PSCR 2021

THE DIGITAL EXPERIENCE

#PSCR2021 • PSCR.GOV

NIST

PSCR ANALYTICS

PSCR

REMOVING PERSONALLY IDENTIFIABLE INFORMATION FROM PUBLIC SAFETY DATA SETS

Gary Howarth, PhD PSCR, Prize Manager

Christine Task, PhD Knexus Research, Computer Scientist

#PSCR2021

NIST

PSCR

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.

Christine Task, Knexus Research Corporation, produced and presented the video on slides 12-36, performing this work under financial assistance award 1333ND21FNB670041 from U.S. Department of Commerce, National Institute of Standards and Technology. Knexus Research Corporation produced and presented the video on slide 14 for publication in the National Institute of Standards and Technology's PSCR 2021 The Digital Experience. The contents of their presentation do not necessarily reflect the views or policies of the National Institute of Standards and Technology or the U.S. Government. Posted with Permission.

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

GOALS

- Identify the difficulty Public Safety faces in sharing data
- Define Privacy
- Introduce Differential Privacy
  - Types of DP outputs
- Some results from the Differential Privacy Temporal Map Challenge
- Resources:
  - Introduce the NIST Privacy Collaboration Space
  - Introduce Tech Demos of software

WHAT’S THE PROBLEM?

Public Safety as Data Generators

- As Public Safety entities has made 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.

DATA COLLECTED BY PUBLIC SAFETY

Calls for Service

- Calls to “911” for emergency assistance
- May include non-emergency calls
- Typically maintained in law enforcement computer-aided dispatch (CAD) systems

Incidents

- Collected by an agency for management
- Stored in Records Management Systems (RMS)
- Officer reports on crimes, situations, concerns, suspects, citizen public safety issues, etc.

Stops, Citations, Arrests

- Proactive and reactive stop of pedestrians or motor vehicles
- May be resolved through warnings, citations, summons, or physical arrests
- Data may be overlapping such as a stop followed by a citation or arrest

Complaints

- Potential mistreatment by authorities
- Policy, procedure, and legal violations
- May include Internal Affairs investigations
- Collection process required by national law and accreditation standards

WHY RELEASE DATA ?

Analytics

Many cities are developing algorithms to analyze crime, fire, and health data. Developers would like to access other localities' data for training, analysis, and validation.

FEDERAL BUREAU OF INVESTIGATION
UNIFORM • CRIME • REPORTING
NIBRS
NATIONAL INCIDENT BASED REPORTING SYSTEM

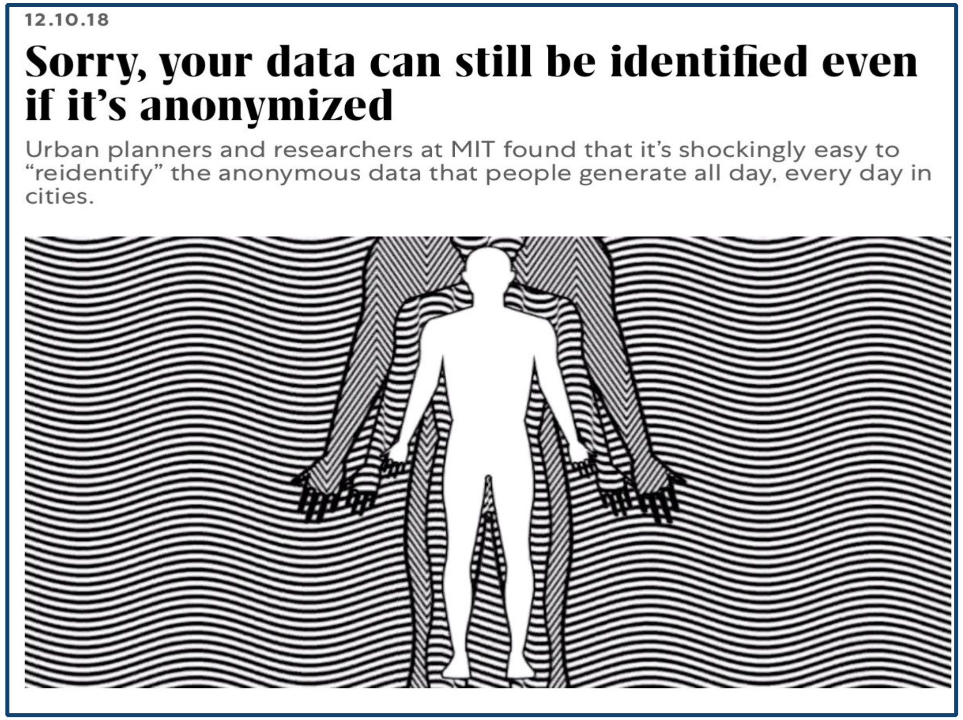
Open Access to Data

Many public safety agencies are required to report certain data. Others wish to share data with the public and researchers.

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



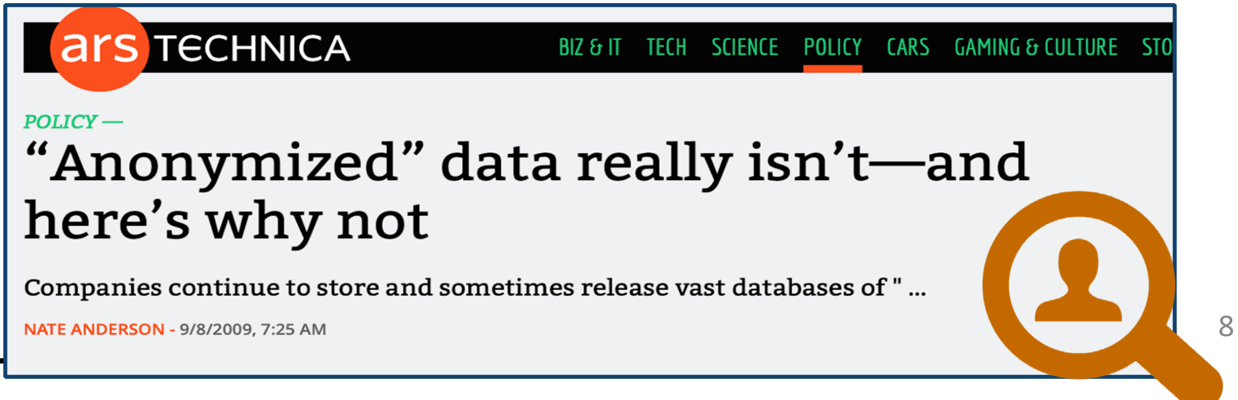
12.10.18

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.

Keeping Secrets: Anonymous Data Isn't Always Anonymous

March 12, 2014 by datascience@berkeley Staff



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

DE-ANONYMIZATION
NEW YORK TAXI DATA

“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.

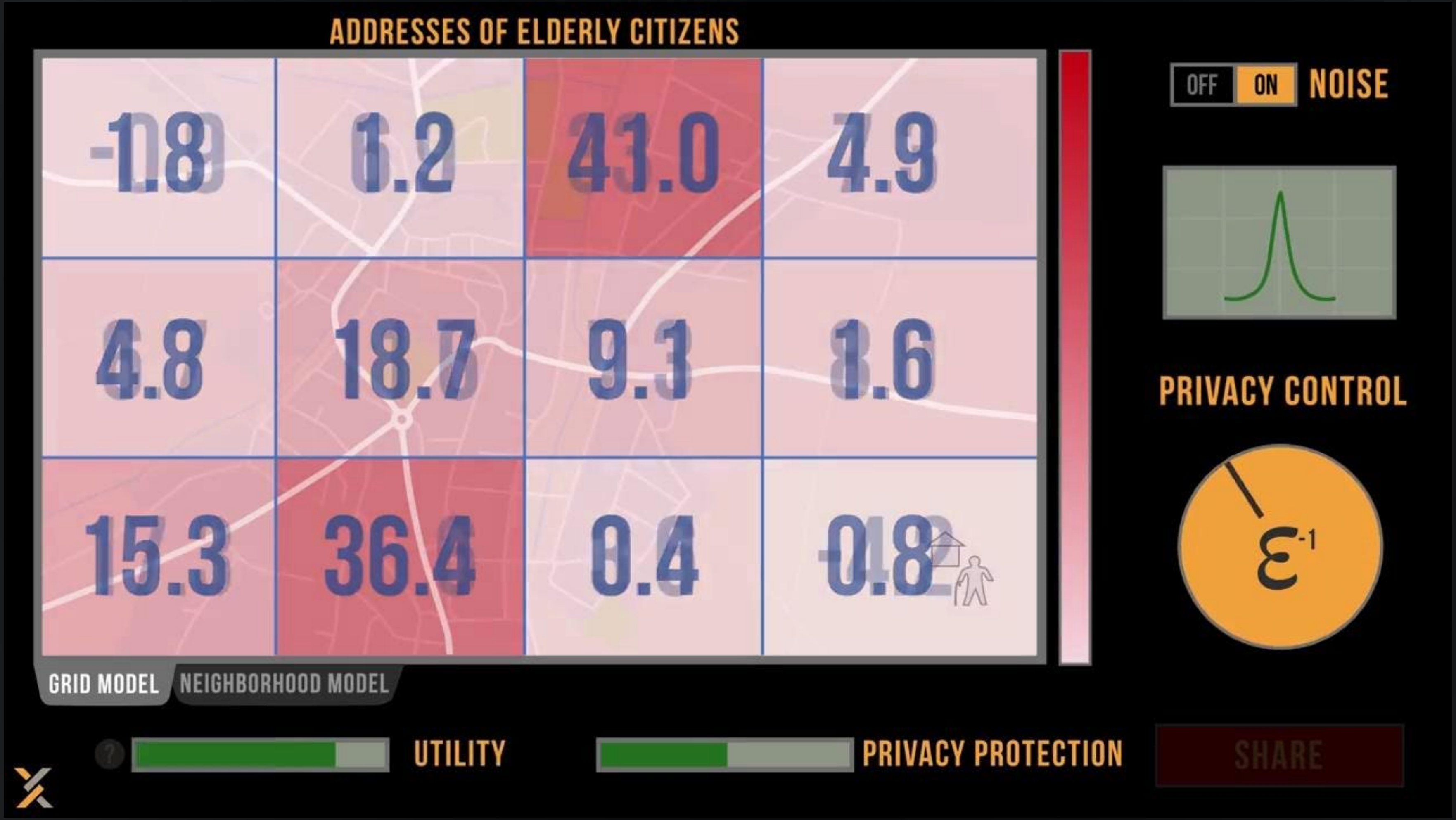
WHAT DO WE MEAN BY PRIVACY?

Privacy-preserving data-mining algorithms allow trusted data-owners to release useful, aggregate information about their data-sets (such as common user behavior patterns) while at the same time protecting individual-level information.

Intuitively, the concept of making large patterns visible while protecting small details makes sense. You just 'blur' things a bit:

If we refine this idea into a mathematically formal definition, we can create a standard for individual privacy.

FORMAL PRIVACY: DIFFERENTIAL PRIVACY GUARANTEE



ADDRESSES OF ELDERLY CITIZENS

1.8	1.2	41.0	4.9
4.8	18.7	9.1	1.6
15.3	36.4	0.4	0.8

NOISE: OFF [✓ ON]

FORMAL PRIVACY: DIFFERENTIAL PRIVACY GUARANTEE

Differential Privacy is a standard that protects privacy no matter what third-party data is available. It does so by strictly limiting what it is possible to learn about any individual in the data set.

HOW PRIVATE IS PRIVATE?

- Differential privacy algorithms privatize records by adding noise to a dataset. The more noise added, the more remote the new dataset is from the original, providing more privacy.
- The amount of noise added to the dataset is characterized by epsilon (***ε***)
- The lower the ***ε***, the more noise is added, producing less precise data
- With ***ε*** = ∞, no noise is added, outputting the original data
- ***ε*** scales exponentially
- The value of ***ε*** must be ‘tuned’ heuristically, balancing the risk of privacy against the utility of the data.

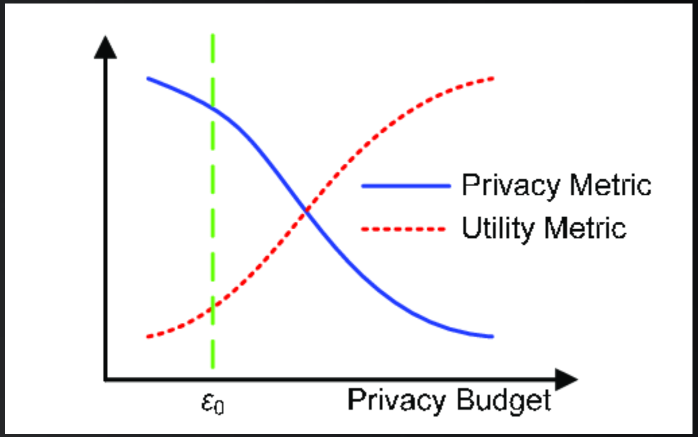


Illustration of the privacy-utility tradeoff. From Liu et al. “Privacy-Preserving Monotonicity of Differential Privacy Mechanisms.” 2018.

ε in The Wild	
2020 Census demo data (people)	10.3
2020 Census demo data (housing)	1.9
Apple emoji prediction	4

TUNING FOR PRIVACY

- Factors to consider when tuning privacy:
  - How ‘rich’ is the data? (e.g., how many and what types of (potential) identifiers are there?)
  - What are the risks of exposure? (e.g., released name v. released bank records, agency credibility)
  - How valuable are precision and accuracy to analysis of the privatized records?
- The Better Meter Stick Contest sought novel ways of evaluating the privacy utility trade-off (more on this later in the presentation).

