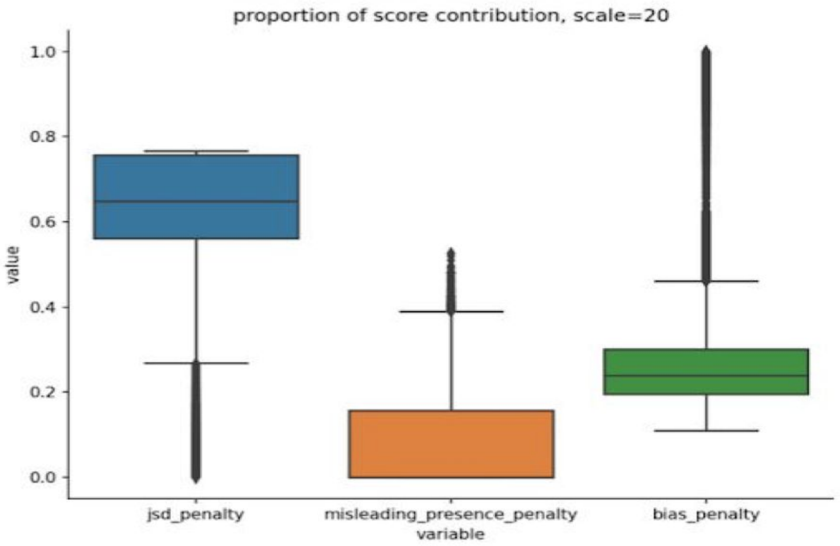
An Example! Pie Chart Parameter Exploration

Score Composition:

The Pie Chart metric has three components (as described in the Metric Definition section above): JSD, MPP, BP. Here we look at the impact each component has on the total score, dependent on the quality of the data.



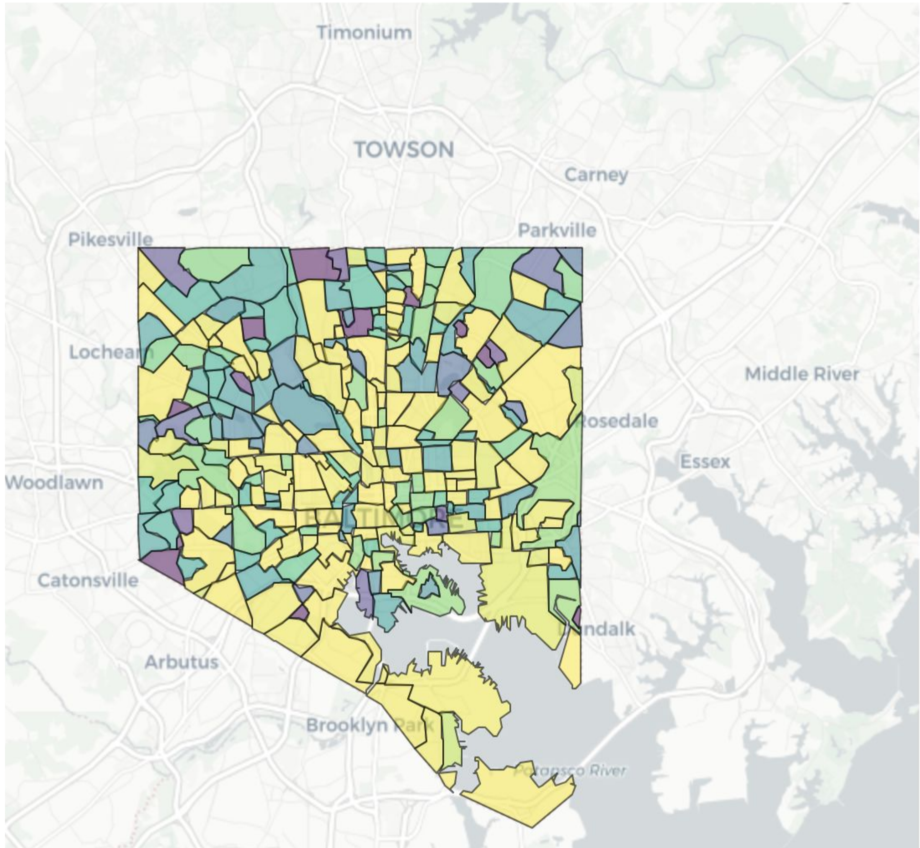




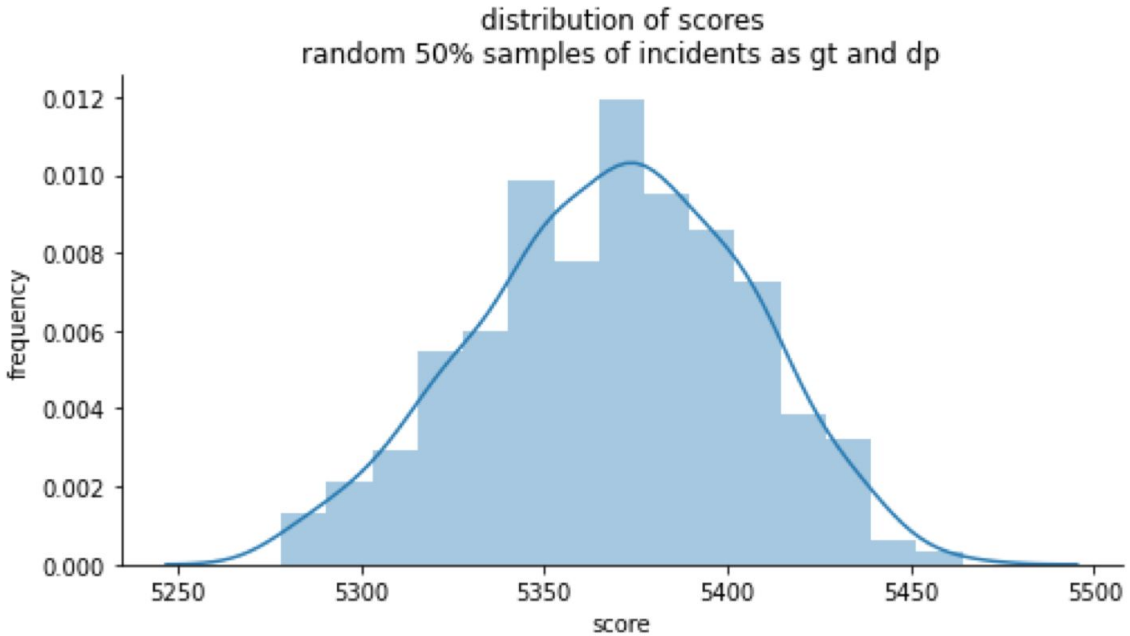
An Example! Pie Chart Sampling Error Benchmark

We can get a benchmark score for good performance by comparing the noise in our privatized results to sampling error. **In this case, we take two uniform random subsamples** of the full incident record data, and then arbitrarily choose one to treat as ground truth and another as the privatized data, and we compute the pie chart metric score (across all neighborhood/months). This effectively gives us the difference between two views of the same ground truth data. If the added privacy error is less than the sampling-error benchmark, the added privacy noise is comparable to variation typically encountered due to sampling.

Two 75% subsamples (a single visualized trial):



Two 50% subsamples (over many trials):



An Example! ...what we didn’t quite get around to:

Further Questions:

At our challenge launch deadline rapidly approaches, our analyses on the tuning properties of the Pie Chart metric are limited to the above. However, there are other interesting questions that could be explored with respect to this metric:

- How does the pie chart metric score change as the number of possible record types increases, given a similar power-law distribution of data?
- If we increase the MPP to 1.0 (effectively having zero tolerance for spurious labels), how does that affect the score distribution? How does it affect the 50% sampling error benchmark?
- What is the Jensen Shannon Distance score for distributions that closely match on all but one record label? What is the maximum error a single record label can have while maintaining a total Jensen Shannon score above the 50% sampling error benchmark?
- How sensitive is the Pie Chart metric to relatively small differences in map/time segments that have very few records? To what extent does setting the FT higher mitigate this effect?
- Can we define a minimum score threshold for “trustworthy” privatized pie charts, and then classify map/time segments as “feasible to privatize” by whether the sampling error benchmark achieves a trustworthy score on those segments?

Submissions to the metric challenge (with ample time between the challenge launch and the January submission deadline), should feel welcome to more fully explore interesting properties of their own metrics, either theoretically or empirically as appropriate. Questions relating to edge cases, impact on practical use cases, and useful or unexpected properties are all of interest.

The Big Objective Here

Every metric will have both capabilities and limitations; no single metric will capture all possible definitions of utility.

