Using and Contributing to the OpenDP Library

Preparation for audience exercises:

1. Open Jupyter notebook: shorturl.at/cimp8
    - Can use in your browser
    - Or File->Download to use local Python installation
2. Fill out audience poll: strawpoll.com/continent

Using and Contributing to the OpenDP Library

Michael Shoemate
Harvard University
shoematem@g.harvard.edu

Salil Vadhan
Harvard University
salil_vadhan@harvard.edu

4th AAAI Workshop on Privacy-Preserving AI
February 13, 2023

Supported by:

Alfred P. Sloan Foundation
Microsoft
mozilla

Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of our funders.

Outline

- Motivation and history (Salil)
- Overview of software & features (Salil)
- The OpenDP Library programming framework (Salil)

- The programming framework in code (Mike)
- Audience exercises (Mike)
- How to contribute (Mike)

- Interactive measurements (Salil)
- OpenDP roadmap (Salil)

OpenDP

A community effort to build a trustworthy and open-source suite of differential privacy tools that can be easily adopted by custodians of sensitive data to make it available for research and exploration in the public interest.

Differential Privacy Deployed

U.S. Census Bureau

- "OnTheMap" commuter data [Machanavajjhala et al. \`06]
- All public-use products from 2020 U.S. Decennial Census [Abowd \`18]

Big Tech

- RAPPOR for Chrome Statistics [Erlingsson et al. \`14]
- iOS10 and Safari [Apple \`16]
- Windows 10 [Ding et al. \`17]
- …

Enterprise Software & Consulting

- Apheris, Canopy, DataFleets/LiveRamp, Decentriq, Hazy, Immuta, LeapYear, Oasis Labs, Oblivious AI, optable, Privitar, SAP, Sarus Technologies, sherpa.ai, TripleBlind, Tumult Labs, …

Open-source DP Software

DP Machine Learning (esp. deep learning)

TensorFlow Privacy [McMahan et al. `18], PySyft [Ryffel et al. `18], Opacus [Testuggine & Mironov `20], …

Academic Proof-of-Concepts

LightDP [Zhang & Kifer `17], Ektelo [Zhang et al. `18], Duet [Near et al. `19], Fuzzi [Zhang et al. `19], Chorus [Johson et al. `20], …

General-purpose repositories

Google DP Library [Wilson et al. `19], IBM Diffprivlib [Holohan et al. `19],
NIST Privacy Engineering Collaboration Space,
OpenMined, Tumult Analytics

The Need for OpenDP

- Trustworthiness
  - Implementing DP correctly is difficult
  - Trustworthy privacy software requires open source *and* vetting
  - OpenDP will get the community of experts engaged in vetting
- Flexibility
  - Every application of DP raises new technical challenges
  - OpenDP Library is designed to grow with science & practice
  - OpenDP can match users with experts to solve their problems
- Community Governance
  - All stakeholders: contributors and users, from industry, government, and academia can have an influence on the roadmap.

OpenDP

A community effort to build a trustworthy and open-source suite of differential privacy tools that can be easily adopted by custodians of sensitive data to make it available for research and exploration in the public interest.

Use Cases

- Archival data repositories (e.g. Dataverse, ICPSR, Zenodo) enabling secondary reuse and replication.
- Government agencies making data available to the public, both for official statistics and open data mandates.
- Data for good programs at companies, sharing data on customers with public and researchers
- Analytics on customer data, internally & with partners
- Machine learning on customer data

How we got started in 2019

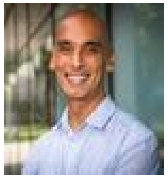
- Grants from the Sloan Foundation

- Collaboration with Microsoft on a DP curator application

- The Privacy Tools Project, funded by NSF, the US Census Bureau, the Sloan Foundation, and Google.