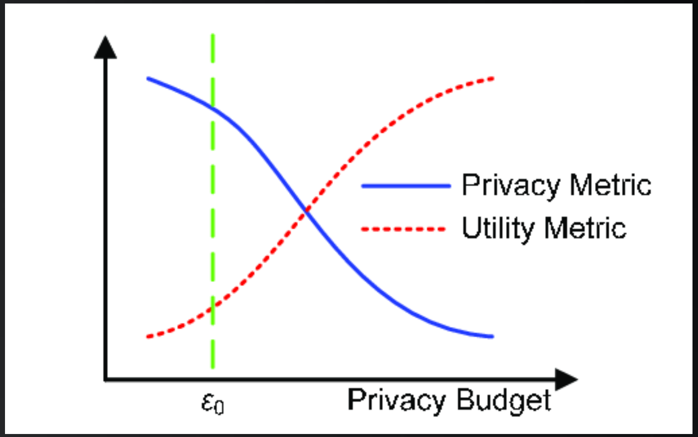


Illustration of the privacy-utility tradeoff.
From Liu et al. “Privacy-Preserving Monotonicity of Differential Privacy Mechanisms.” 2018.

CHALLENGE OBJECTIVE

In the Differential Privacy Temporal Map Challenge (DeID2), your task 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

We evaluate all submissions at three different levels of privacy. We chose to test at privacy-loss levels used by both academic research ($\varepsilon = 0.1$, $\varepsilon = 1$) and industry products ($\varepsilon = 10$).

1

Privacy write-ups

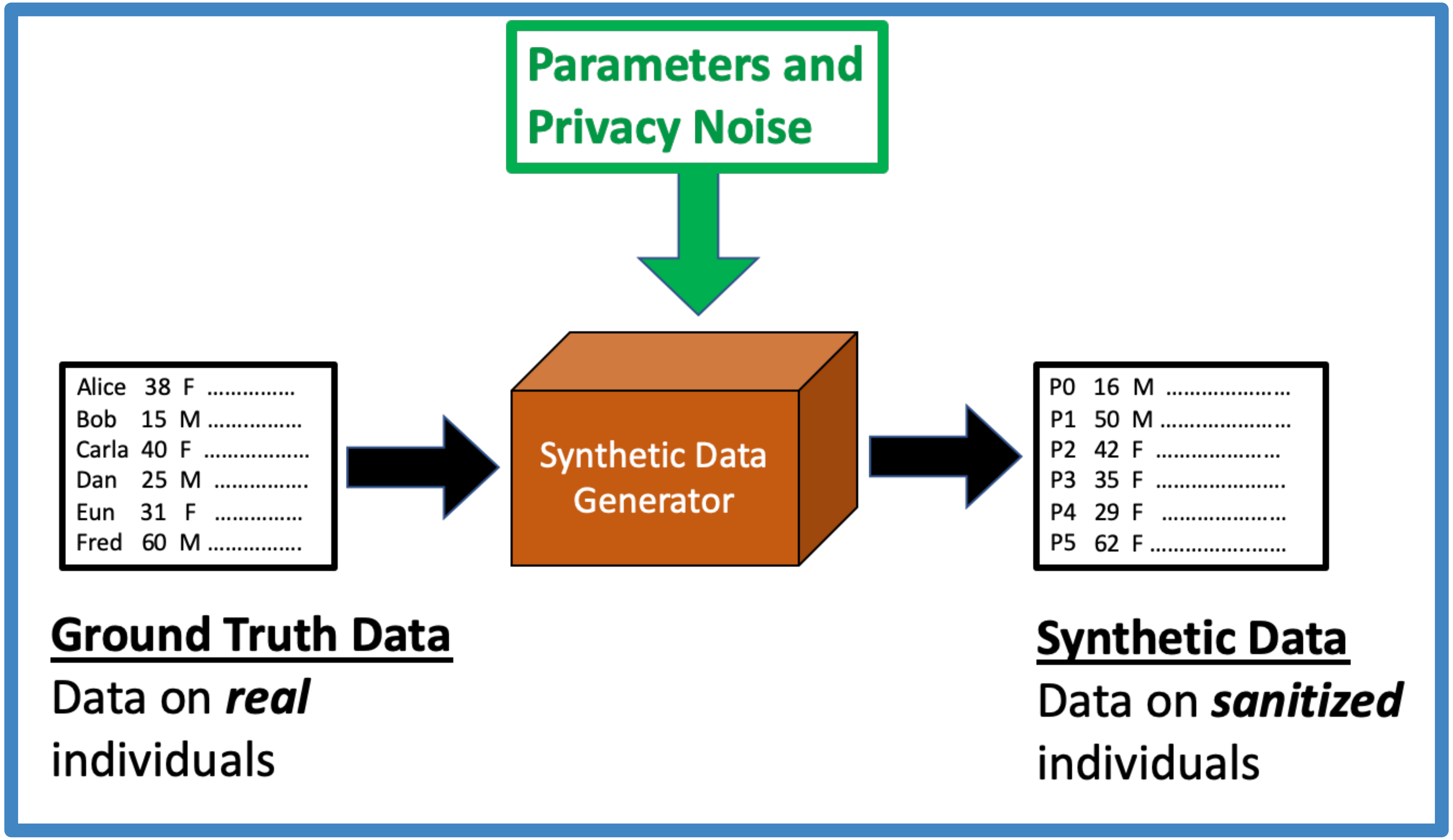
Confirmed by subject matter experts

2

Algorithm submissions

Evaluated by published performance metric

2018 PSCR DIFFERENTIAL PRIVACY CHALLENGE: Synthetic Data



Parameters and Privacy Noise

Alice	38	F	....
Bob	15	M	....
Carla	40	F	....
Dan	25	M	....
Eun	31	F	....
Fred	60	M	....

Synthetic Data Generator

P0	16	M	....
P1	50	M	....
P2	42	F	....
P3	35	F	....
P4	29	F	....
P5	62	F	....

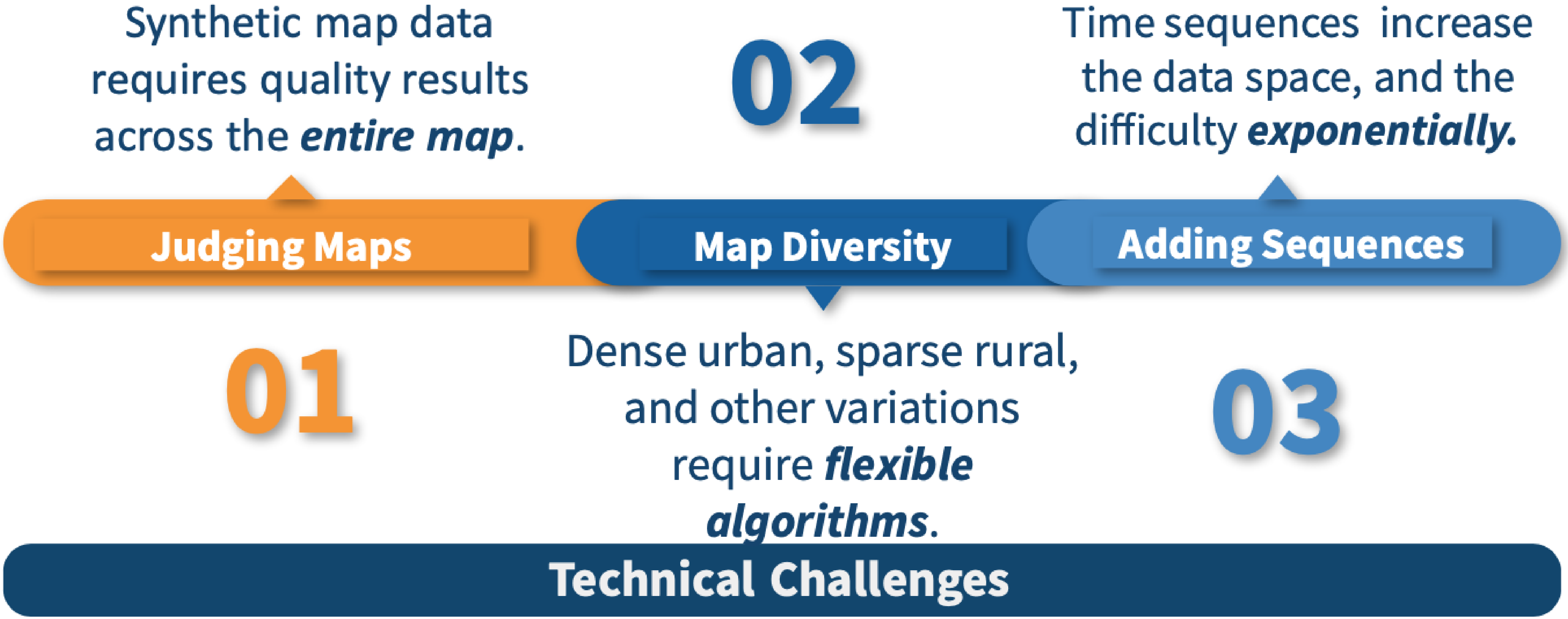
Ground Truth Data

Data on ***real*** individuals

Synthetic Data

Data on ***sanitized*** individuals

WHY IS THE TEMPORAL MAP PROBLEM DIFFICULT TO SOLVE?
PSCR Better Meter Stick Contest



01

Judging Maps

Synthetic map data requires quality results across the ***entire map***.

02

Map Diversity

Dense urban, sparse rural, and other variations require ***flexible algorithms***.

03

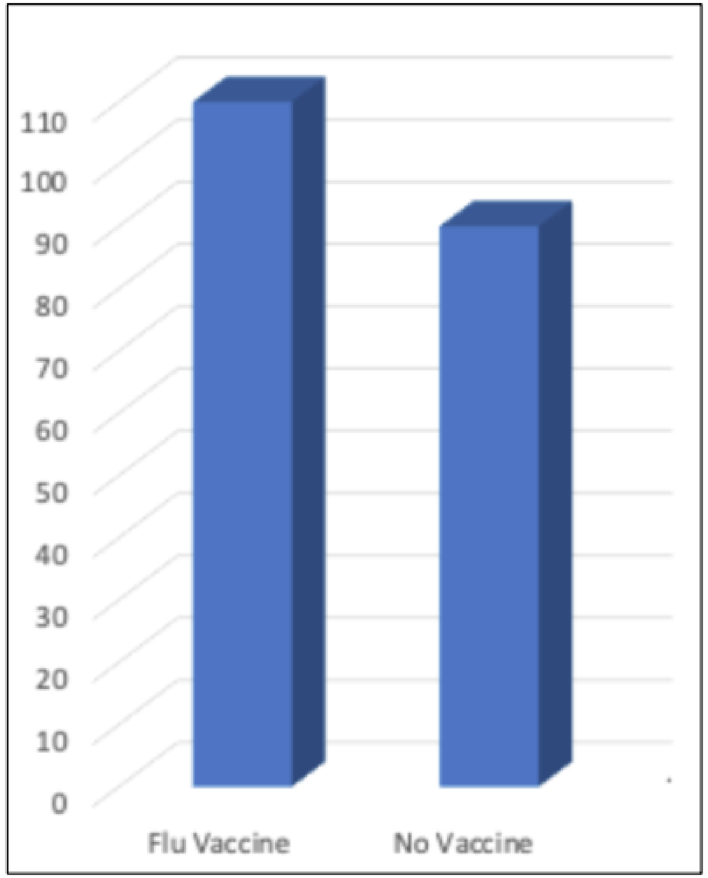
Adding Sequences

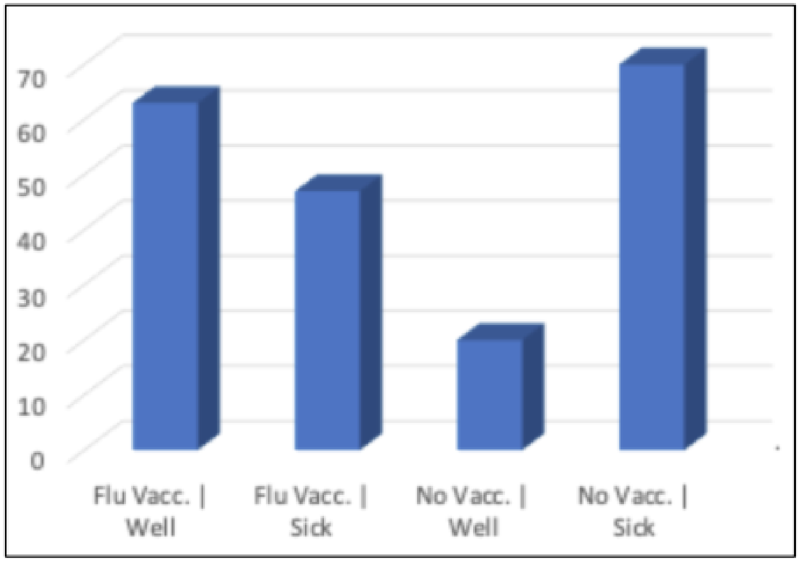
Time sequences increase the data space, and the difficulty ***exponentially.***

Technical Challenges

WHY IS THE TEMPORAL MAP PROBLEM DIFFICULT TO SOLVE?
PSCR Better Meter Stick Contest

Problem size and complexity increase with amount of information shared and number of map locations