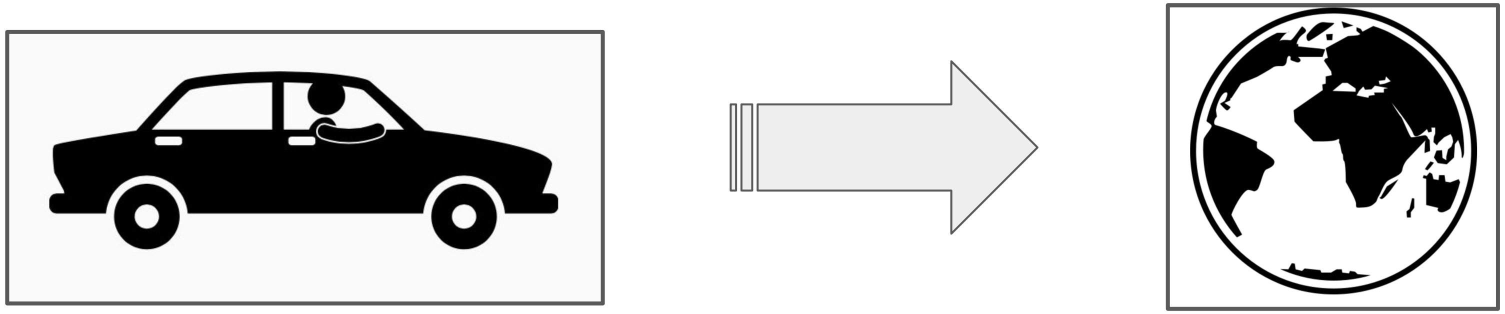
The overall objective of this challenge is to collect metrics that:

(1) Capture real world use cases and data stakeholder needs

(2) Are ***well defined***, and clearly written so that they are straightforward to implement correctly.

(3) Are ***well understood***, with analysis that explores both capabilities and limitations--blindspots, instability, biases, comparability properties….

Tips & Tricks: Defensive Driving for Metric Developers



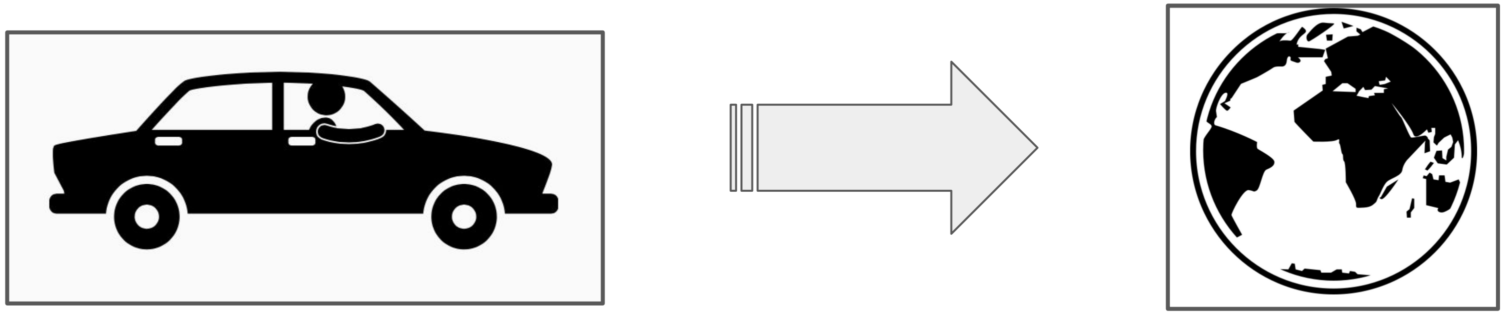
This is for real. We are really, really going to use these, and (if you consent) we are really, really going to put them in front of lots of other important people in the privacy research community so that they can use them too. You get a chance to do this write-up, we let you use all the color and pictures and words and pages and everything else you might want to use to get the word out about your idea, and then once you're done….

Your idea is going into other people's hands. It'll be passed around, pointed out over beers at conferences, mentioned briefly in undergrad lectures, cited in papers…. **and at some point it's going to get misused**.

How do you make sure your metric survives intact in the grapevine of a rapidly changing, rapidly growing, bleeding edge R&D field? By trying to find and clearly identify all the potential pitfalls yourself, and include them with the metric's definition so that people using your metric understand not just *how to implement it*, but also *how it works* and *where it doesn't*.

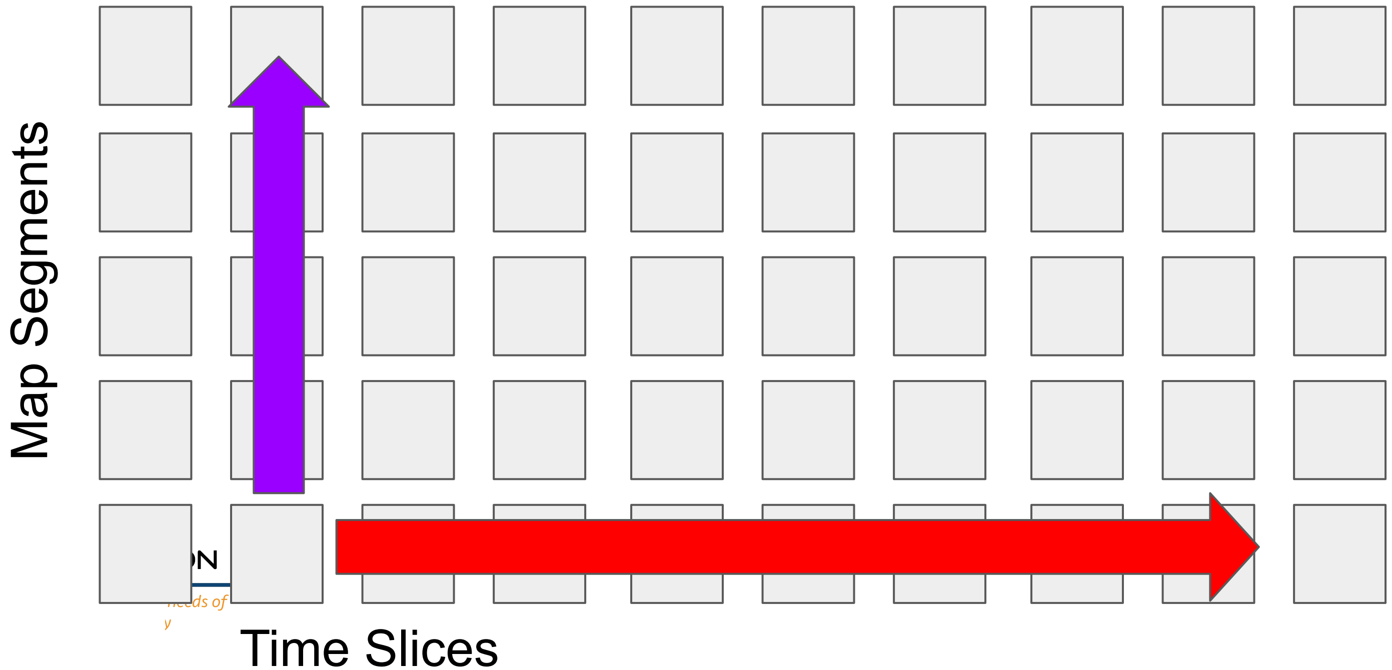
So here's some tips and things to keep in mind for how to do this.

Tips & Tricks: Time vs. Space



<u>Evaluation Space:</u>

Aggregation of Event Types by Time Slice and Map Segment

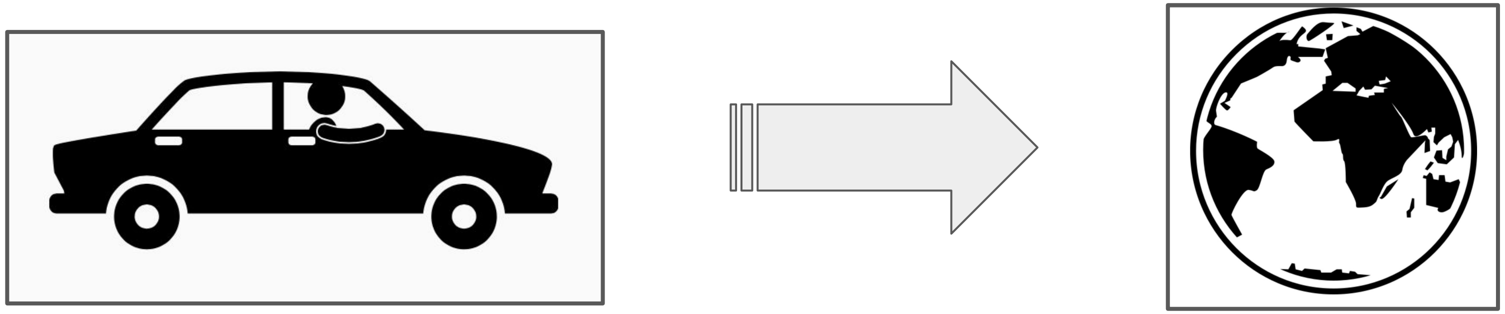


Remember the fun part of this challenge-- **Adventures in Space Time!**

How is your metric handling the difference between space and time? There will be geographic correlations in the data and temporal ones, and we want to make sure that all are preserved in the privatized data.

What part of this problem are you tackling? Are you focused only on map segments, and simply averaging across time? Or are you looking at trends through time and only averaging across map segments? Or are you handling both together?

Tips & Tricks: Ordinal vs Categorical



Data features come in two basic types:

Ordinals that have a natural order to them like numbers, dollar amounts, ages, poverty percentages, times, years, and even highest grade of education.