OpenDP Executive Committee

Gary King
Faculty Director



Salil Vadhan
Faculty Director

Stefano Iacus
Director of Data Science, IQSS

Annie Wu
Program Director

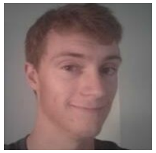
James Honaker
Chief Privacy Engineer

Andrew Vyrros
Senior Library Architect

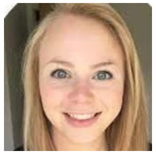
Development Team & Staff


Silvia Casacuberta Puig
Intern


Raman Prasad
Technical Lead for Research Software


Michael Shoemate
Senior Software Developer


Connor Wagaman
Intern


Georgina Evans
PhD Candidate


Zachary Ratliff
Graduate Student


Patrick Song
Undergraduate Student


Vicki Xu
Undergratuate Student


Lindsay Froess
Project Coordinator


Jayshree Sarathy
Graduate Student

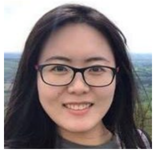

Grace Tian
Intern


Hanwen Zhang
Graduate Student


Ellen Kraffmiller
Technical Lead


Koissi Savi
Postdoc


Andy Vyrros
Senior Library Architect


Wanrong Zhang
Postdoc

Our Microsoft Collaborators

Joshua Allen
Principal Data Scientist

John Kahan
VP & Chief Data Analytics Officer

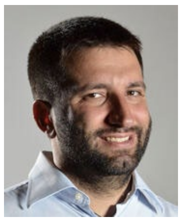
Mayana Pereria
Senior Data Scientist

Sarah Bird
Principal Program Manager

Scott McCullers
Program Manager

Kevin White
Sr. Director Program Management

2019-20 Ad Hoc Design Committee



Marco Gaboardi
Boston University

Merce Crosas
IQSS Chief Data Science & Technology Officer

Michael Hay
Colgate University

Gary King
Faculty Co-Director

Aleksandra Korolova
University of Southern California

James Honaker
Chief Privacy Engineer

Ilya Mironov
Facebook AI

Salil Vadhan
Faculty Co-Director

Advisory Board

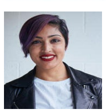
John Abowd
US Census Bureau,
Cornell University

Gilles Barthe
Max Planck Institute,
IMDEA Software Institute

Barbara Bierer
Brigham & Women's Hospital
Harvard Medical School

Sarah Bird
Microsoft

danah boyd
Microsoft Research



Rumman Chowdury
CEO, Founder of Parity AI

Cynthia Dwork
Harvard University
Radcliffe Institute
Microsoft Research

Gerome Miklau
(co-chair)
UMass, Amherst

John Friedman
Brown University

Jeff Gill
American University

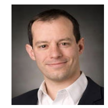
Daniel Goroff
Alfred P. Sloan
Foundation
(ex officio)

Frauke Kreuter
University of Mannheim
Institute for Employment
Research, Germany

Orran Krieger
PI Mass Open Cloud
Boston University

David Lazer
Northeastern University

Margaret Levenstein
University of Michigan

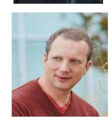


Adam Smith
(co-chair)
Boston University

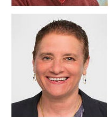
Katrina Ligett
Hebrew University

Carlos Maltzahn
UC Santa Cruz
CROSS

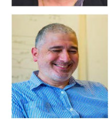
Kenneth Mandl
Harvard Medical School
Boston Children's
Hospital