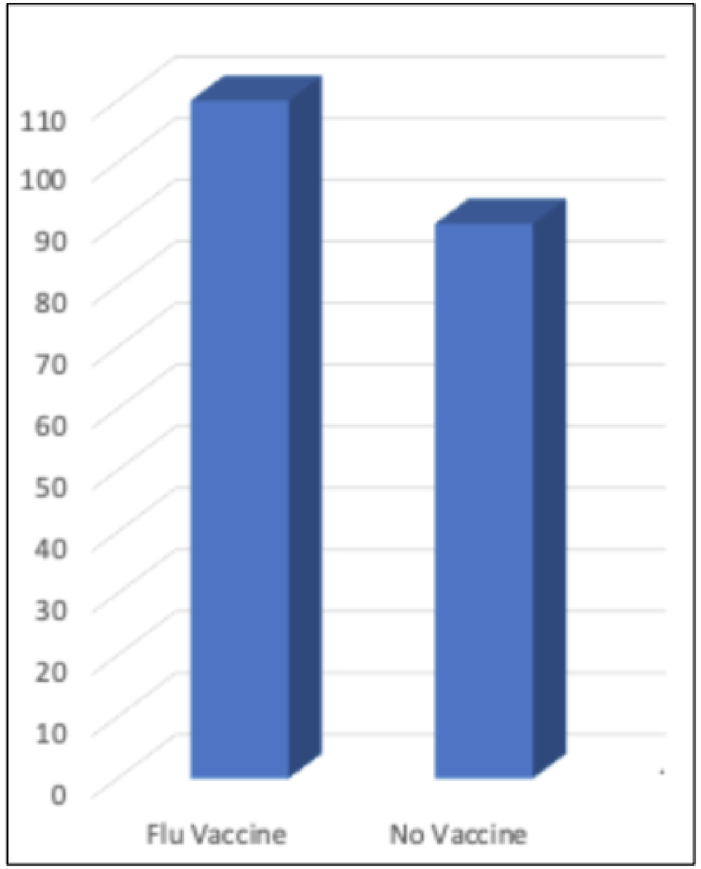


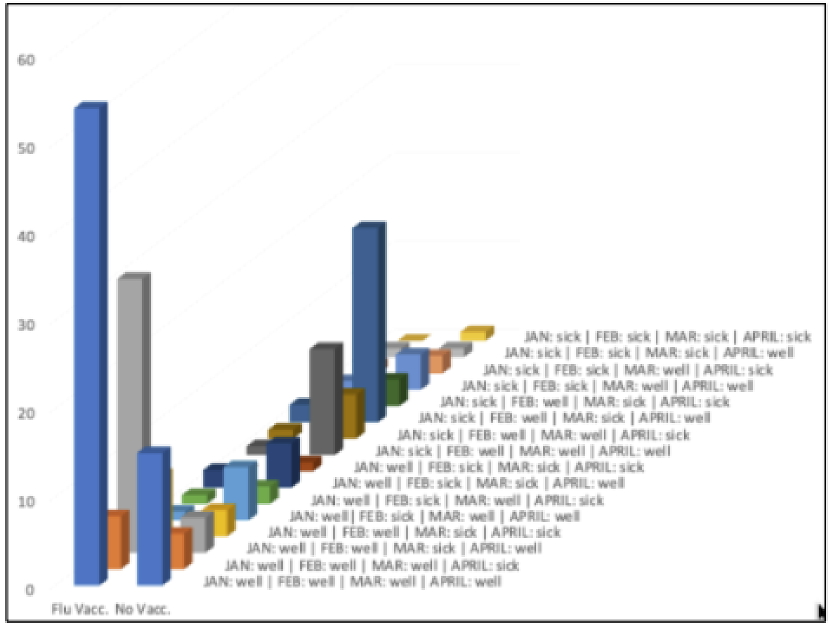


WHY IS THE TEMPORAL MAP PROBLEM DIFFICULT TO SOLVE?
PSCR Better Meter Stick Contest

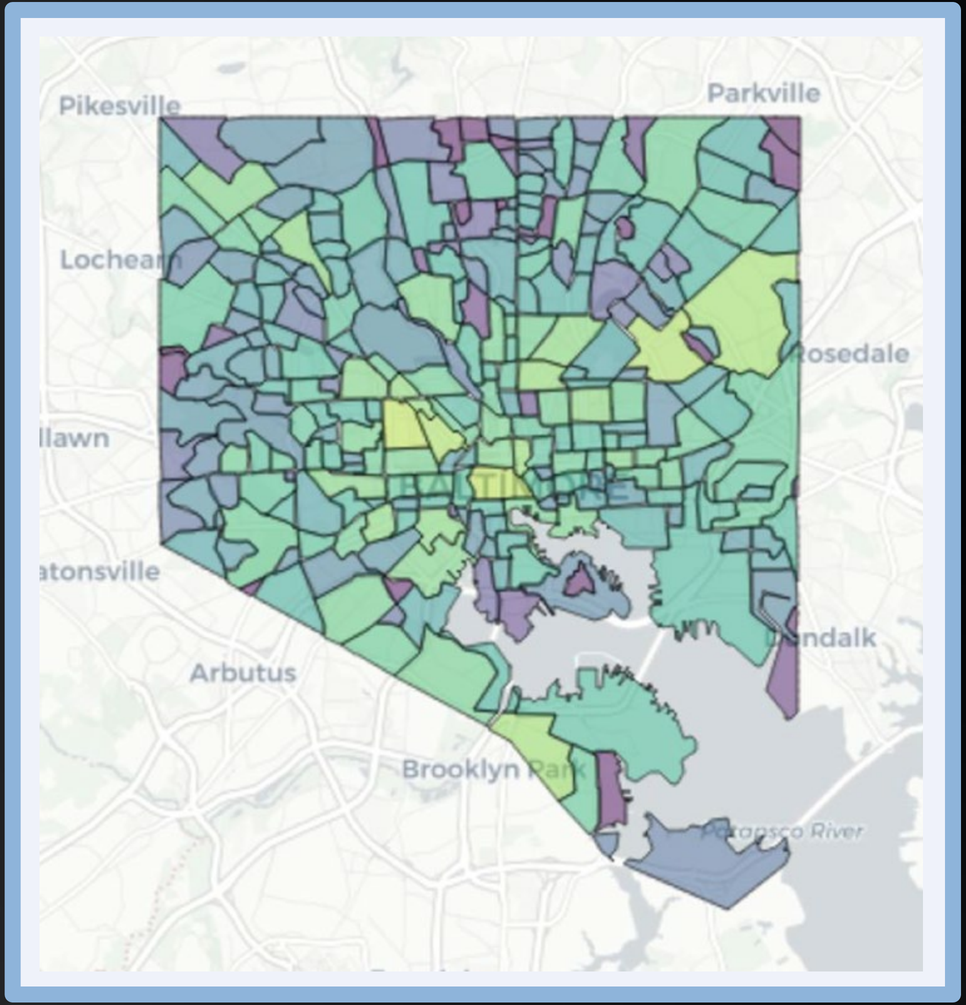
Problem size and complexity increase with amount of information shared and number of map locations

Problem size and complexity increase exponentially with number of time steps (per individual).





2020 PSCR DIFFERENTIAL PRIVACY CHALLENGE:
Temporal Map Data



SPRINT 1

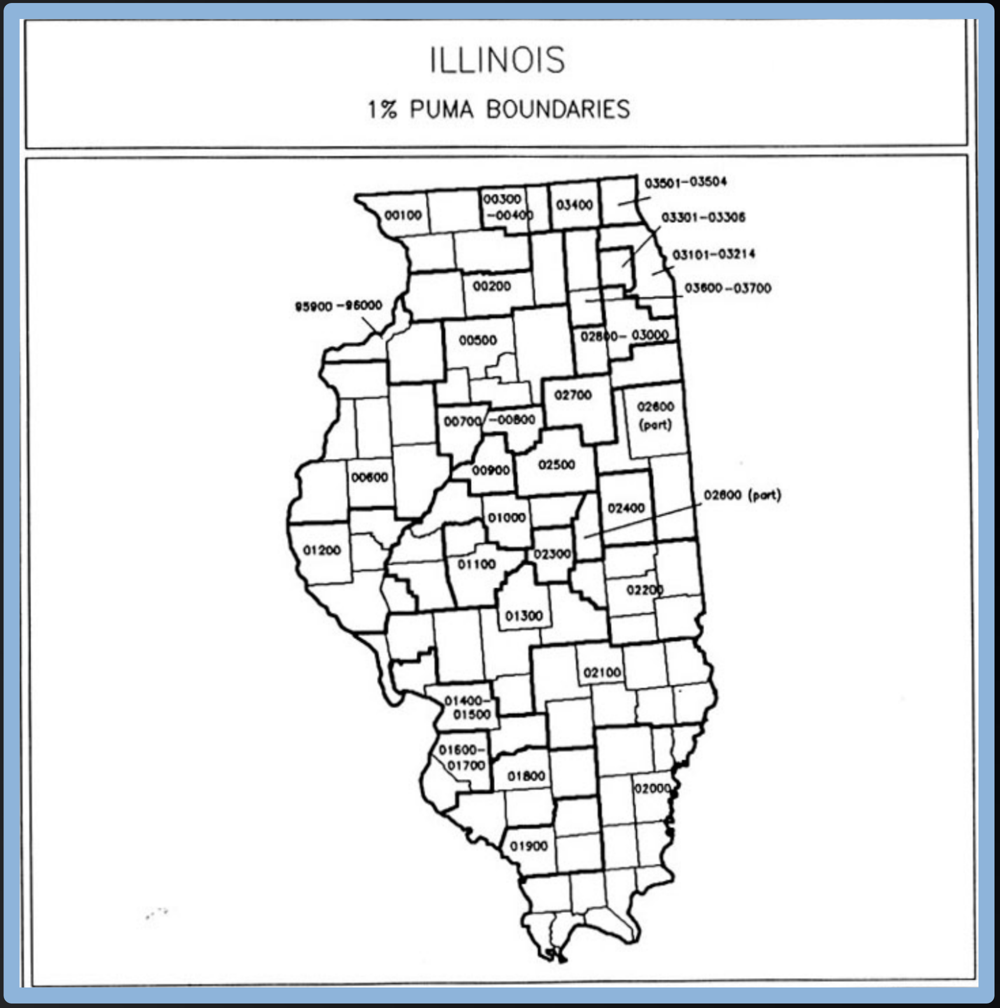
Competitors modeled:

- City of Baltimore 911 call data
- Baltimore neighborhoods
- Police incident and response

Competitors developed:

- Differentially aggregate methods
- Information dashboard shows trends by neighborhood and month

2020 PSCR DIFFERENTIAL PRIVACY CHALLENGE: Temporal Map Data



SPRINT 2

Competitors modeled:

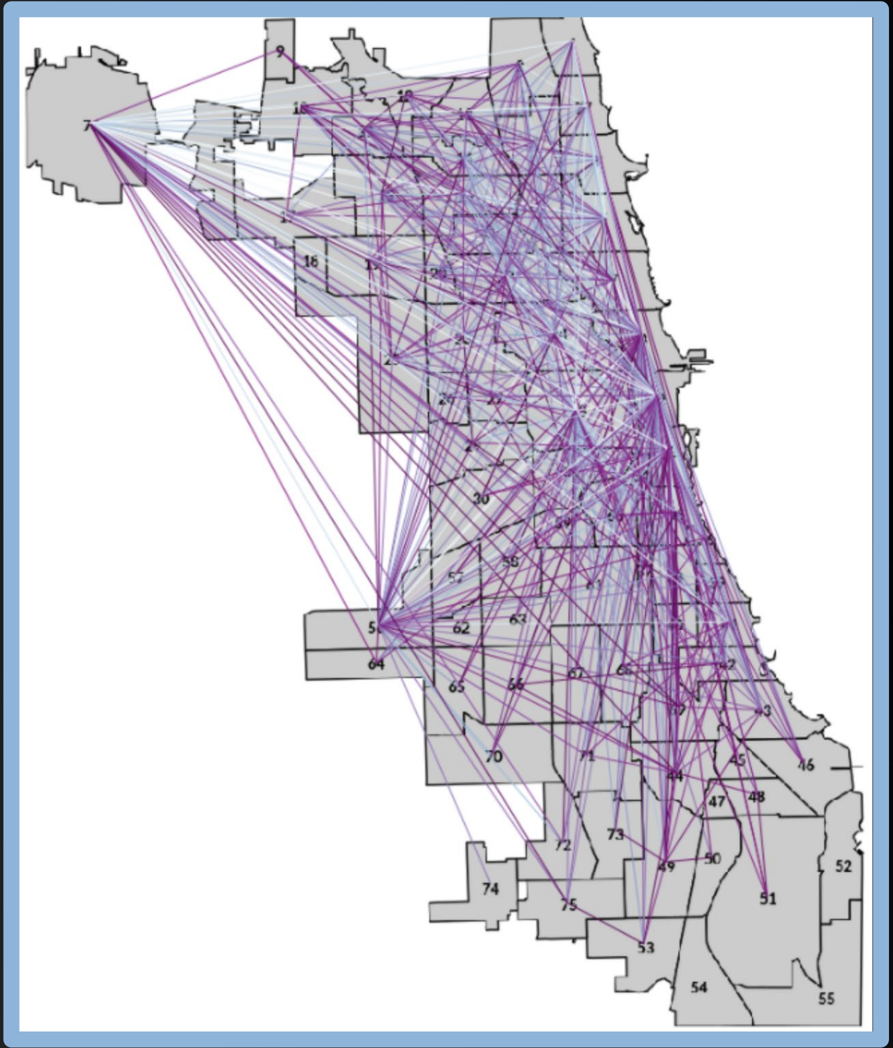
- American Community Survey Data
- PUMA[1] districts in IL, OH, NY, PA, GA, NC and SC
- Demographics, income, health insurance, work status

Competitors developed:

- Synthetic Records over a decade
- Privacy-preserving data can be used to analyze neighborhood shifts and population trends over time

[1] Public Use Microdata Areas (**PUMAs**) are non-overlapping, statistical geographic areas that partition each state or equivalent entity into geographic areas containing no fewer than 100,000 people each.