Categoricals that have no natural ordering: sex, race, language, ancestry, favorite websites, event code, map segment (with caveats).

How does your metric use these two types of variables? Does it only work with one type or the other? (← that's fine). As always, be clear.

Actual Data Table

Age (Number)	Gender (M/F)	Income (Number)	Attended University (T/F)
23	M	$73K	F
32	F	$65K	T
45	M	$84K	T
68	F	$112K	T
54	F	$91K	F

Synthetic Data Table

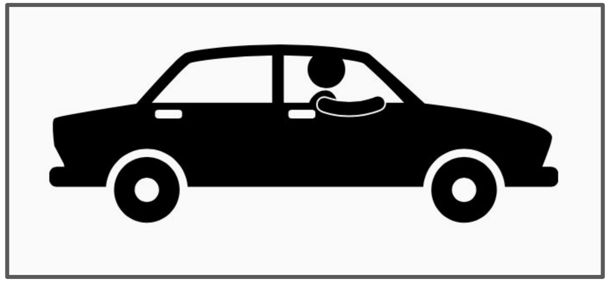
Age (Number)	Gender (M/F)	Income (Number)	Attende University (T/F)
23	M	$73K	F
32	F	$65K	T
45	M	$84K	T
68	F	$112K	T
54	F	$91K	F

Three Marginals Output from Step 1: *Actual and Synthetic Person Data Sources*

Gender (M/F)	Income (Number)	Attended University (T/F)	Actual Count	Synthetic Count
M	$0-33K	F		
F	$0-33K	F		
M	$0-33K	T		
F	$0-33K	T		
M	$34-66K	F		

3-marginal metric from the NIST Differential Privacy Synthetic Data Challenge
Uses binning to treat numerical variables like categorical variables.

Tips & Tricks: Ordinal vs Categorical *Error*



NOTE!

Ordinals have a natural definition of error, how far apart two values are, (A - B). .

Categoricals don't necessarily. You can look at things like edit distance, counts of the number of records with each value (as in pie chart and marginal-based techniques), or using them as ***class values in classification techniques***.

Understanding clearly how your metric operates on these two feature types is important.

Actual Data Table

Age (Number)	Gender (M/F)	Income (Number)	Attended University (T/F)
23	M	$73K	F
32	F	$65K	T
45	M	$84K	T
68	F	$112K	T
54	F	$91K	F

Synthetic Data Table

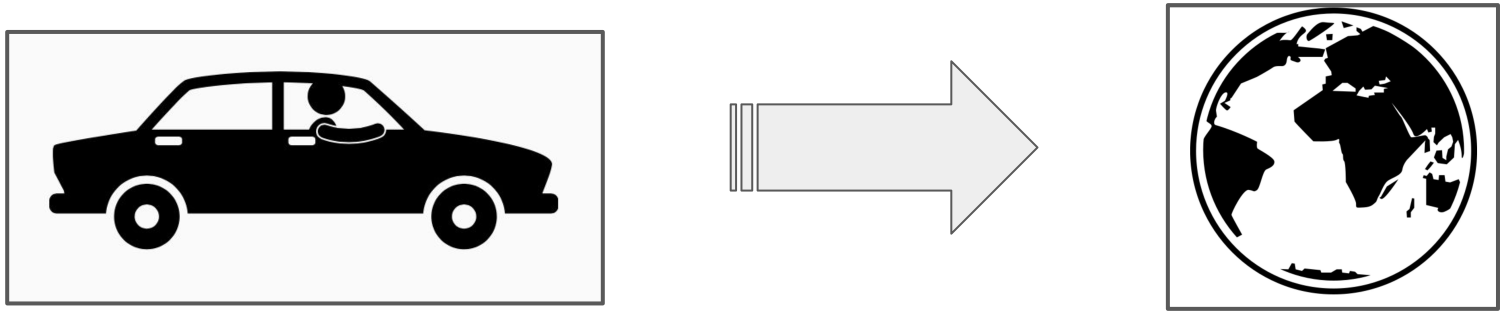
Age (Number)	Gender (M/F)	Income (Number)	Attende University (T/F)
23	M	$73K	F
32	F	$65K	T
45	M	$84K	T
68	F	$112K	T
54	F	$91K	F

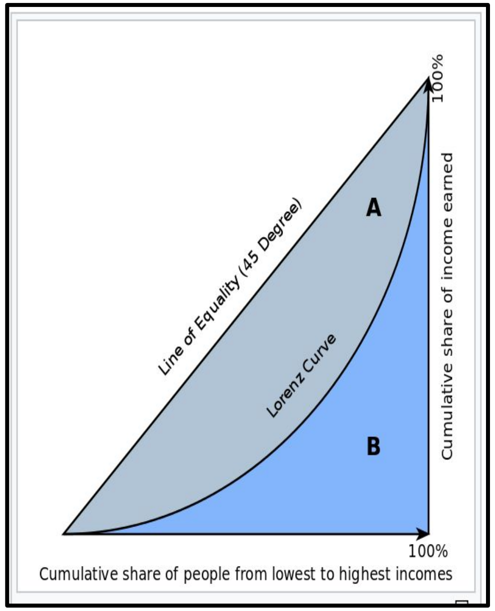
Three Marginals Output from Step 1: *Actual and Synthetic Person Data Sources*

Gender (M/F)	Income (Number)	Attended University (T/F)	Actual Count	Synthetic Count
M	$0-33K	F		
F	$0-33K	F		
M	$0-33K	T		
F	$0-33K	T		
M	$34-66K	F		

3-marginal metric from the NIST Differential Privacy Synthetic Data Challenge
Uses binning to treat numerical variables like categorical variables.

Tips & Tricks: Generalization and Configuration





https://en.wikipedia.org/wiki/Gini_coefficient

Income Inequality metric from the
NIST Differential Privacy Synthetic Data Challenge

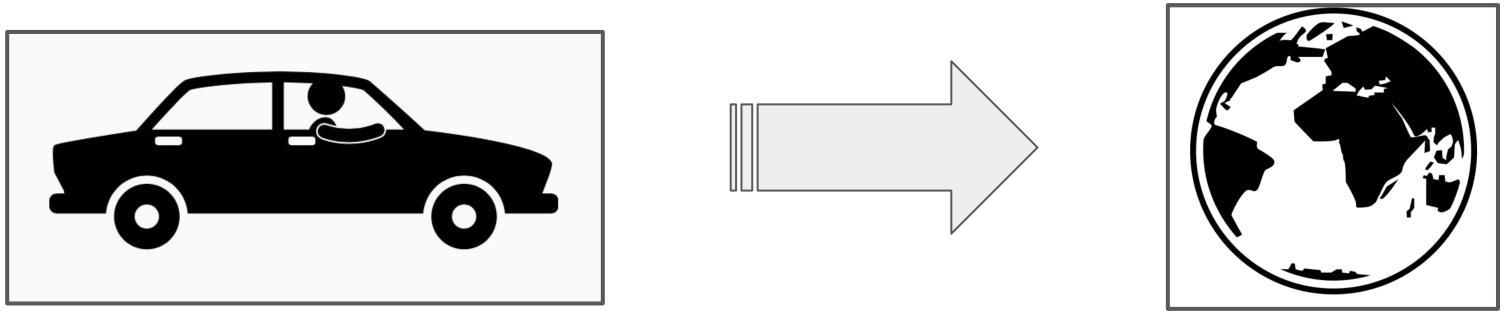
Do you have a great, specific real world use case in mind, such an income inequality, the pay gap, or anti-gerrymanding analytics… but it's highly dependent on the schema containing a specific set of features?

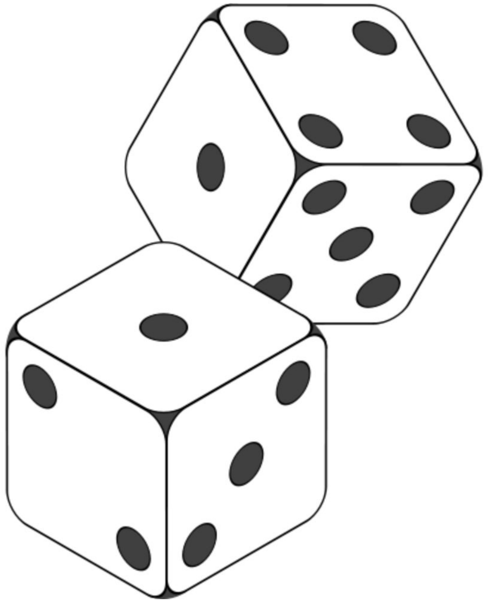
Consider generalizing it! If a use case generally runs on income, can it be run on any financial variable? Or even any numerical variable?

If a use case generally runs on sex or race, can it also be run on any demographic variable?

Merics that can be configured to run on many different schema can provide more comprehensive analysis and much better coverage.