Ilya Mironov
Facebook



Helen Nissenbaum
Cornell University



Kobbi Nissim
Georgetown University

Dina N. Paltoo
National Heart, Lung and
Blood Institute

Merce Crosas
Government
Generalitat de Catalunya

Jules Polonetsky
Future of Privacy Forum

Aaron Roth
University of Pennsylvania

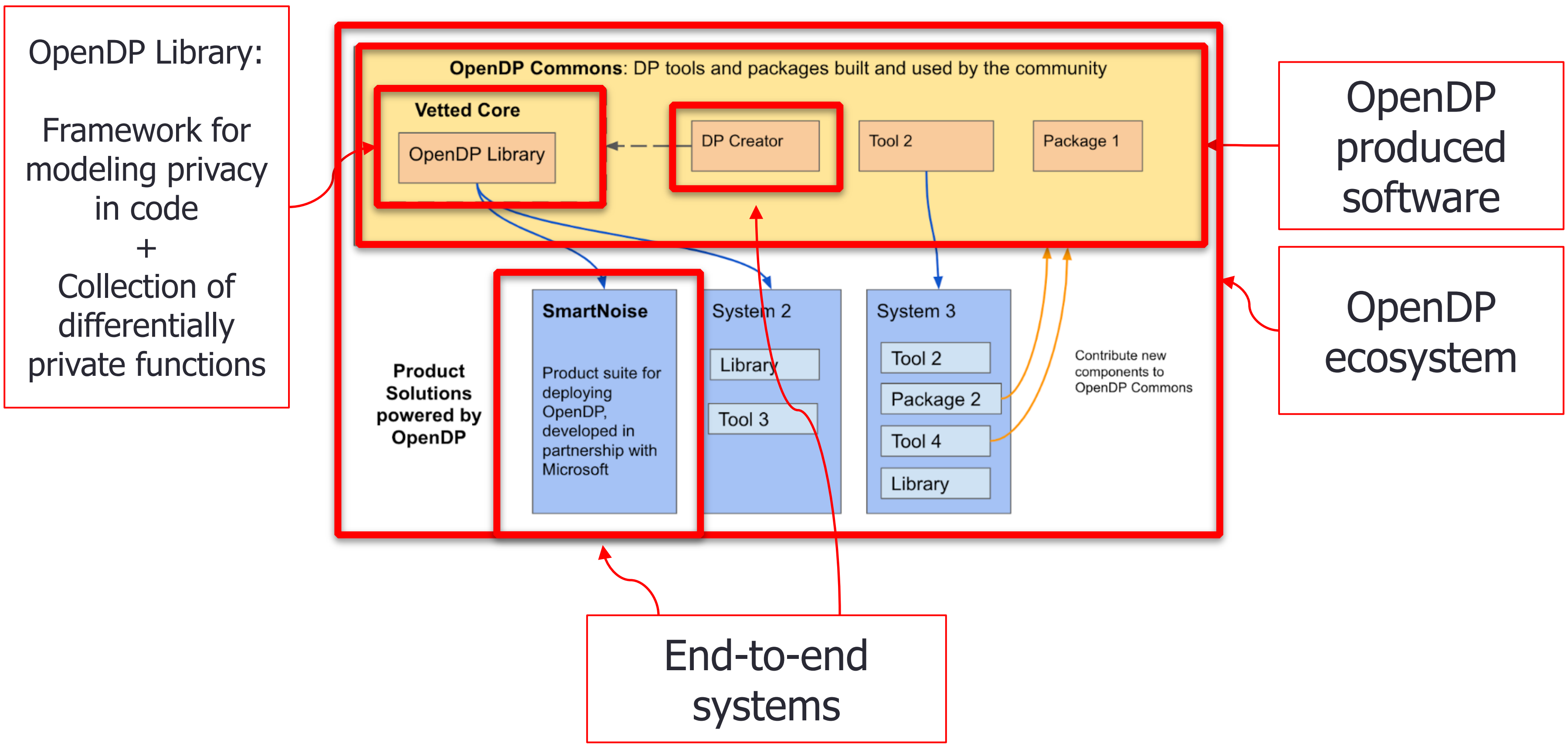
Aleksandra Slavkovic
Pennsylvania State
University

Dawn Song
University of California,
Berkeley

Latanya Sweeney
Harvard University

Omer Tene
International Association
of Privacy Professionals

OpenDP Software Elements



SmartNoise (OpenDP+Microsoft)

README.md

License MIT

SmartNoise SDK: Tools for Differential Privacy on Tabular Data

The SmartNoise SDK includes 2 packages:

- smartnoise-sql: Run differentially private SQL queries
- smartnoise-synth: Generate differentially private synthetic data

To get started, see the examples below. Click into each project for more detailed examples.