2020 PSCR DIFFERENTIAL PRIVACY CHALLENGE: Temporal Map Data



SPRINT 3

Competitors modeled:

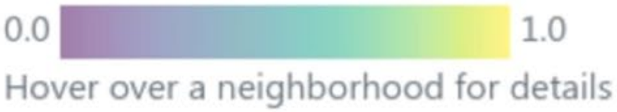
- City of Chicago taxi data
- Chicago Community Areas
- Traffic and trip patterns by time of day and day of week

Competitors developed:

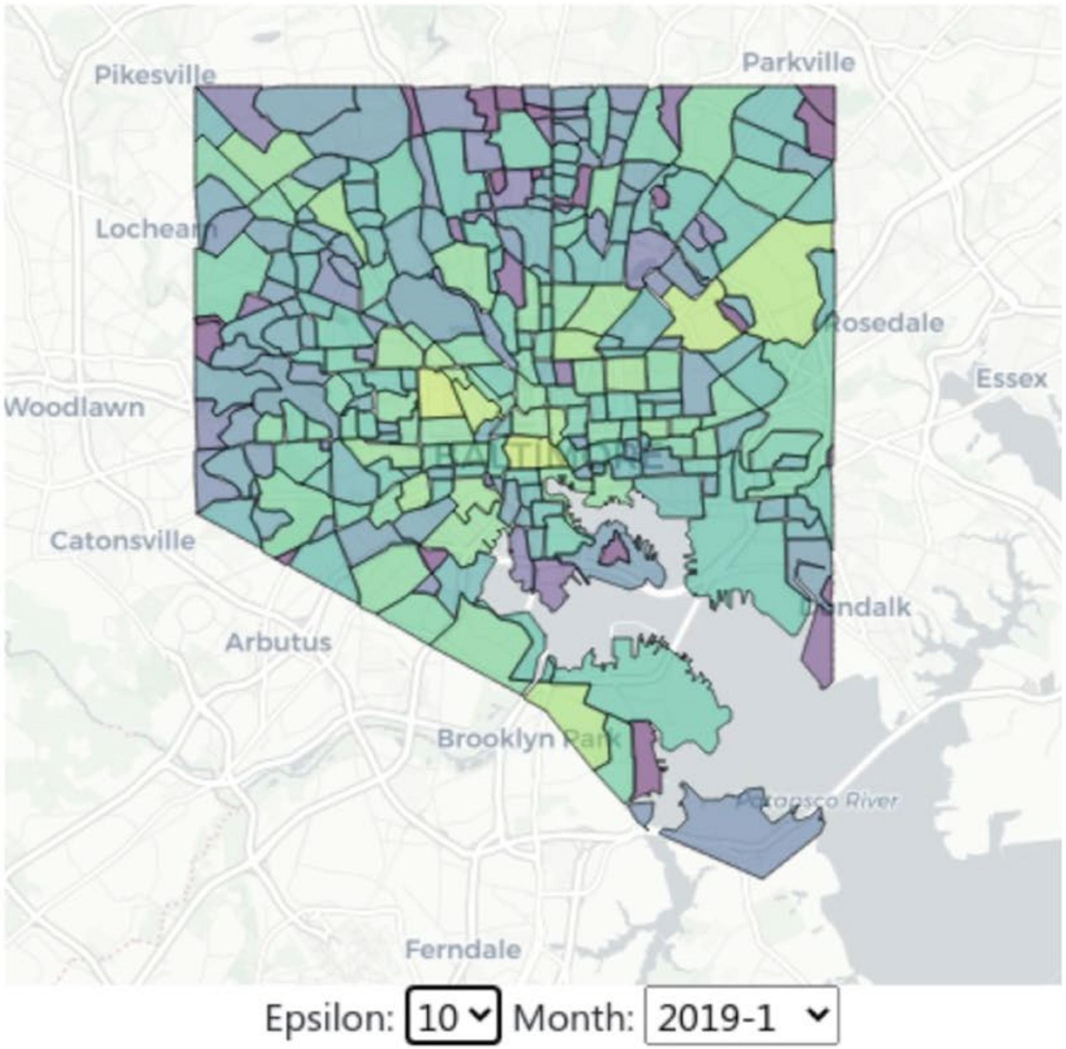
- Synthetic individual taxi drivers
- Synthetic trip data
- Privatized information can be used to analyze, traffic, activity and flow between community areas.

WHAT ABOUT QUALITY?
PSCR Better Meter Stick Contest

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.



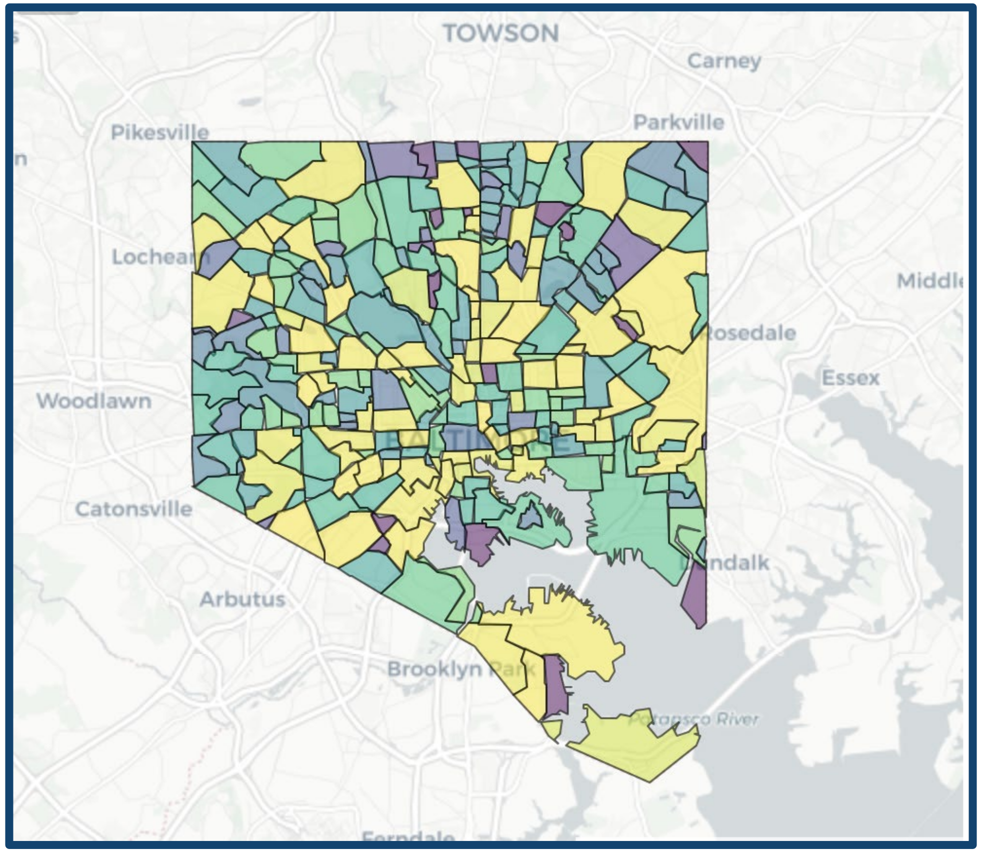
Hover over a neighborhood for details



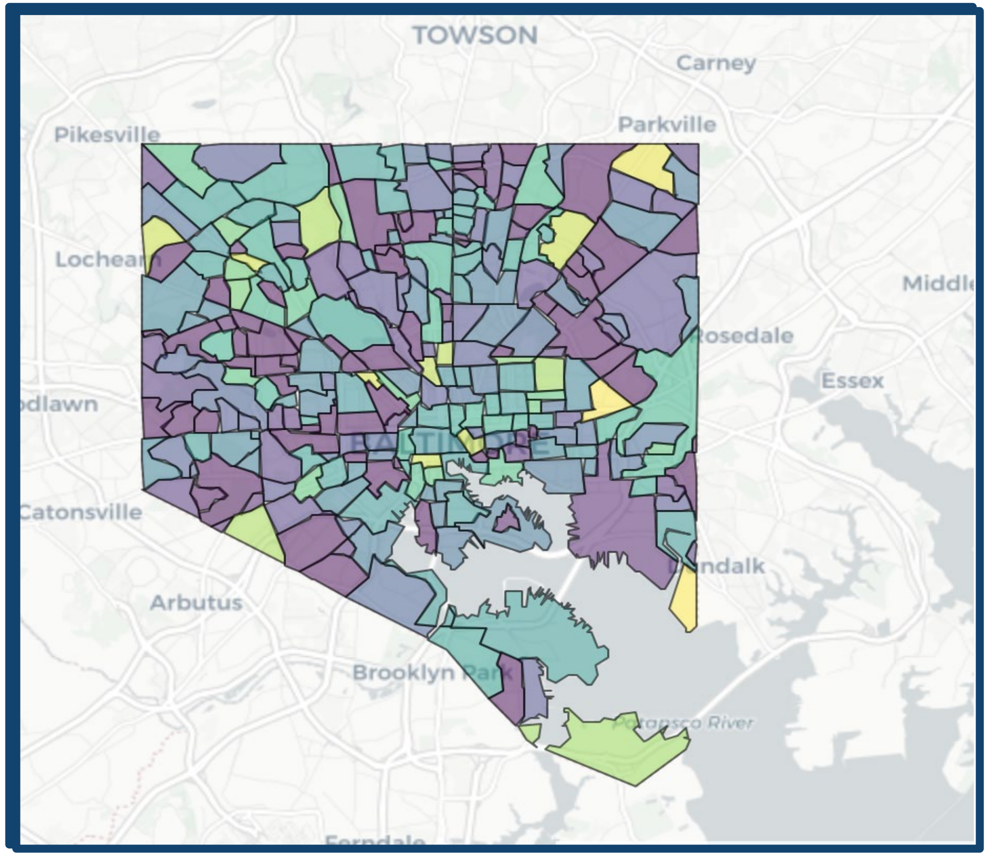
Epsilon: 10 Month: 2019-1

APPROACHES AND RESULTS
PSCR Better Meter Stick Contest

Baltimore 911: 1st Place Winner



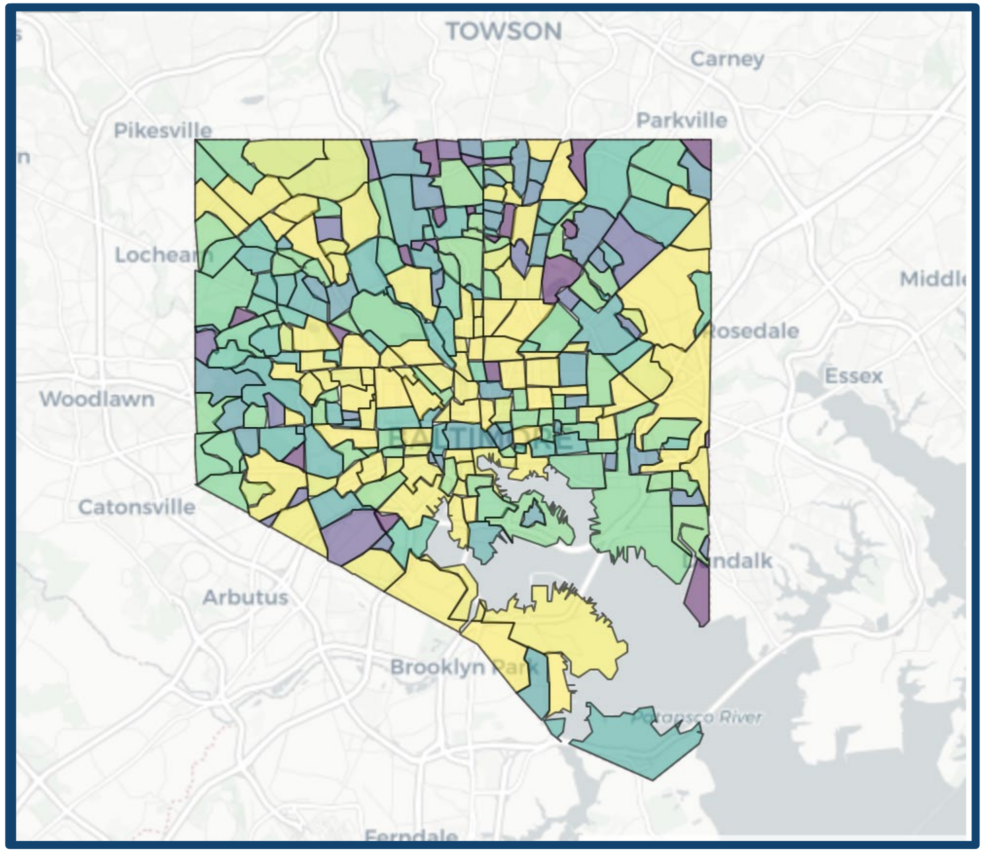
Baltimore 911: 5th Place Winner



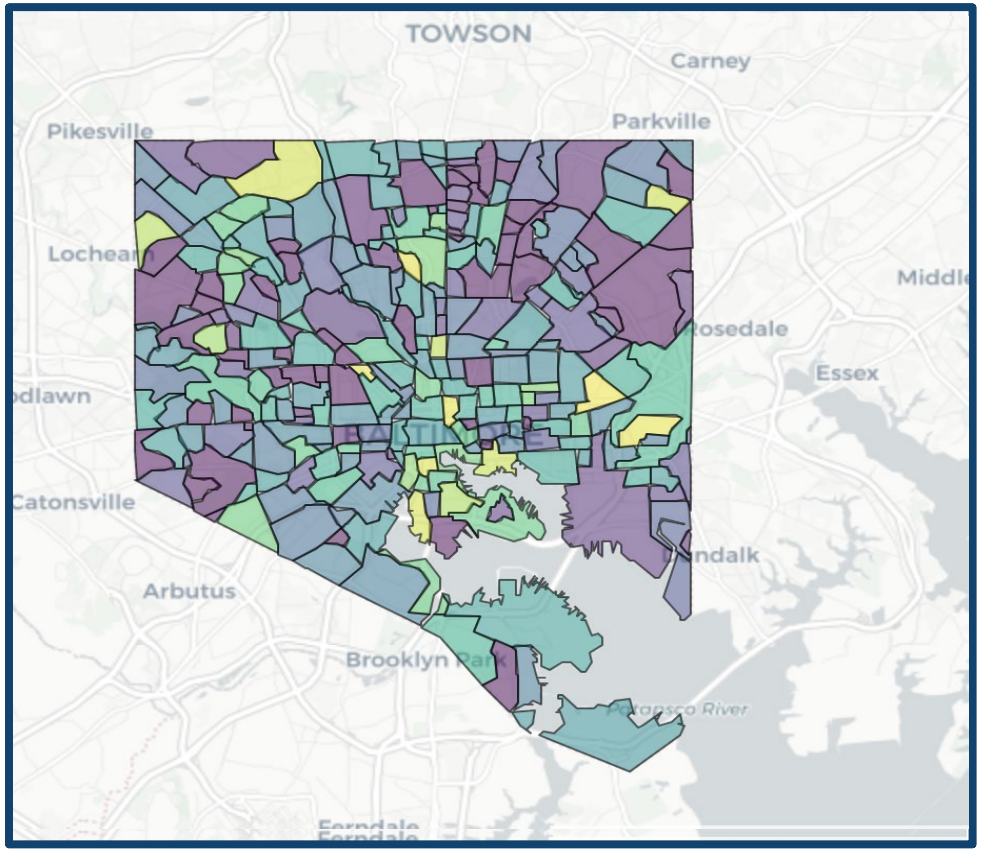
Data with poor utility 🟪 🟦 🟩 🟨 **Data with more utility**