Tips & Tricks: Randomization





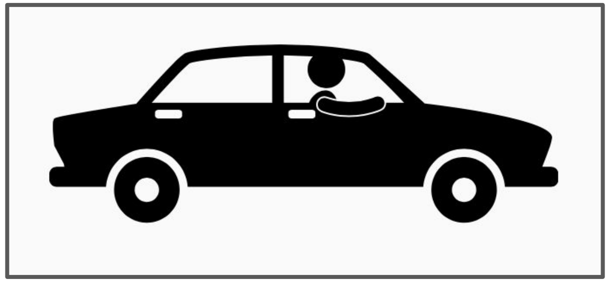
If you have a great idea for a comprehensive metric to evaluate the data, but it takes to long to run, and tends to choke and die if there’s too many features or too many records--

Consider Randomization!

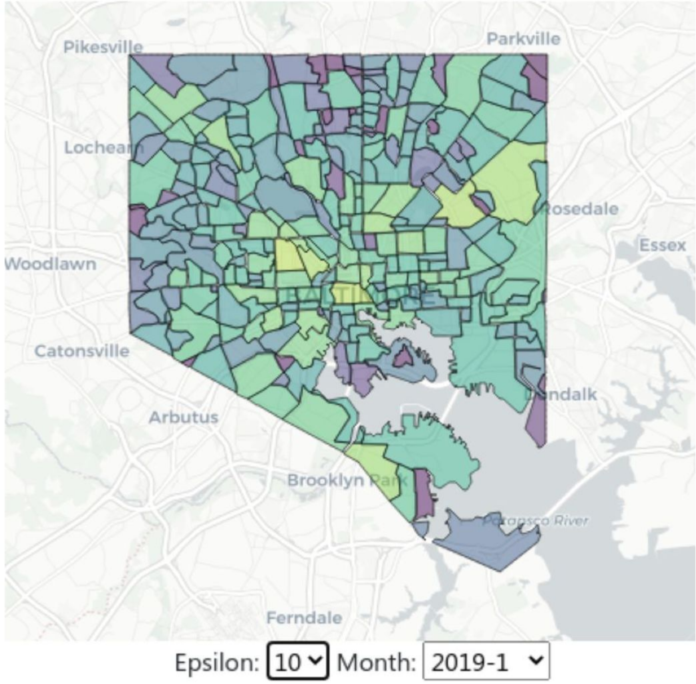
By randomly subsampling features or records, you can create a metric that gets a rapid high-level snapshot of the whole data set quality without exhaustively checking every possible combination.

Be careful to explore sampling ratios and stability, though! (more later)

Tips & Tricks: Snapshot and Deep Dive



The **Interactive Map** allows you to see your scores geographically (across all map segments). Here we see that dense urban neighborhoods closer to the city center, which generally contain more records, have better scores than rural and suburban neighborhoods where records may be more sparse. These are challenges that will need to be creatively overcome to achieve good performance on the Sprint 1 task.



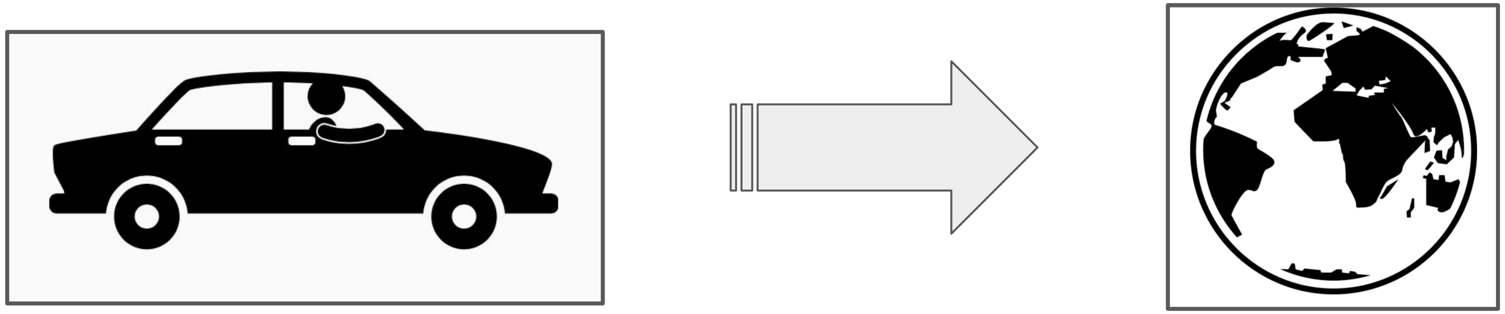
Snapshot vs Deep Dive!

Often metrics can be designed to either give a single total data quality score for a privatized temporal map (Snapshot Mode), or to investigate and pinpoint sources of disparity between the privatized and ground truth data (Deep Dive Mode).

How does your metric produce its single score?

Can you unroll your aggregation or refocus your metric to give more detailed information about specific points of failure?

Tips & Tricks: Snapshot and Deep Dive



The **Temporal Scores Chart** allows you to select a given neighborhood and see the change in your pie chart scores in that neighborhood over each of the time segments. Here we see the scores are relatively uniform across months for our baseline privacy algorithm. However, a privacy algorithm that leverages the temporal aspect of the problem, for example by aggregating counts across multiple time segments, might see more interesting variation here.

Remington (213)

year	month	score
2019	1	0.6073
2019	2	0.6631
2019	3	0.6635
2019	4	0.6849
2019	5	0.6235
2019	6	0.6508
2019	7	0.6536
2019	8	0.5685
2019	9	0.5944
2019	10	0.6263
2019	11	0.6558
2019	12	0.6480

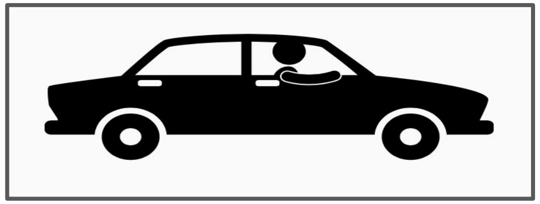
Snapshot vs Deep Dive!

Often metrics can be designed to either give a single total data quality score for a privatized temporal map (Snapshot Mode), or to investigate and pinpoint sources of disparity between the privatized and ground truth data (Deep Dive Mode).

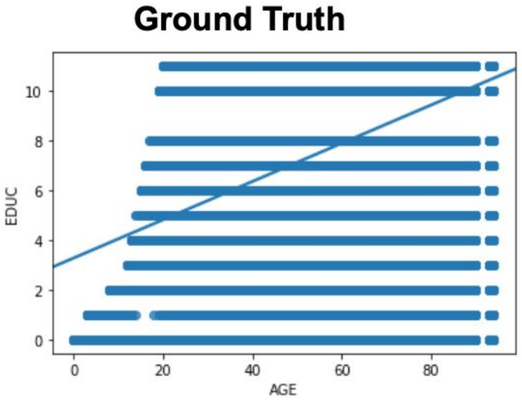
How does your metric produce its single score?

Can you unroll your aggregation or refocus your metric to give more detailed information about specific points of failure?

Tips & Tricks: Checking Blindspots (and Decision Boundaries)



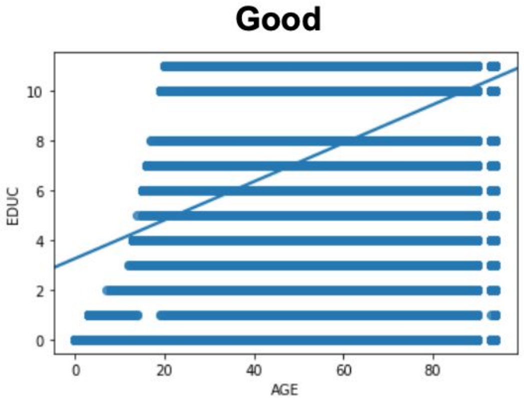
Age vs. Education



Coefficients:
[[0.07710821]]

Intercept:
[3.27240961]

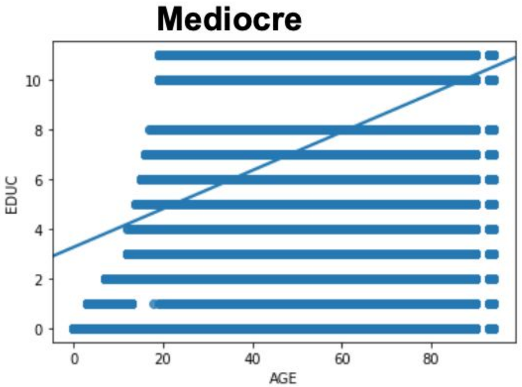
R-squared:
0.27983908685879266



Coefficients:
[[0.07751626]]

Intercept:
[3.25282563]

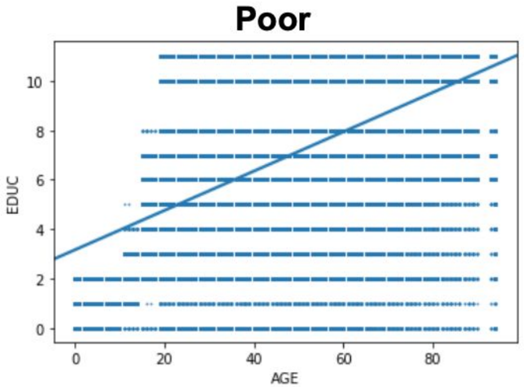
R-squared:
0.2825870270758932



Coefficients:
[[0.07741602]]