SQL

python 3.7 | 3.8 |3.9 | 3.10

Install

```
pip install smartnoise-sql
```

+ 14 contributors

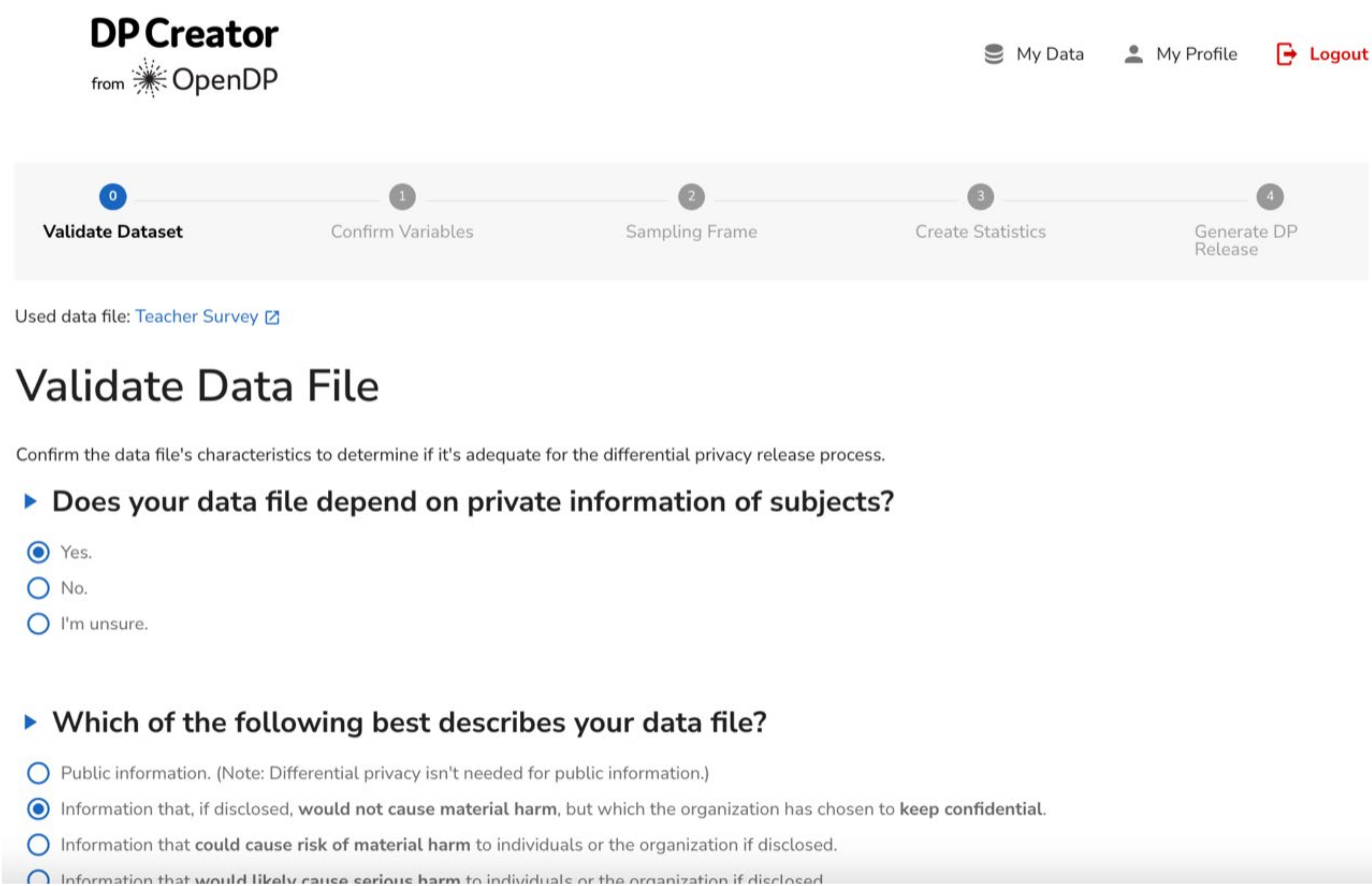
Environments 1

github-pages Active

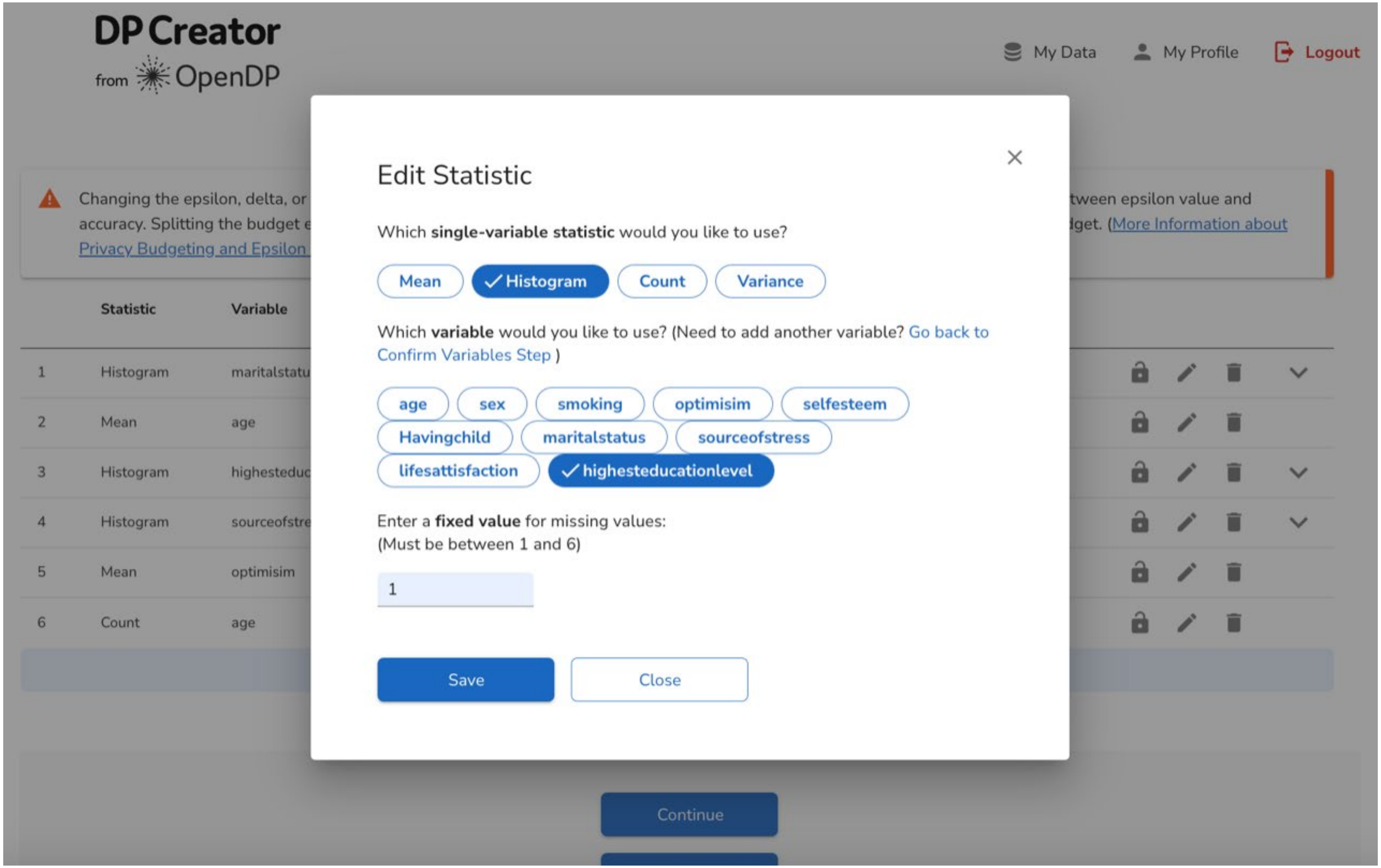
Languages

- Python 97.3%
- ANTLR 1.0%
- HTML 0.6%
- Makefile 0.5%
- HCL 0.3%
- TSQL 0.3%

DP Creator: a Tool for Non-Experts



Creating Statistics



Partitioning the Privacy-Loss Budget

DP Creator
from OpenDP

My Data   My Profile   Logout

Create Statistics

Create the statistics you would like to release. Confirm the default levels or edit them to change the degree of noise or interference you'd like to add. The default values distribute epsilon evenly across variables.

Epsilon (ε)	Delta (δ)	Confidence Level
1	0	95%

More information about Epsilon