APPROACHES AND RESULTS
PSCR Better Meter Stick Contest

Baltimore 911: 1st Place Winner



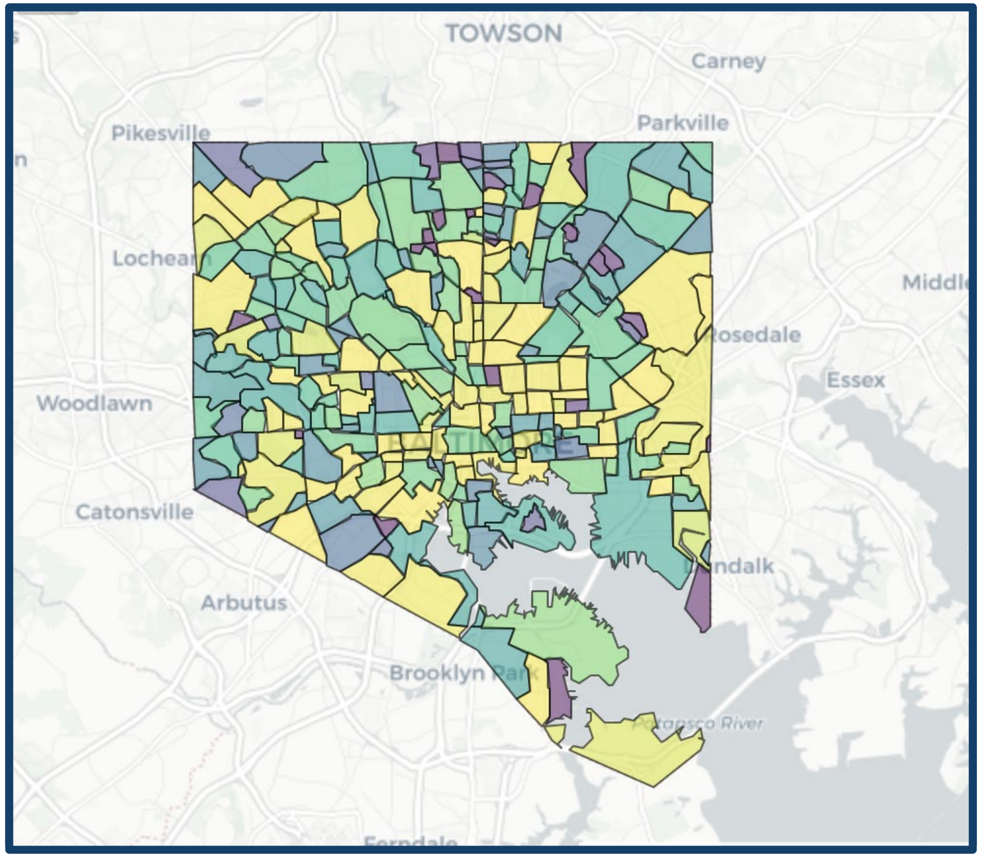
Baltimore 911: 5th Place Winner



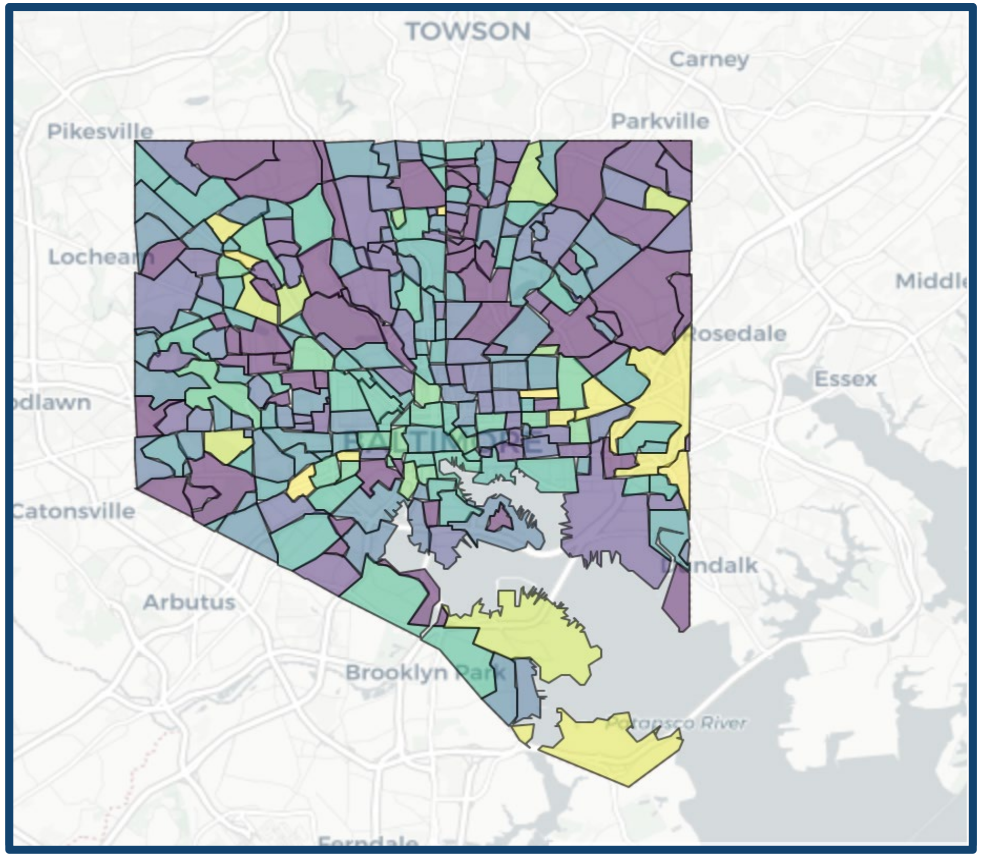
Data with poor utility ▨ **Data with more utility**

APPROACHES AND RESULTS
PSCR Better Meter Stick Contest

Baltimore 911: 1st Place Winner



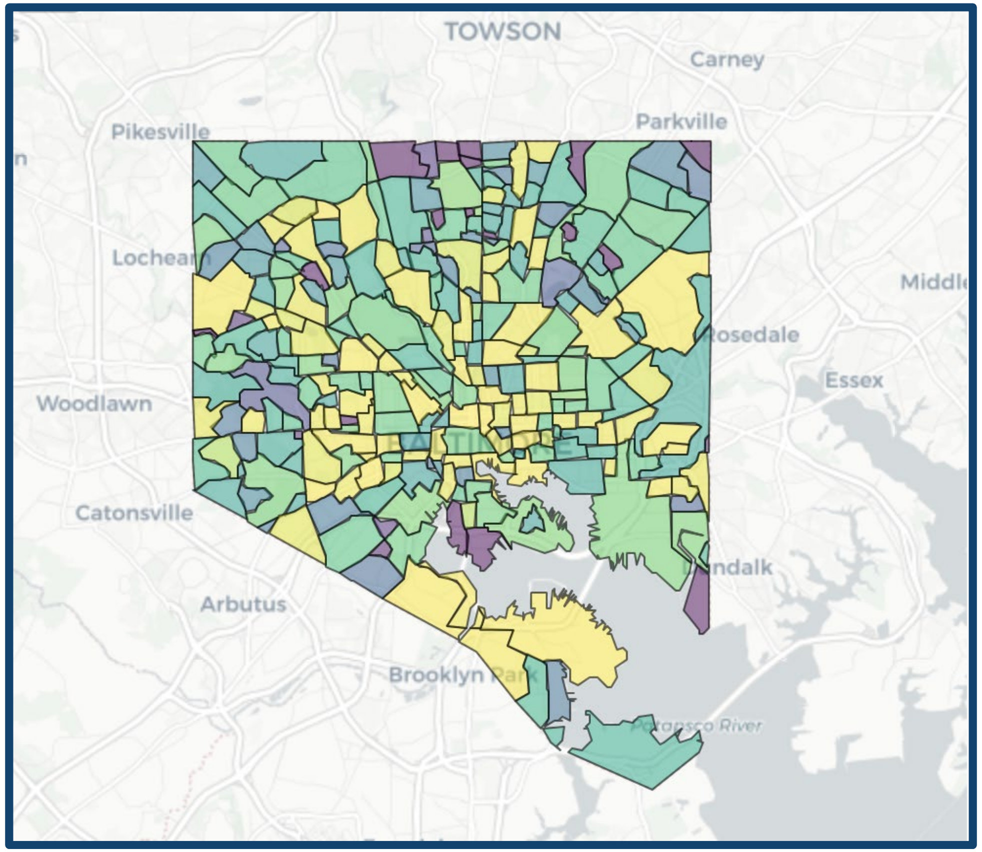
Baltimore 911: 5th Place Winner



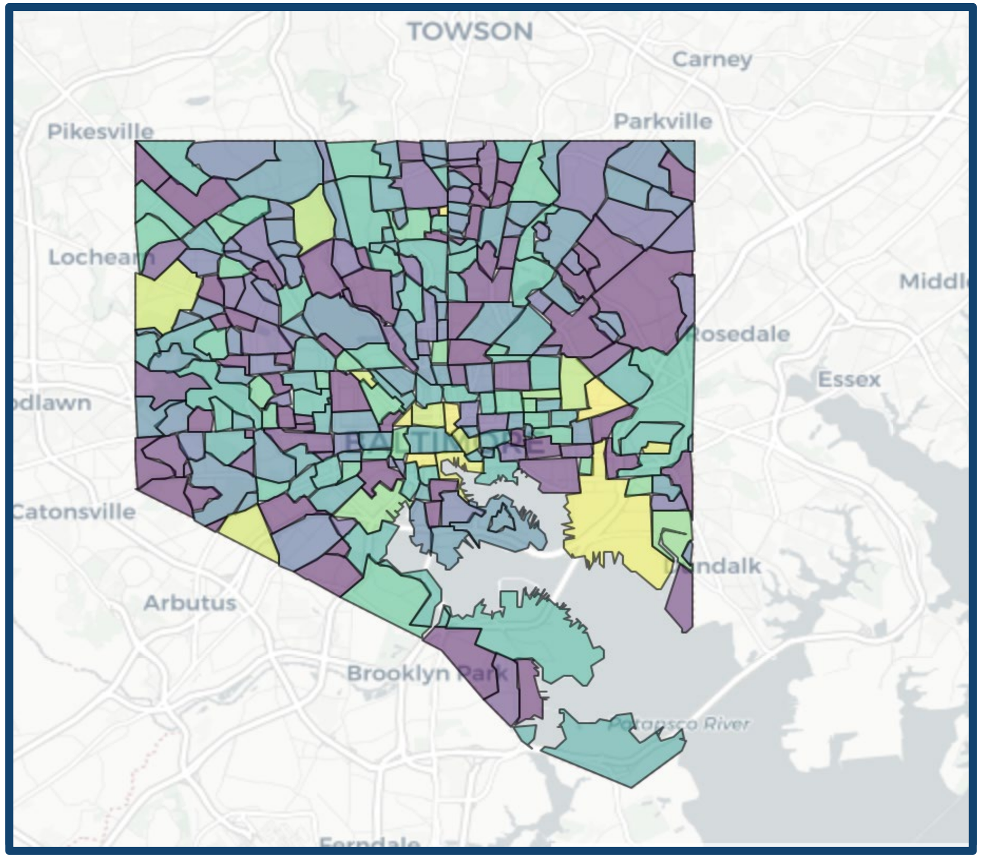
Data with poor utility ▬ **Data with more utility**