Intercept:
[3.25825086]

R-squared:
0.2820458675394747



Coefficients:
[[0.07959727]]

Intercept:
[3.16305652]

R-squared:
0.29869253706847976

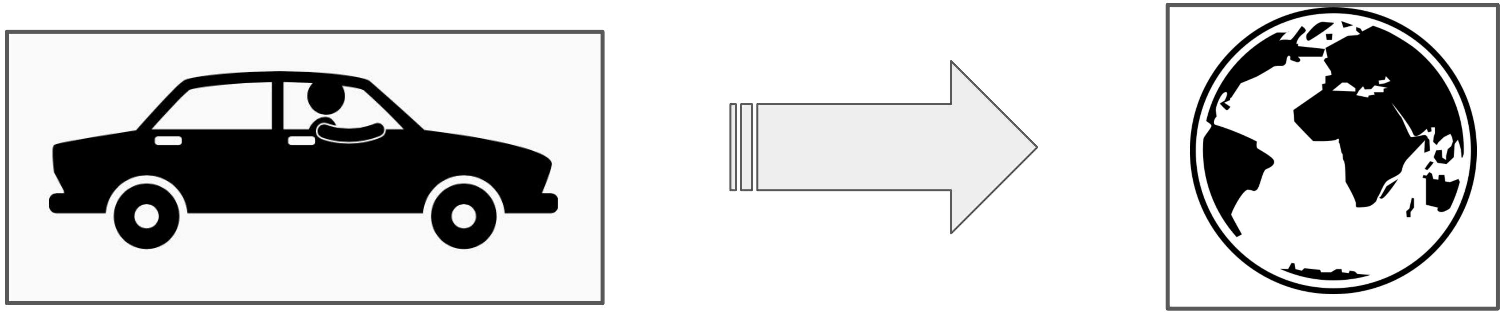
All reasonable metrics provide imperfect discriminative power, and that's fine-- **Do you know where your metrics blindspots are?**

Do you use binning on numerical variable, or threshold cut-offs like the pie chart metric? Bin sizes and thresholds are decision boundaries that create blind spots.

How does your metric aggregate information? Does it take an average, find a precentile, or fit a curve? What type of details is it glossing over when it does this?

Does your metric project data into euclidean (cartesian/vector) space? What information might be lost in that projection.

Tips & Tricks: Checking Edge Cases



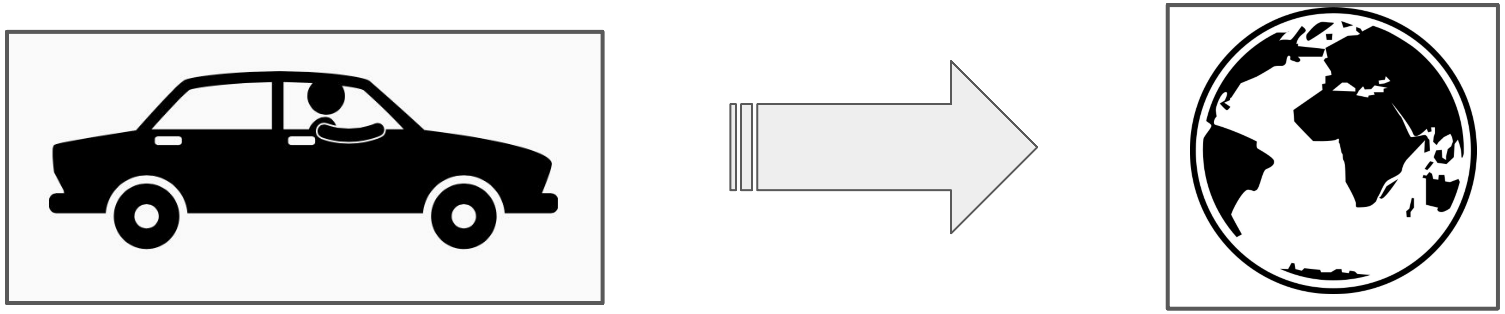
neighborhood	year	month	0	1	2	...	171	172	173
0	2019	1	0	0	0	...	0	0	0
0	2019	2	0	0	0	...	0	0	0
0	2019	3	0	0	0	...	0	0	0
...	...	...	...	...	...	...	...	...	...
277	2019	10	0	0	0	...	0	0	0
277	2019	11	0	0	0	...	0	0	0
277	2019	12	0	0	0	...	0	0	0

What happens to your metric when the ground truth is full of zeros, and the privatized data isn't? What about when there's only a single record? What happens when the privatized data has many, many more records than the ground truth?

What if the input schema only has a single numerical feature, and the rest are categorical? What if it only has one categorical feature and the rest are numerical?

Doing a good debugging on your metric is a good idea to avoid unexpected and alarming behavior down the road. Think carefully through how your metric behaves at extreme or unusual inputs. Make sure you clearly identify any assumptions you're making about what inputs are valid.

Tips & Tricks: Checking Stability



$$\frac{3}{4} = 0.75$$

$$\frac{399}{400} = 0.9975$$

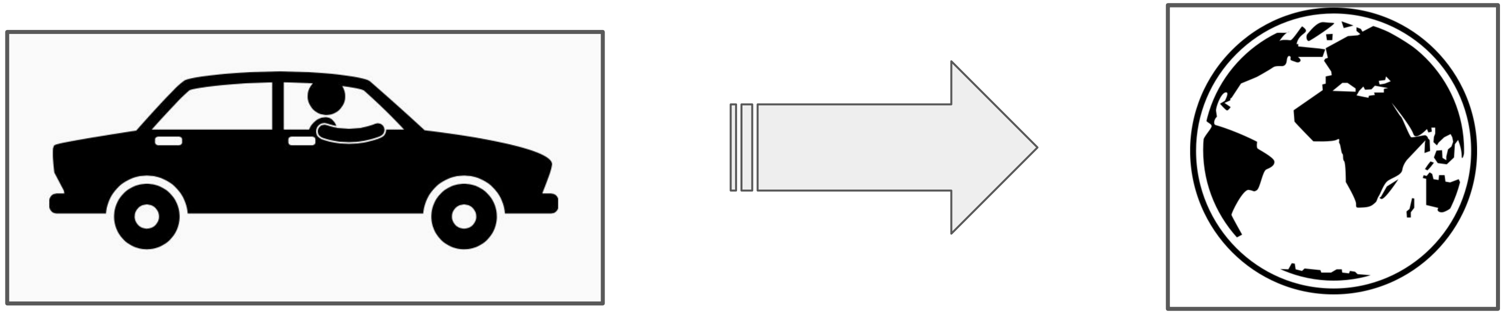
Ratios get strange when the numbers are small.
Randomization, if you're using too small a sampling ratio, can produce wildly different answers depending on what sample you get.

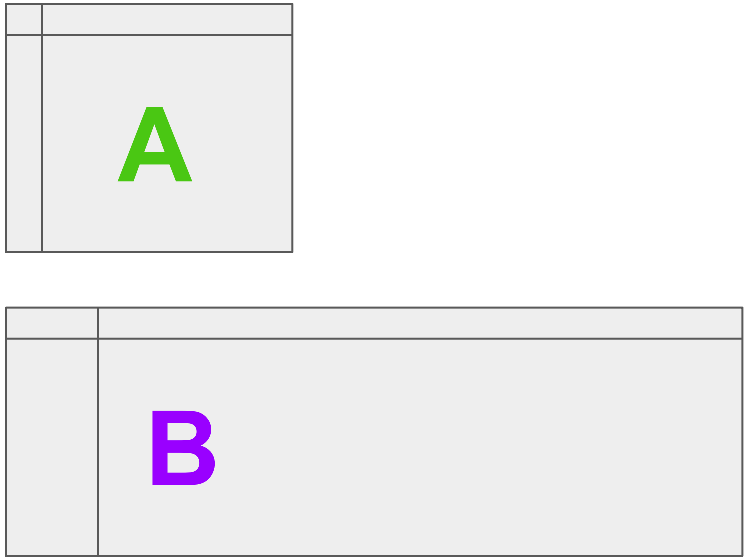
How stable is your metric?

Run it multiple times on the same input (if randomized) and check the distribution.
See how it behaves on data sets at the extremes (very sparse data, very dense data).

It doesn't need to work perfectly everywhere, but we need to understand in what contexts the results are stable and dependable, and in what contexts we may need to run multiple trials, or go with a different metric.

Tips & Tricks: Checking Comparability





Take a look at your metric and check this real quick-- How do the numbers change depending on the size of the input data? The number of possible record types? The number of numerical features vs. categorical features? How many zeros (sparseness) there is in the ground truth data?

When you get a score of 700 on a data-set in Schema A, and a score of 600 on a data-set in Schema B, does it really mean that the second data set is worse quality? Or does it just mean that the second data-set is *larger*?

How do your metric scores change dependent on the schema of the data, independent of the data quality itself?

It's fine if your metric isn't comparable between different data schemas, but understanding those properties is important to ensuring your metric isn't accidentally misused to produce misleading or invalid performance rankings.

Let’s Get Going!