More information about Delta

Changing the epsilon, delta, or significance level will directly impact the privacy settings. Every DP statistic has a different tradeoff between epsilon value and accuracy. Splitting the budget evenly can diminish the usefulness of statistics. More complex statistics will generally require more budget. (More Information about Privacy Budgeting and Epsilon (ε))

	Statistic	Variable	Handle Missing Values	Epsilon	Delta	Error
1	Histogram	maritalstatus	Insert Fixed Value: 1	0.167	NA	30.5
2	Mean	age	Insert Fixed Value: 42	0.167	NA	0.141
3	Histogram	highesteducationlevel	Insert Fixed Value: 1	0.167	NA	28.7
4	Histogram	sourceofstress	Insert Fixed Value: 9	0.167	NA	31.2
5	Mean	optimisim	Insert Fixed Value: 15	0.167	NA	0.0616
6	Count	age		0.167	NA	18.0

Publishing the Release

DP Release

FultonPUMS5full (1).csv

Current Status: Release Completed

Created: December 7, 2022 at 18:59:18:793264

DP Creator
from OpenDP

Differentially Private Release
FultonPUMS5full (1).csv
7 December, 2022

This report contains differentially private (DP) statistics calculated by the DP Creator application using the file "FultonPUMS5full (1).csv." file named "FultonPUMS5full (1).csv" which was uploaded by user Dev Administrator.

Please read the report carefully, especially in regard to the usage of these statistics. If you have any questions, please email us info@opendp.org.

Note: If you are using Adobe Acrobat, a JSON version of this data is attached to this PDF as a file named "release_data_2ea6c085-e168-45df-8b6d-ffda6209c3ce.json."