APPROACHES AND RESULTS
PSCR Better Meter Stick Contest

Baltimore 911: 1st Place Winner



Baltimore 911: 5th Place Winner

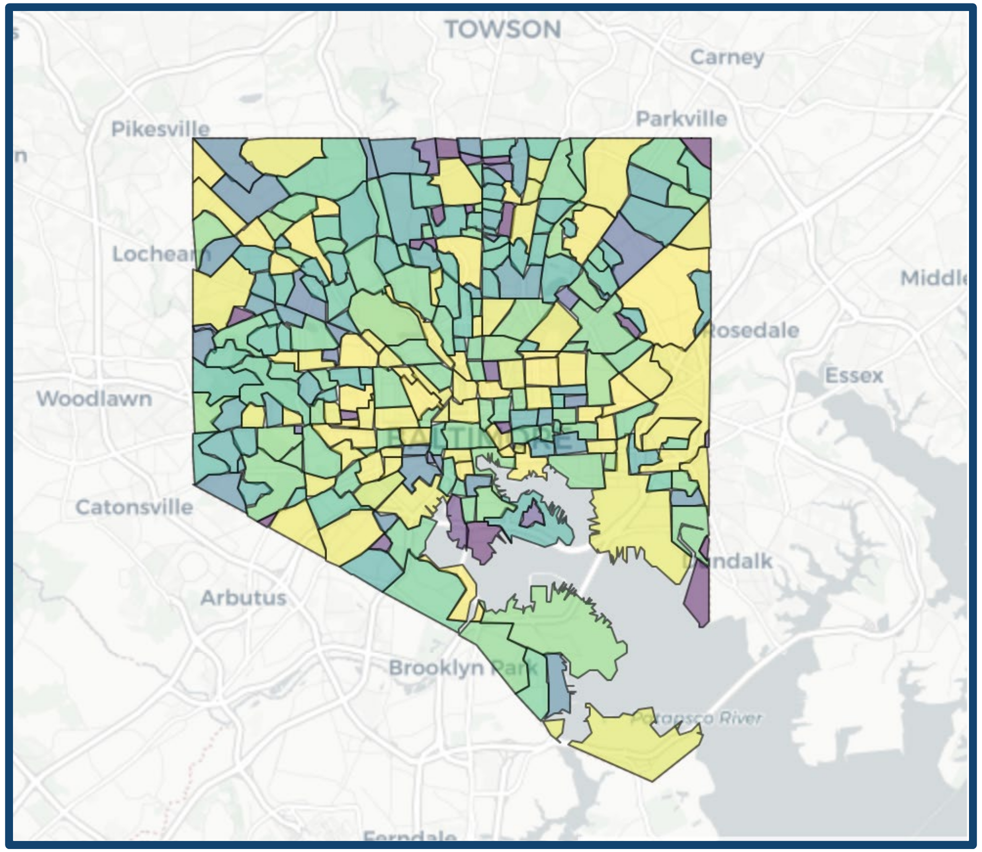


Data with less utility

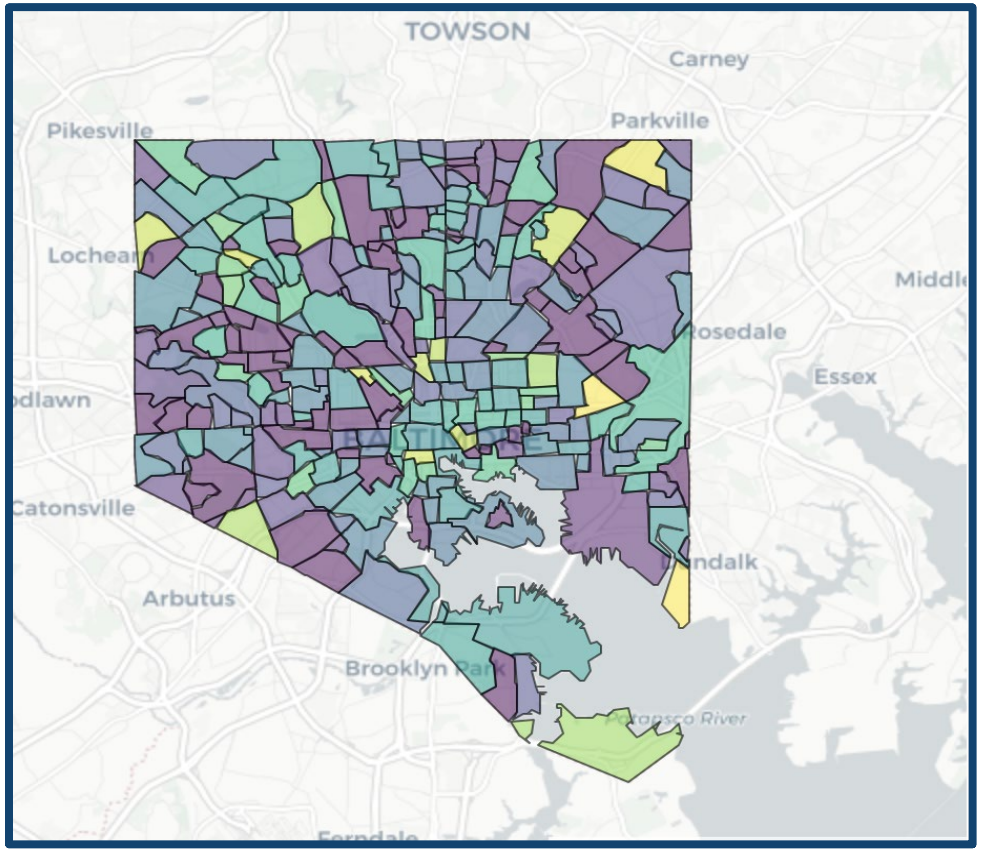
Data with more utility

APPROACHES AND RESULTS
PSCR Better Meter Stick Contest

Baltimore 911: 1st Place Winner



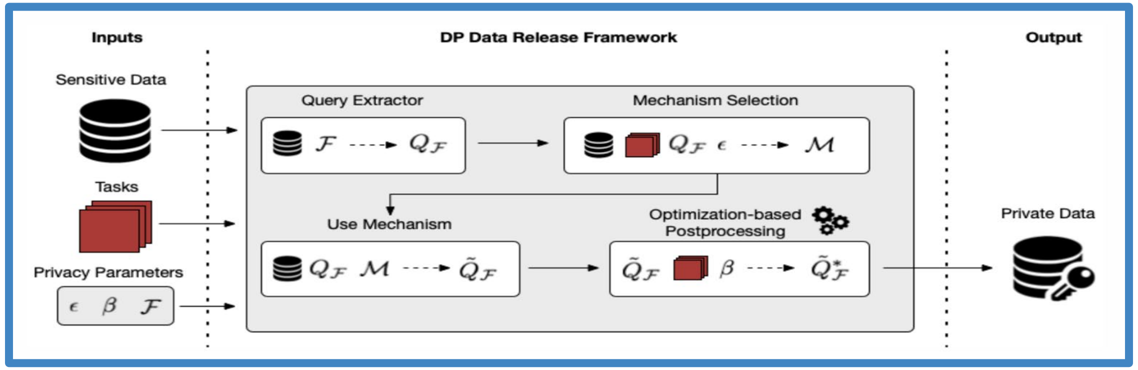
Baltimore 911: 5th Place Winner



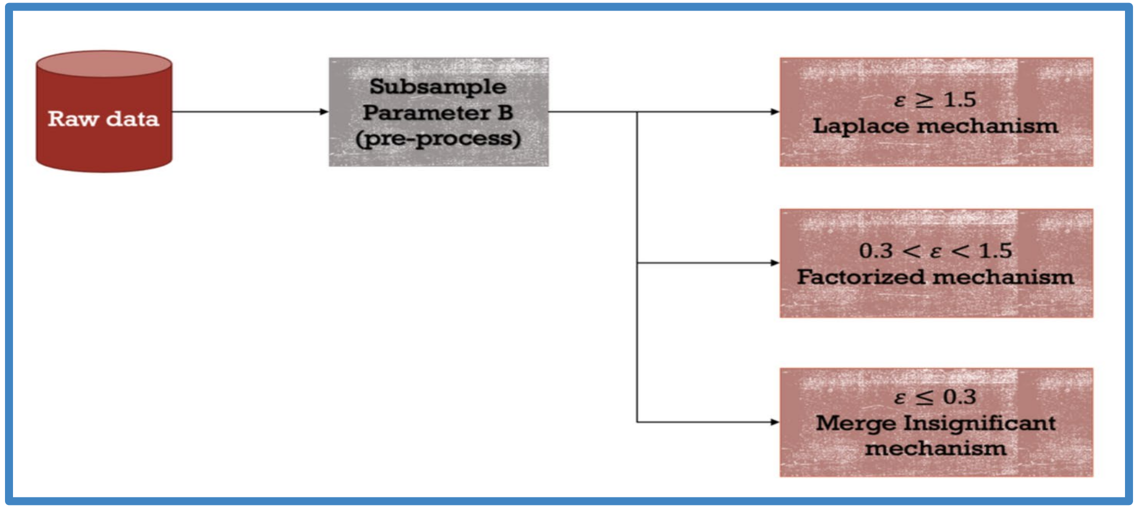
Data with less utility ▨ **Data with more utility**

APPROACHES AND RESULTS: How’d they do it?
PSCR Better Meter Stick Contest

APPROACHES AND RESULTS: How’d they do it?
PSCR Better Meter Stick Contest

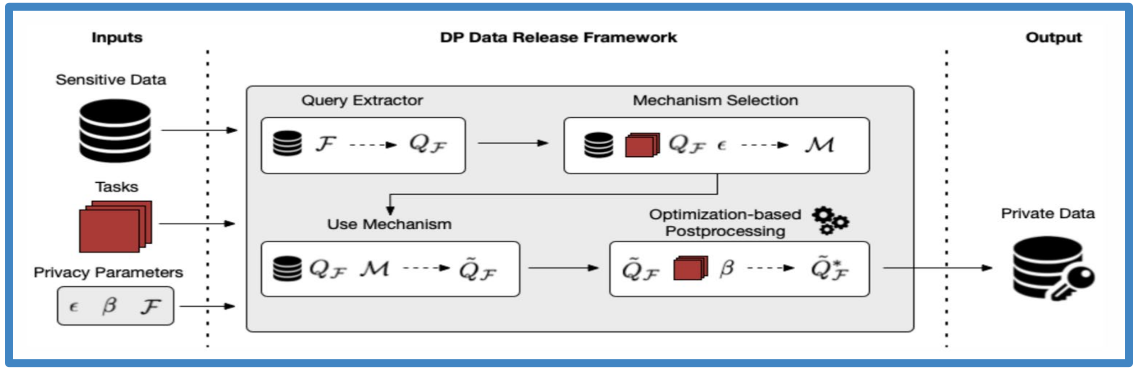


Inputs
DP Data Release Framework
Output
Sensitive Data
Tasks
Privacy Parameters
$\epsilon \quad \beta \quad \mathcal{F}$
Query Extractor
$\mathcal{F} \dashrightarrow Q_{\mathcal{F}}$
Mechanism Selection
$Q_{\mathcal{F}} \; \epsilon \dashrightarrow \mathcal{M}$
Use Mechanism
$Q_{\mathcal{F}} \; \mathcal{M} \dashrightarrow \tilde{Q}_{\mathcal{F}}$
Optimization-based Postprocessing
$\tilde{Q}_{\mathcal{F}} \; \beta \dashrightarrow \tilde{Q}^{*}_{\mathcal{F}}$
Private Data



Raw data
Subsample Parameter B (pre-process)
$\varepsilon \geq 1.5$ Laplace mechanism
$0.3 < \varepsilon < 1.5$ Factorized mechanism
$\varepsilon \leq 0.3$ Merge Insignificant mechanism

APPROACHES AND RESULTS: How’d they do it?
PSCR Better Meter Stick Contest



Inputs

DP Data Release Framework

Output

Sensitive Data

Tasks

Privacy Parameters

ϵ β $\mathcal{F}$

Query Extractor

$\mathcal{F} \dashrightarrow Q_{\mathcal{F}}$

Mechanism Selection

$Q_{\mathcal{F}}$ $\epsilon \dashrightarrow \mathcal{M}$

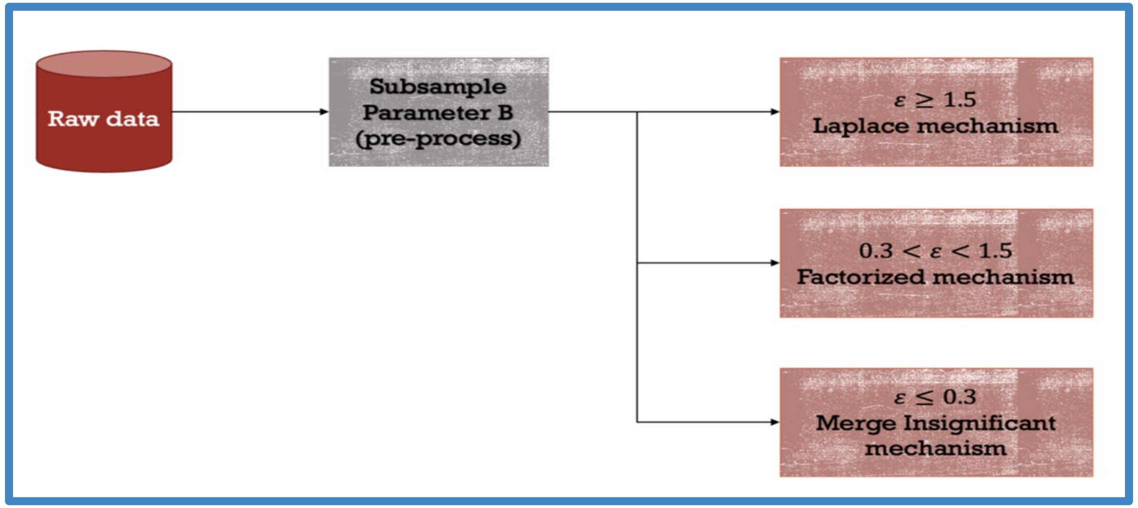
Use Mechanism

$Q_{\mathcal{F}}$ $\mathcal{M} \dashrightarrow \tilde{Q}_{\mathcal{F}}$

Optimization-based Postprocessing

$\tilde{Q}_{\mathcal{F}}$ $\beta \dashrightarrow \tilde{Q}^*_{\mathcal{F}}$

Private Data



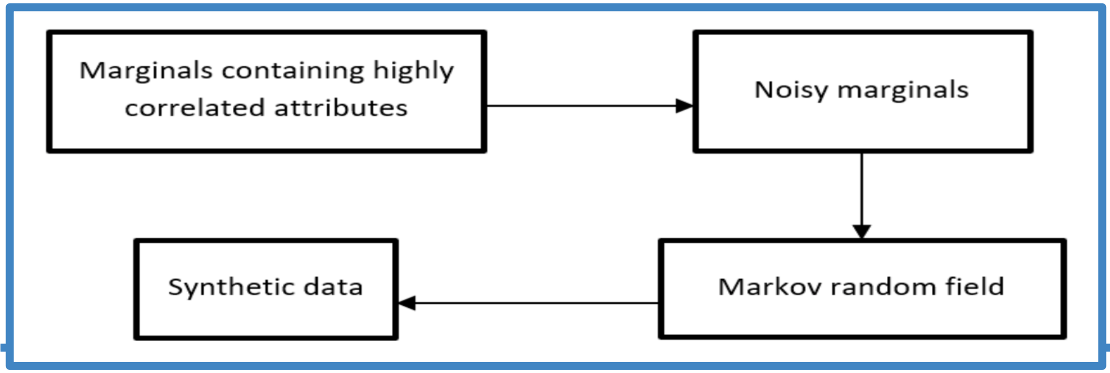
Raw data

Subsample Parameter B (pre-process)

$\varepsilon \geq 1.5$ Laplace mechanism

$0.3 < \varepsilon < 1.5$ Factorized mechanism

$\varepsilon \leq 0.3$ Merge Insignificant mechanism



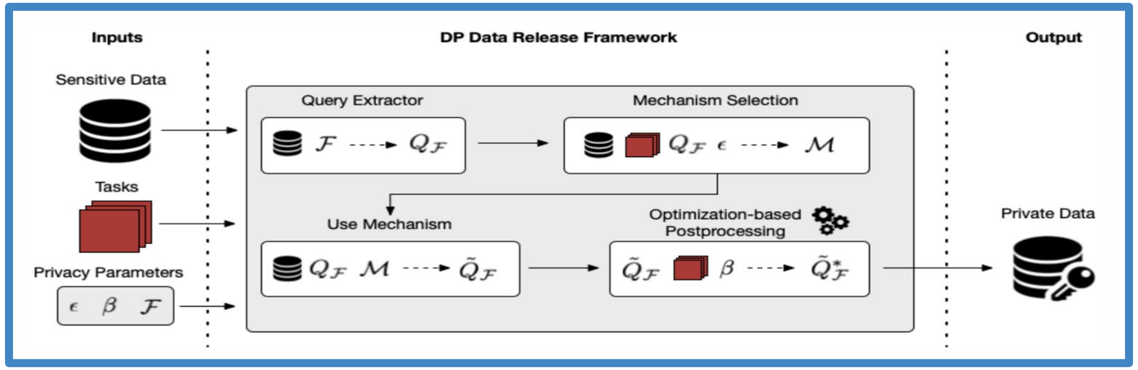
Marginals containing highly correlated attributes