NASA Tournament Lab | DRIVENDATA | herox

Overview | Guidelines | Timeline | Updates 5 | Forum 2 | Community 47 | Entries | Resources | FAQ

Challenge Overview

This challenge is Part 1 of a multi-phased challenge. To participate in the other stages please visit https://deid.drivendata.org. See the complete challenge rules here. Additionally, further details about the other stages can be found below.

The Public Safety Communications Research Division (PSCR) of the National Institute of Standards and Technology (NIST) invites members of the public to join the Differential Privacy Temporal Map Challenge (DeID2). This multi-stage challenge will award up to $276,000 to advance differential privacy technologies by building and measuring the accuracy of algorithms that de-identify data sets containing temporal and geographic information with provable differential privacy.

The DeID2 Challenge is composed of three contests:

PSCR OPEN INNOVATION
Solving the Public Safety needs of tomorrow...today

Let’s Get Going!

Overview | Guidelines | **Timeline** | Updates 5 | Forum 2 | Community 47 | Entries | Resources | FAQ

Challenge Timeline

Start

Sept. 17, 2020, 9:17 a.m. PDT
Date Launched

Oct. 1, 2020, 6 a.m. PDT
Enter

Oct. 2, 2020, 6:24 a.m. PDT
You registered for challenge

Oct. 20, 2020, 9:20 a.m. PDT
You are here

Nov, 30 2020, 7pm PDT
Executive Summaries due for optional preliminary review

PSCR OPEN INNOVATION
Solving the Public Safety needs of tomorrow...today

Let's Get Going!

Overview | Guidelines | Timeline | Updates 5 | Forum 2 | Community 47 | Entries | Resources | FAQ

Challenge Forum

Search forum

New topic

Forum Sections	Threads	Total Posts
Challenge Guideline Clarifications Post any requested clarifications to the challenge guidelines here.		
Challenge Eligibility, Feedback or Website Issues Post any potential website issues here, or general feedback on the challenge (like you want to submit multiple submissions).	1 Thread	4 Posts

Let's Get Going!

Overview | Guidelines | Timeline | Updates 5 | Forum 2 | Community 47 | Entries | Resources

Challenge Resources

Resources

+ Add resource

Example Temporal Map Data
Oct. 1, 2020
Leave a comment

Tips and Tricks to Submitting
Sept. 30, 2020
Leave a comment

Sample Submission - Pie Chart Metric
Sept. 30, 2020
Leave a comment

Submission Template
Sept. 30, 2020
Leave a comment

Important Dates

Registration Opens	October 1, 2020
Executive Summaries due for optional preliminary review	November 30, 2020
Webinar 2	December 4, 2020
Submissions due	January 5, 2021
NIST PSCR Compliance check (for public voting)	January 5-6, 2021
Public voting	January 8-21, 2021
Judging and Evaluation	January 5 - February 2, 2021
Winners Announced	February 4, 2021

PSCR OPEN INNOVATION
Solving the Public Safety needs of tomorrow...today

Questions?

Competition Details and Official Rules

Challenge.gov
https://www.challenge.gov/challenge/differential-privacy-temporal-map-challenge/

HeroX
https://www.herox.com/bettermeterstick

DrivenData
https://deid.drivendata.org/

Challenge Questions
PSPrizes@nist.gov

PSCR OPEN INNOVATION
Solving the Public Safety needs of tomorrow...today

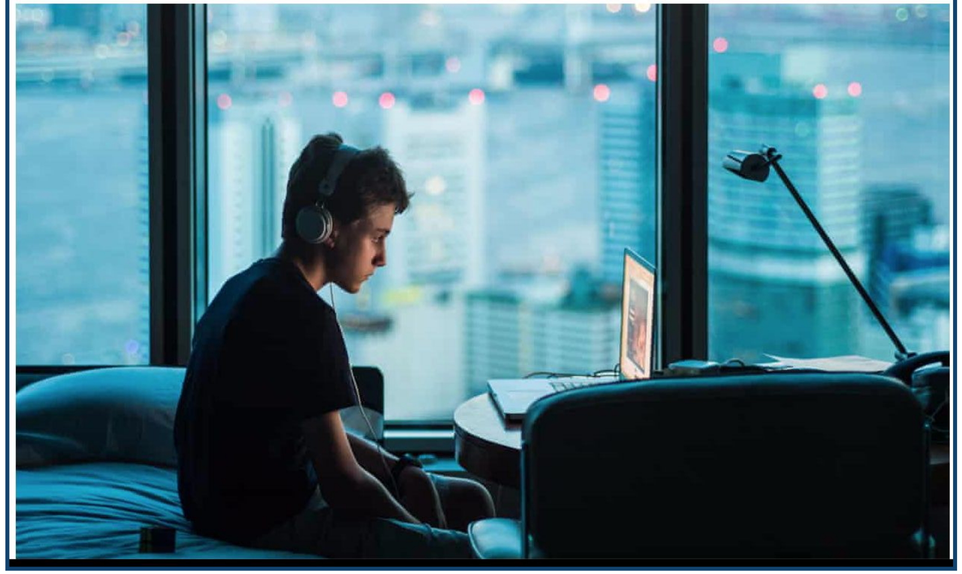
Thank you!

Appendix

Attacks on Privacy: De-anonymization

'Data is a fingerprint': why you aren't as anonymous as you think online

So-called 'anonymous' data can be easily used to identify everything from our medical records to purchase histories



Keeping Secrets: Anonymous Data Isn't Always Anonymous

March 12, 2014 by datascience@berkeley Staff

ars TECHNICA BIZ & IT TECH SCIENCE POLICY CARS GAMING & CULTURE STO

POLICY —

"Anonymized" data really isn't—and here's why not

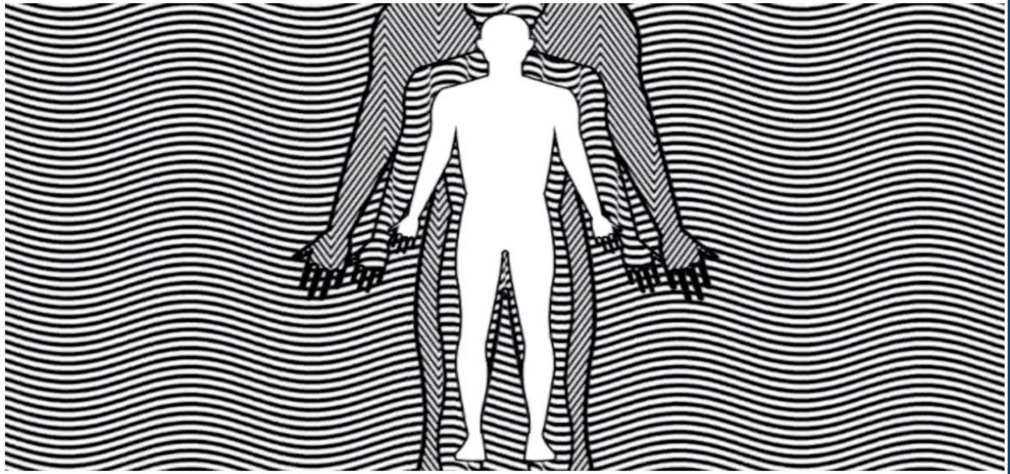
Companies continue to store and sometimes release vast databases of " ...

NATE ANDERSON - 9/8/2009, 7:25 AM

12.10.18

Sorry, your data can still be identified even if it's anonymized

Urban planners and researchers at MIT found that it's shockingly easy to "reidentify" the anonymous data that people generate all day, every day in cities.



De-anonymization New York Taxi Data

New York taxi details can be extracted from anonymised data, researchers say

FoI request reveals data on 173m individual trips in US city - but could yield more details, such as drivers' addresses and income

▲ Data about New York city taxi drivers and rides could be de-anonymised, researchers warn. Photograph: Jan Johannessen/Getty Images Photograph: Jan Johannessen/Getty Images

Alex Hern

@alexhern

Fri 27 Jun 2014 10.57 EDT

"Using a simulation of the medallion data, we show that our attack can re-identify over 91% of the taxis that ply in NYC even when using a perfect pseudonymization of medallion numbers."

Douriez, Marie, et al. "Anonymizing nyc taxi data: Does it matter?." *2016 IEEE international conference on data science and advanced analytics (DSAA)*. IEEE, 2016.