Contents

Publishing the Release

DP Release

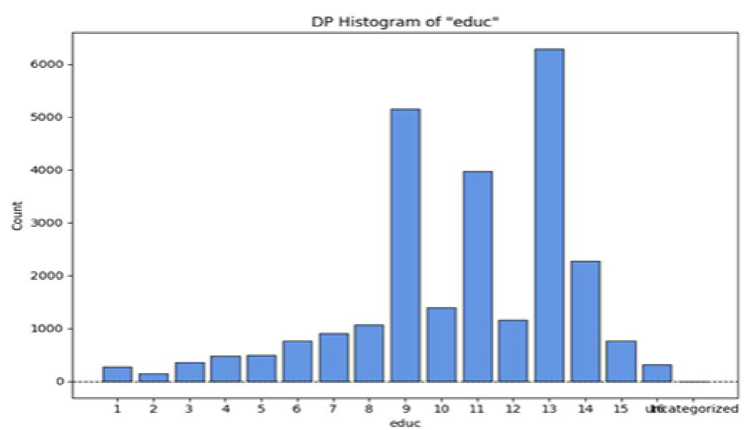
FultonPUMS5full (1).csv

Current Status: Release Completed

Created: December 7, 2022 at 18:59:18:793264

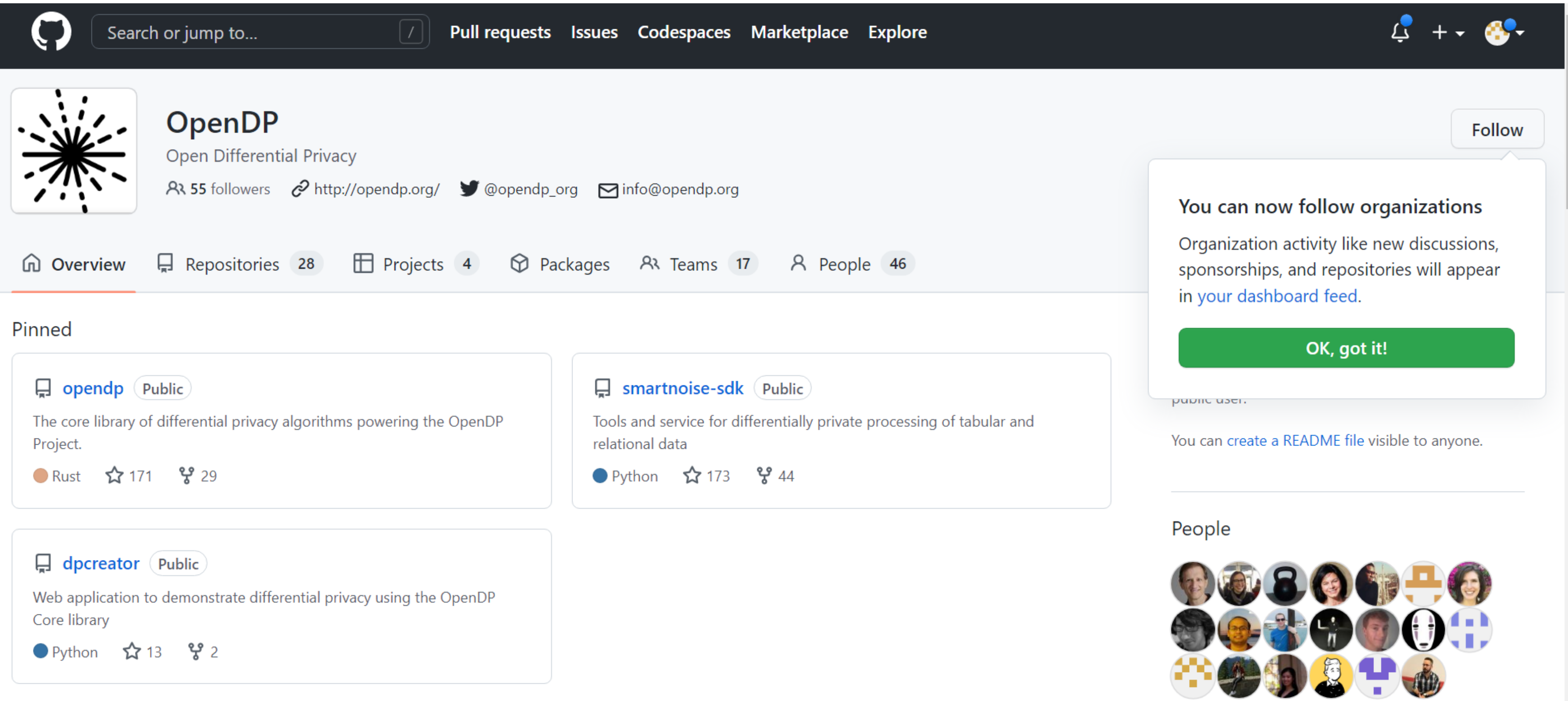
1.2. educ - histogram

Result. A **DP Histogram** has been calculated for the variable **educ**. The histogram is shown below:



Negative values. The histogram contains negative values. For more information on how to use this data, please see the section **5. Negative Values**