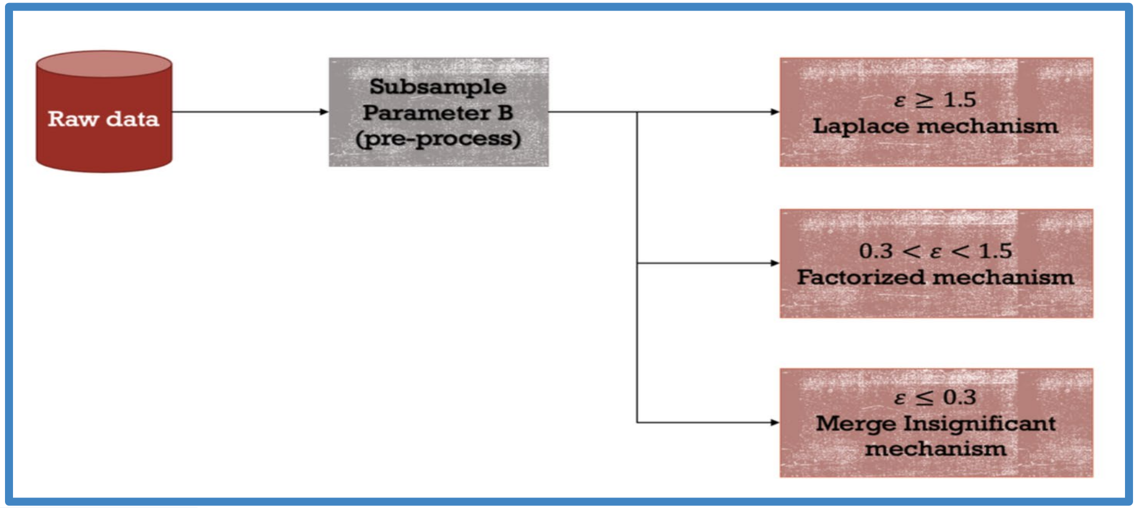
Noisy marginals

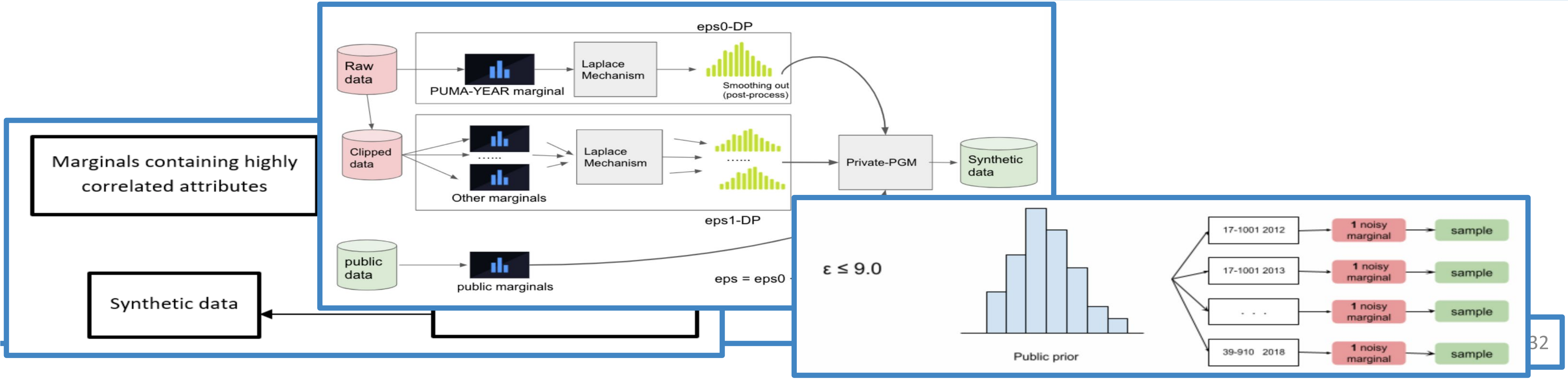
Markov random field

Synthetic data

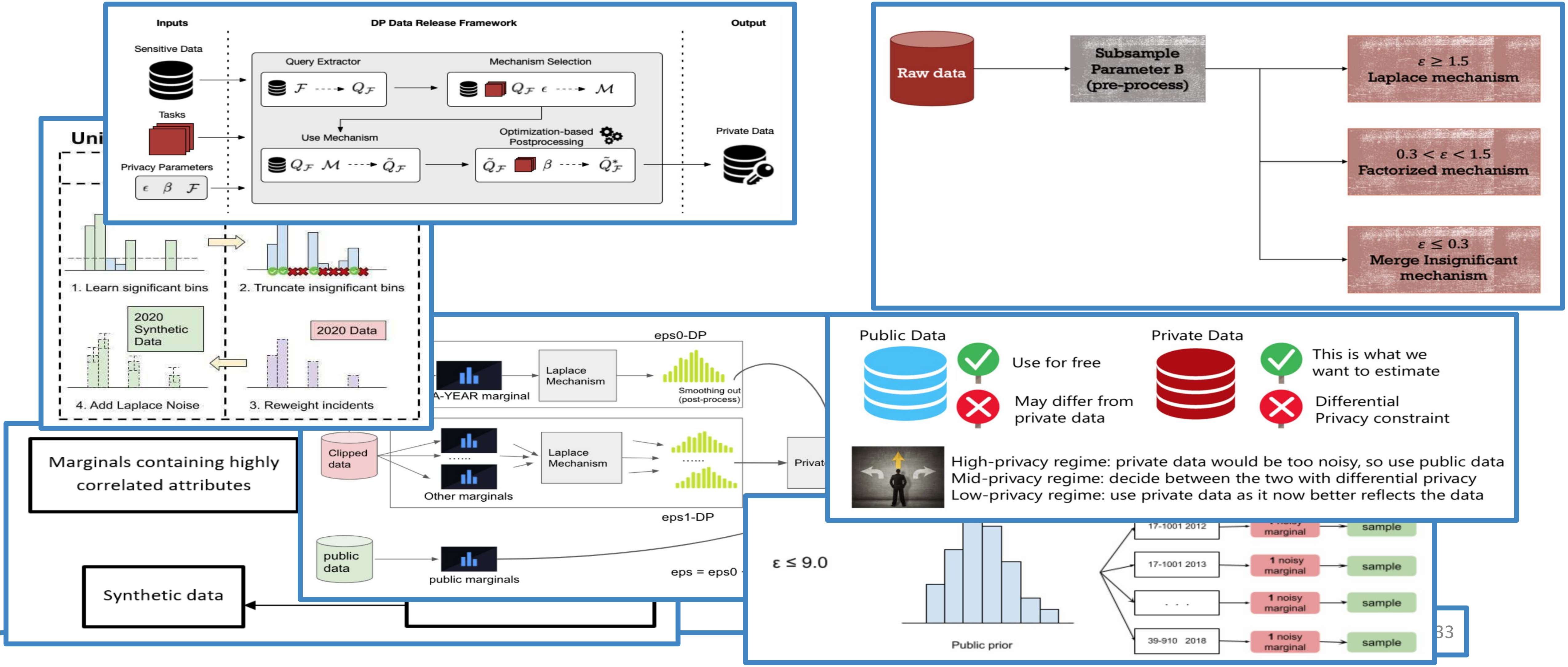
APPROACHES AND RESULTS: How’d they do it?
PSCR Better Meter Stick Contest


Inputs
DP Data Release Framework
Output
Sensitive Data
Tasks
Privacy Parameters
ϵ β ℱ
Query Extractor
ℱ ----> $Q_{\mathcal{F}}$
Mechanism Selection
$Q_{\mathcal{F}}$ ϵ ----> ℳ
Use Mechanism
$Q_{\mathcal{F}}$ ℳ ----> $\tilde{Q}_{\mathcal{F}}$
Optimization-based Postprocessing
$\tilde{Q}_{\mathcal{F}}$ β ----> $\tilde{Q}^*_{\mathcal{F}}$
Private Data


Raw data
Subsample Parameter B (pre-process)
ε ≥ 1.5 Laplace mechanism
0.3 < ε < 1.5 Factorized mechanism
ε ≤ 0.3 Merge Insignificant mechanism



APPROACHES AND RESULTS: How'd they do it?
PSCR Better Meter Stick Contest



Inputs
DP Data Release Framework
Output
Sensitive Data
Tasks
Privacy Parameters
ϵ β $\mathcal{F}$
Query Extractor
$\mathcal{F} \dashrightarrow Q_{\mathcal{F}}$
Mechanism Selection
$Q_{\mathcal{F}}$ $\epsilon \dashrightarrow \mathcal{M}$
Use Mechanism
$Q_{\mathcal{F}}$ $\mathcal{M} \dashrightarrow \tilde{Q}_{\mathcal{F}}$
Optimization-based Postprocessing
$\tilde{Q}_{\mathcal{F}}$ $\beta \dashrightarrow \tilde{Q}^{*}_{\mathcal{F}}$
Private Data
Raw data
Subsample Parameter B (pre-process)
$\varepsilon \geq 1.5$ Laplace mechanism
$0.3 < \varepsilon < 1.5$ Factorized mechanism
$\varepsilon \leq 0.3$ Merge Insignificant mechanism
Uni
1. Learn significant bins
2. Truncate insignificant bins
2020 Synthetic Data
2020 Data
4. Add Laplace Noise
3. Reweight incidents
eps0-DP
A-YEAR marginal
Laplace Mechanism
Smoothing out (post-process)
Clipped data
Other marginals
eps1-DP
Privat
public data
public marginals
eps = eps0
Marginals containing highly correlated attributes
Synthetic data
ε ≤ 9.0
Public prior
17-1001 2012
17-1001 2013
. . .
39-910 2018
1 noisy marginal
sample

Public Data

✓ Use for free

✗ May differ from private data

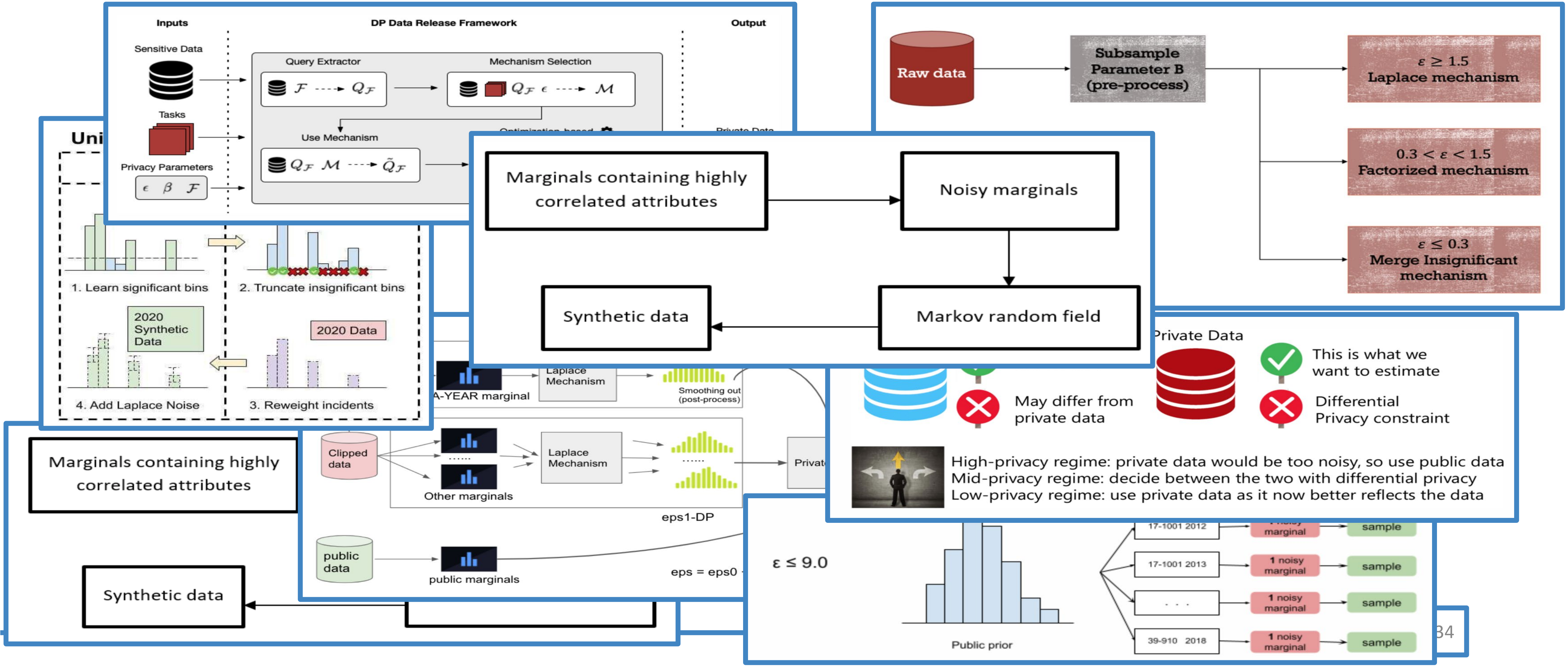
Private Data

✓ This is what we want to estimate

✗ Differential Privacy constraint

High-privacy regime: private data would be too noisy, so use public data
Mid-privacy regime: decide between the two with differential privacy
Low-privacy regime: use private data as it now better reflects the data

APPROACHES AND RESULTS: How’d they do it?
PSCR Better Meter Stick Contest



Inputs

DP Data Release Framework

Output

Sensitive Data

Tasks

Privacy Parameters

ϵ β $\mathcal{F}$

Query Extractor

$\mathcal{F} \dashrightarrow Q_{\mathcal{F}}$

Mechanism Selection

$Q_{\mathcal{F}}$ $\epsilon \dashrightarrow \mathcal{M}$

Use Mechanism

$Q_{\mathcal{F}}$ $\mathcal{M} \dashrightarrow \tilde{Q}_{\mathcal{F}}$

Optimization-based

Private Data

Raw data

Subsample Parameter B (pre-process)