Temporal Map Data

Public Safety Uses:

- Policy (e.g. resource allocation)
- Incident Management (e.g. evacuation plan)
- Analytics

Privacy Risks:

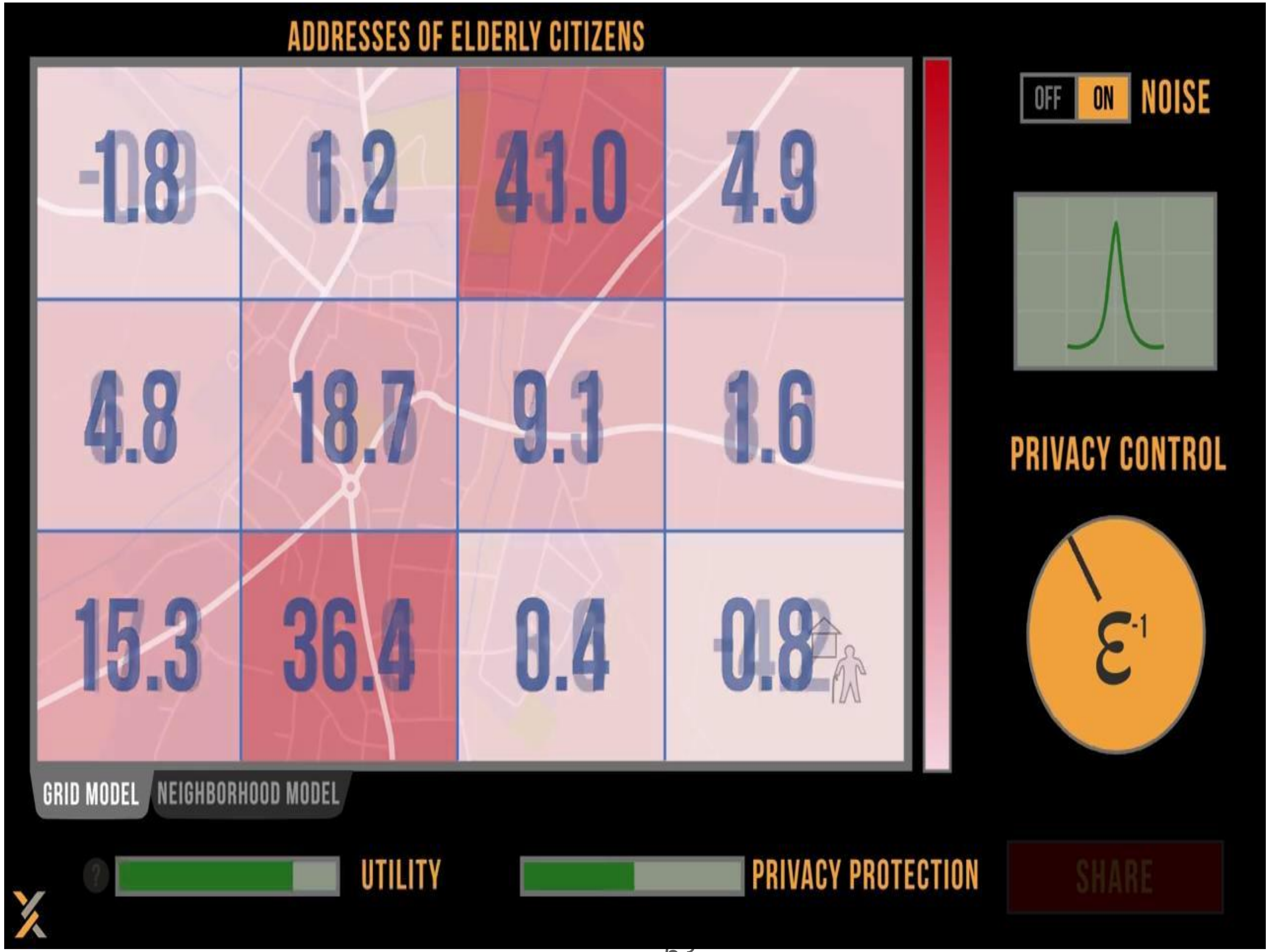
- Data sets may contain PII
- Linkage attacks can use location data to find a person
- Location history may contain sensitive information

Data Challenges:

- Data space scales with number of locations
- Data space scales *exponentially* with individual sequence length.
- Variability in map segments require flexible solutions

PSCR OPEN INNOVATION
Solving the Public Safety needs of tomorrow...today

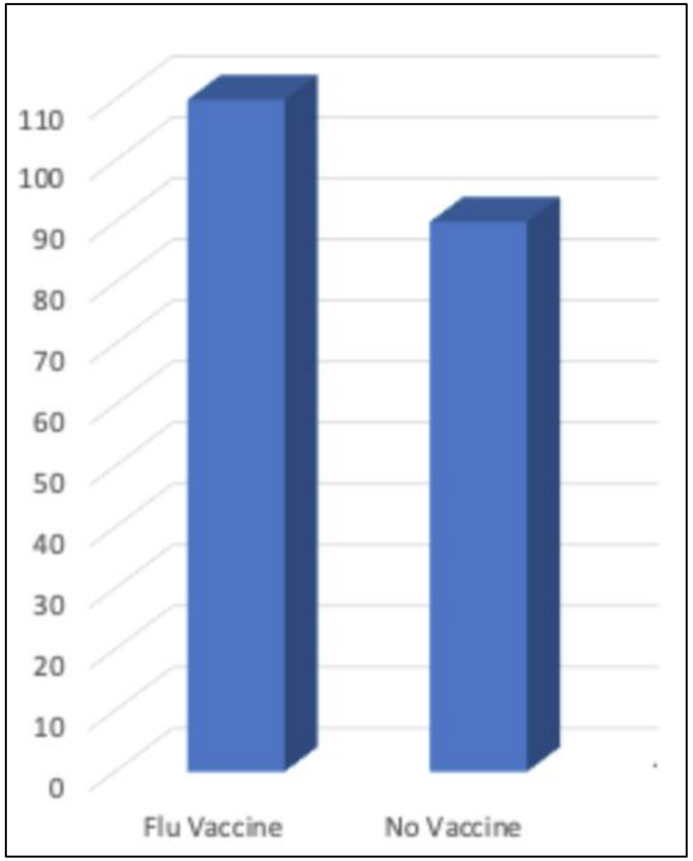
Differential Privacy Explainer Video

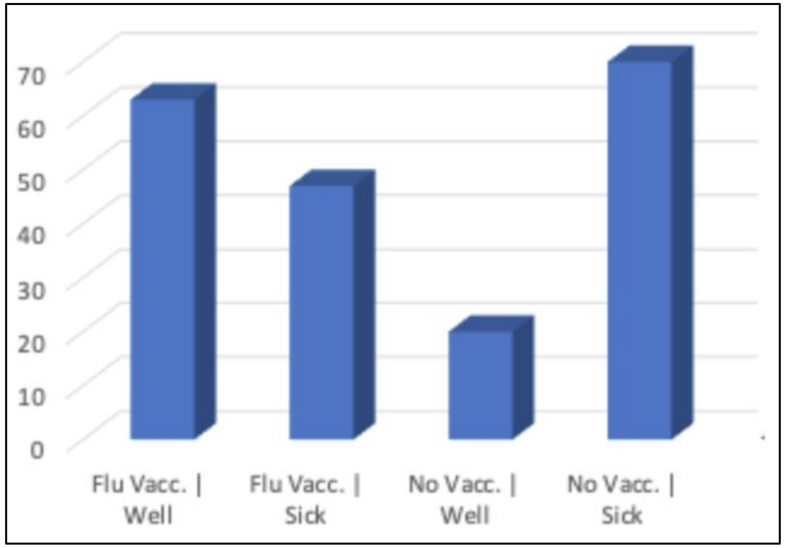


Temporal Map Data Technicalities

Data Challenges:

- Data ***space scales with number of locations***
- Data space scales exponentially with length of individual time sequences.
- Variability in map segments require flexible solutions



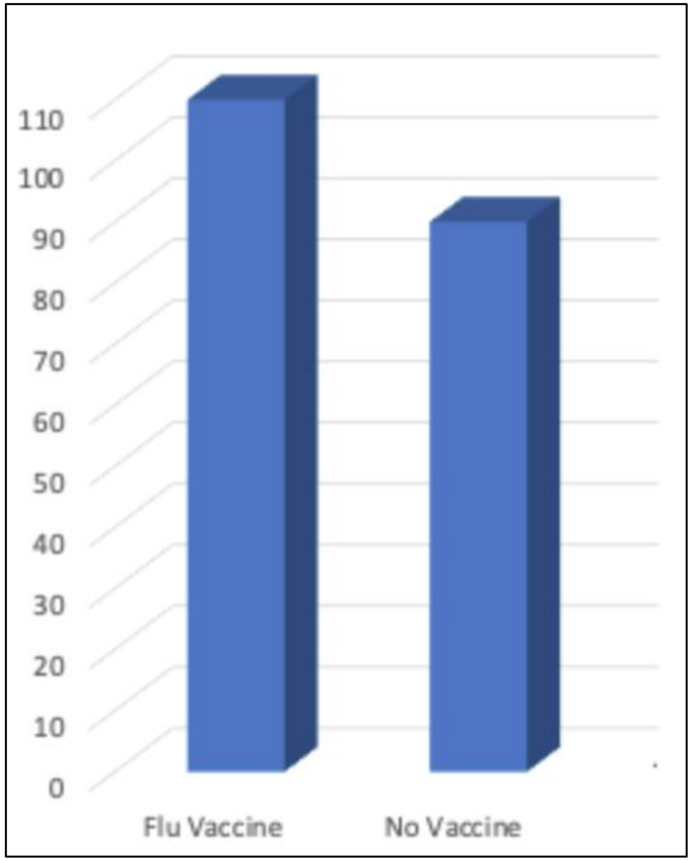


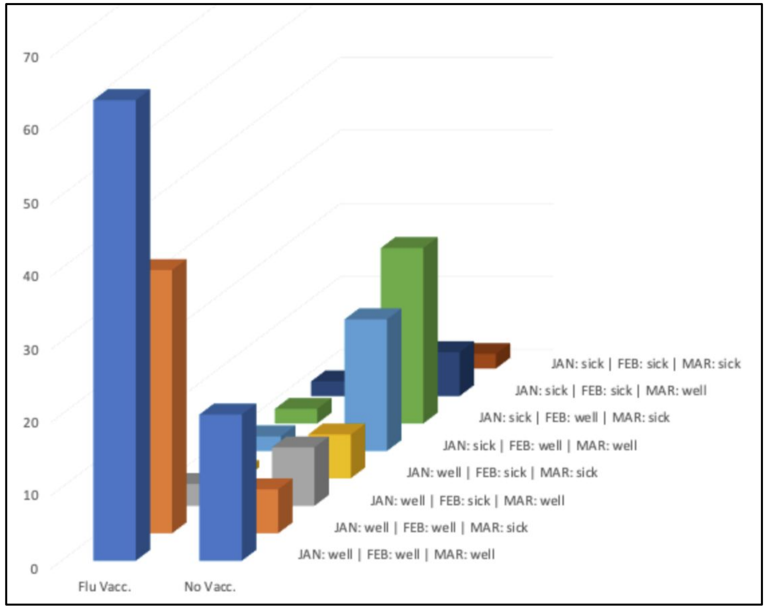
PSCR OPEN INNOVATION
Solving the Public Safety needs of tomorrow...today

Temporal Map Data Technicalities

<u>Data Challenges:</u>

- Data space scales with number of locations
- Data space scales ***exponentially*** with length of individual time sequences.
- Variability in map segments require flexible solutions



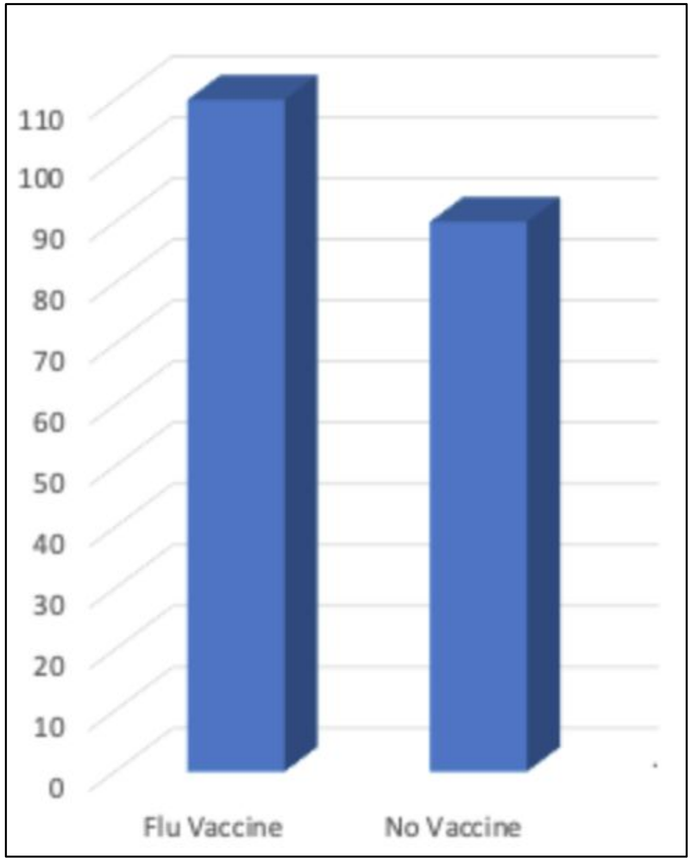


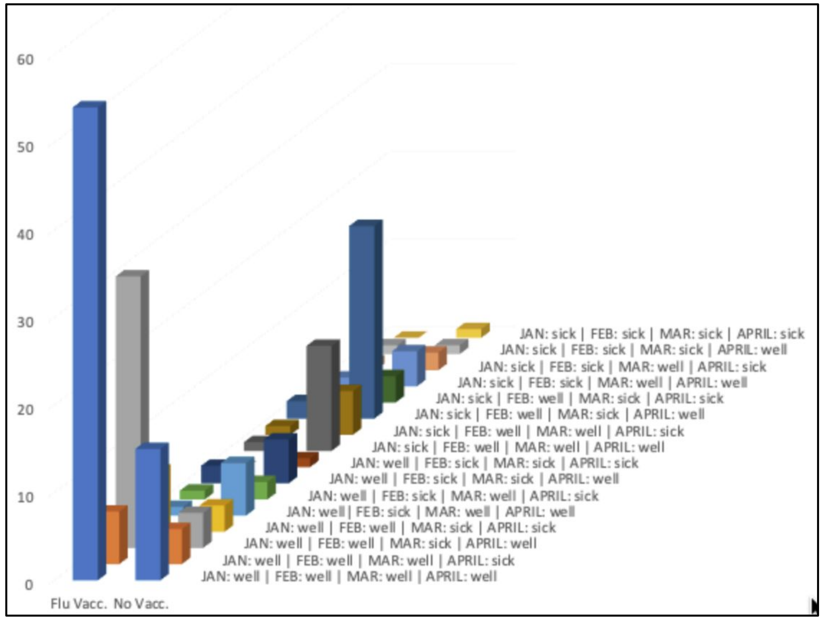
PSCR OPEN INNOVATION
Solving the Public Safety needs of tomorrow...today

Temporal Map Data Technicalities

Data Challenges:

- Data space scales with number of locations
- Data space scales ***exponentially*** with length of individual time sequences.
- Variability in map segments require flexible solutions





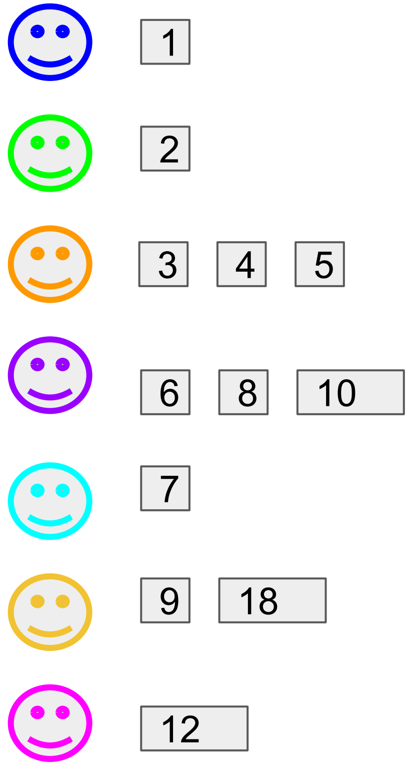
PSCR OPEN INNOVATION
Solving the Public Safety needs of tomorrow...today

Problem Definitions: Temporal Map Data

- In Sprint 1, using the Baltimore 911 Call Database for data. Time segments are months and map segments are neighborhoods.
- Input data given to competitors as a CSV file, Event Record File
- Event event record will include a tag/serial number for an ‘Individual’ (artificially generated). Privacy is protected at the Individual level.
- Note that max records per individual determines ‘sensitivity’ for differential privacy-- the amount of noise needed to privatize, and the difficulty of the problem. More records/individual is much harder.
- Output scored as aggregated call record types in each neighborhood, in each month.
- SME-proposed scoring function tested on (SME-proposed) naive baseline privatization code.

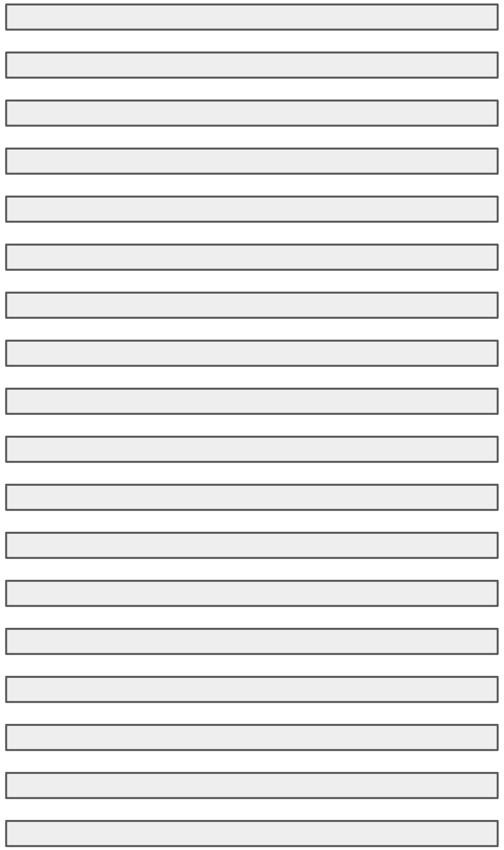
Individual Sequences:

Mapping between Individuals and Event Records. This will be the unit of privacy protection.



Raw Event Records:

Serial Number, Timestamp, Map segment ID, Event Info



Evaluation Space:

Event Info Aggregation, per Map Segment x Time Range