$\varepsilon \geq 1.5$ Laplace mechanism

$0.3 < \varepsilon < 1.5$ Factorized mechanism

$\varepsilon \leq 0.3$ Merge Insignificant mechanism

Uni

1. Learn significant bins

2. Truncate insignificant bins

2020 Synthetic Data

2020 Data

4. Add Laplace Noise

3. Reweight incidents

Marginals containing highly correlated attributes

Noisy marginals

Markov random field

Synthetic data

A-YEAR marginal

Laplace Mechanism

Smoothing out (post-process)

Clipped data

Other marginals

Laplace Mechanism

Privat

eps1-DP

public data

public marginals

eps = eps0

Private Data

May differ from private data

This is what we want to estimate

Differential Privacy constraint

High-privacy regime: private data would be too noisy, so use public data
Mid-privacy regime: decide between the two with differential privacy
Low-privacy regime: use private data as it now better reflects the data

Marginals containing highly correlated attributes

Synthetic data

ε ≤ 9.0

Public prior

17-1001 2012

17-1001 2013

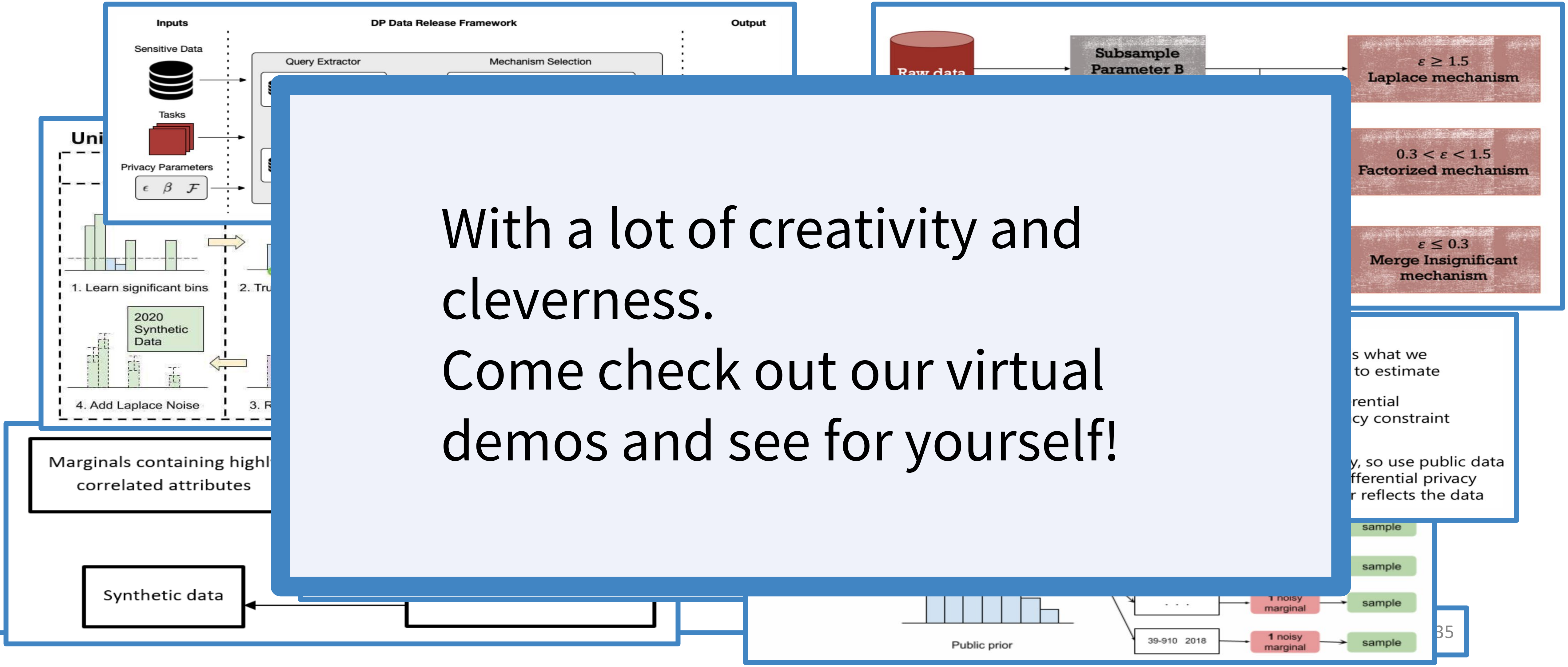
. . .

39-910 2018

1 noisy marginal

sample

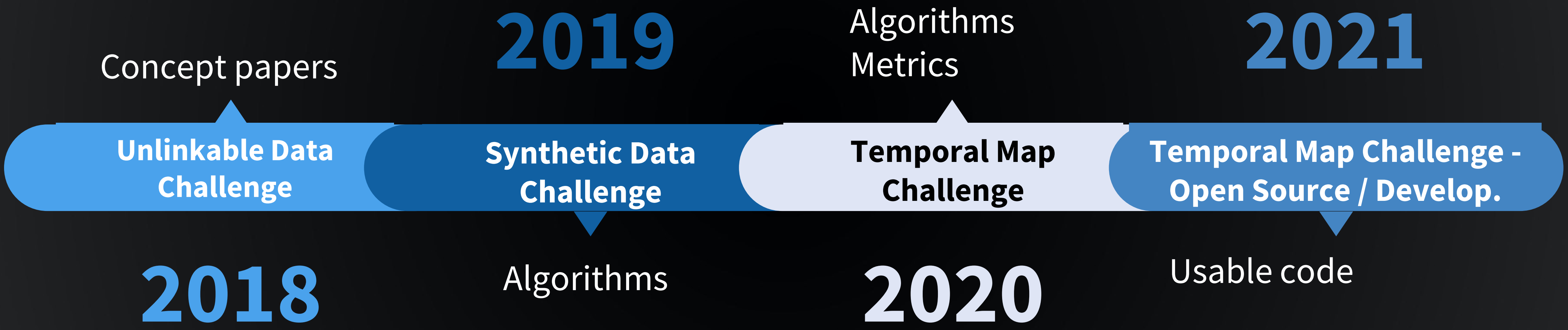
APPROACHES AND RESULTS: How'd they do it?
PSCR Better Meter Stick Contest



With a lot of creativity and cleverness.

Come check out our virtual demos and see for yourself!

Differential Prize Challenge Series



2018

Unlinkable Data Challenge

Concept papers

2019

Synthetic Data Challenge

Algorithms

2020

Temporal Map Challenge

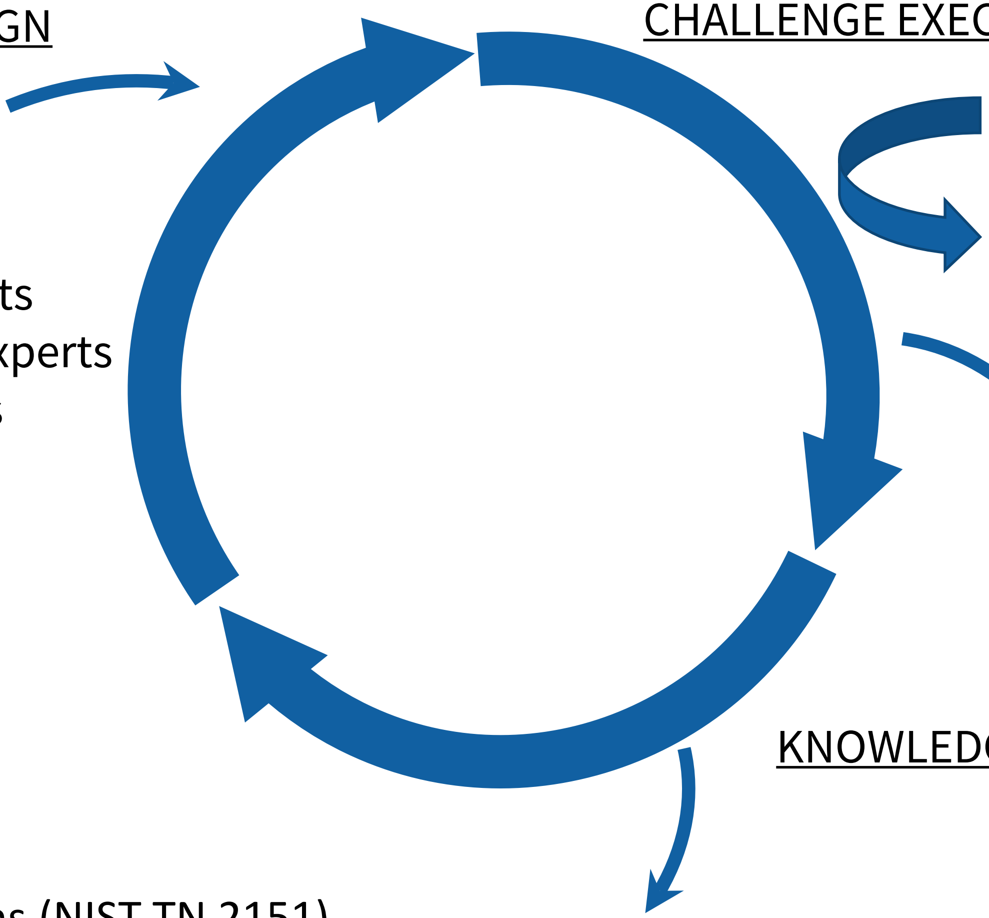
Algorithms
Metrics

2021

Temporal Map Challenge - Open Source / Develop.

Usable code

DIFFERENTIAL PRIVACY CHALLENGE LOGIC MODEL



CHALLENGE DESIGN

- Privacy literature
- Feedback from:
  - Past contestants
  - Public safety experts
  - Privacy experts

CHALLENGE EXECUTION

Algorithms

Improved algorithms

- Datasets
- Performance metrics
- Participation data

KNOWLEDGE INCUBATION & TRANSMISSION

- NIST publications (NIST TN 2151)
- Open-source software (NIST Privacy Engineering Collaborative Space)
- Symposium (Fall 2021)

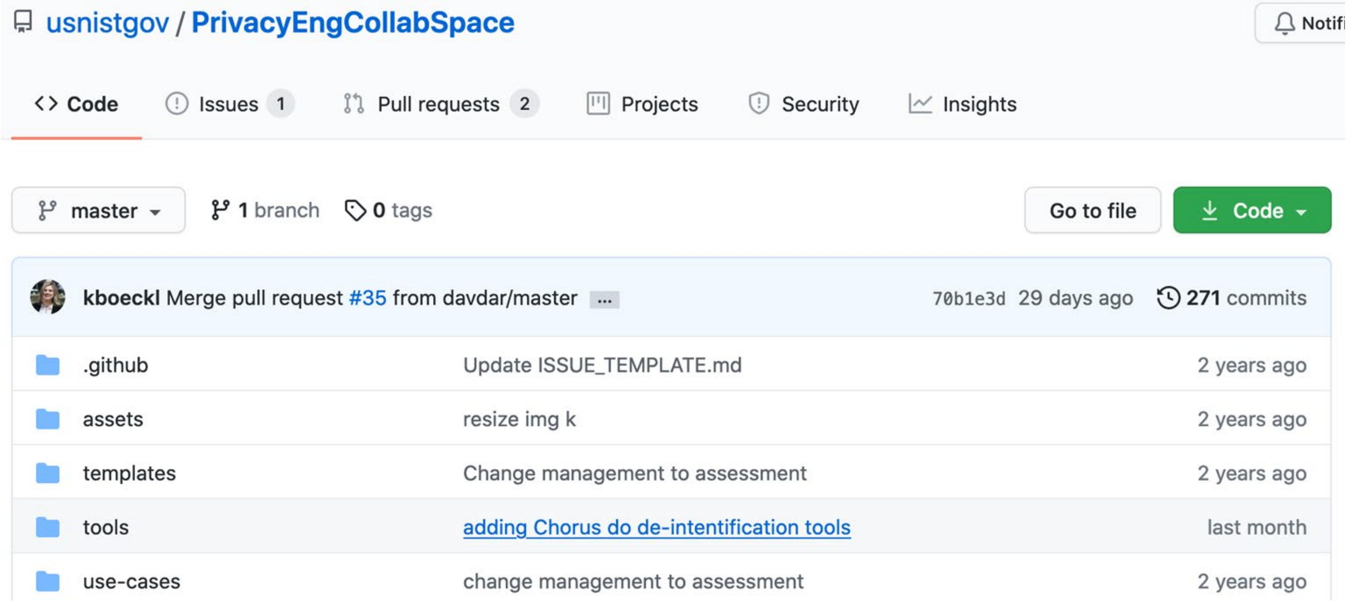
FUTURE DIRECTIONS

- PSCR is organizing differential privacy algorithms and metrics into an open-source repository.
- PSCR is incentivizing 2020 DP Challenge solvers to develop robust, open-source code.
- PSCR will be hosting a symposium (Nov 2021) focusing on how to make DP algorithms more extensible and robust.

WHERE DO I FIND RESOURCES?

Visit the NIST Privacy Collaboration Space on Github:

https://github.com/usnistgov/PrivacyEngCollabSpace



kboeckl Merge pull request #35 from davdar/master ... 70b1e3d 29 days ago 271 commits

.github	Update ISSUE_TEMPLATE.md	2 years ago
assets	resize img k	2 years ago
templates	Change management to assessment	2 years ago
tools	adding Chorus do de-intentification tools	last month
use-cases	change management to assessment	2 years ago

See the Technical Demo section of the Stakeholders Meeting to see demos from winners of the Differential Privacy Temporal Map Challenge.

TECHNICAL DEMOS

Please see the Technical Demo section of the 2021 PSCR Stakeholder Meeting Digital Experience for examples of differential privacy implementations.

THANK YOU!

Gary Howarth
Prize Challenge Manager
NIST, PSCR
Gary.Howarth@nist.gov

Christine Task
Computer Scientist
Knexus Research
christine.task@knexusresearch.com