The OpenDP Repository



Users of OpenDP/SmartNoise SDK

Microsoft

UN PET Lab
United Nations
Privacy Enhancing Technologies Lab

Oblivious

Mass.gov

LINCOLN LABORATORY
MASSACHUSETTS INSTITUTE OF TECHNOLOGY

CrisisReady

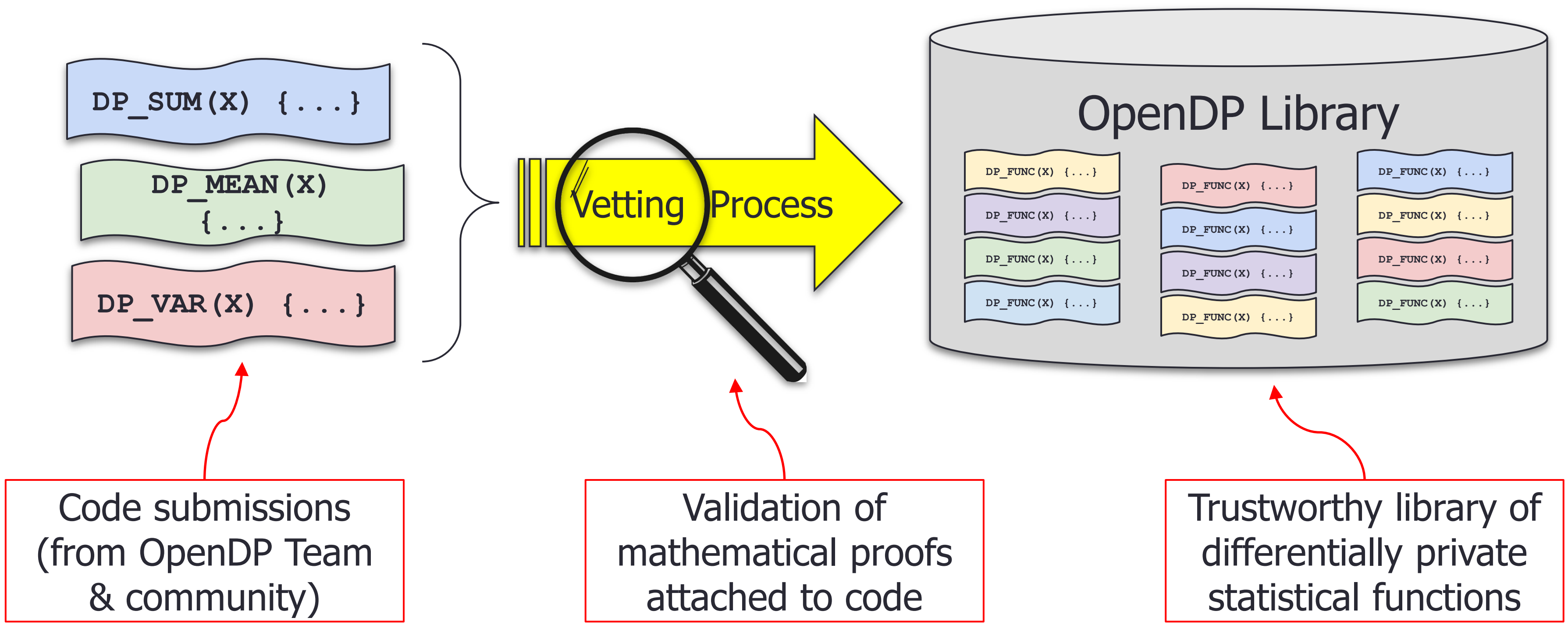
URBAN
INSTITUTE

DATAFLEETS

Sarus

cuebiq

Building a Trustworthy Library



Widespread Underestimation of Sensitivities in DP Libraries [CCS `22]

Sílvia Casacuberta, Michael Shoemate,
Salil Vadhan, and Connor Wagaman

OpenDP

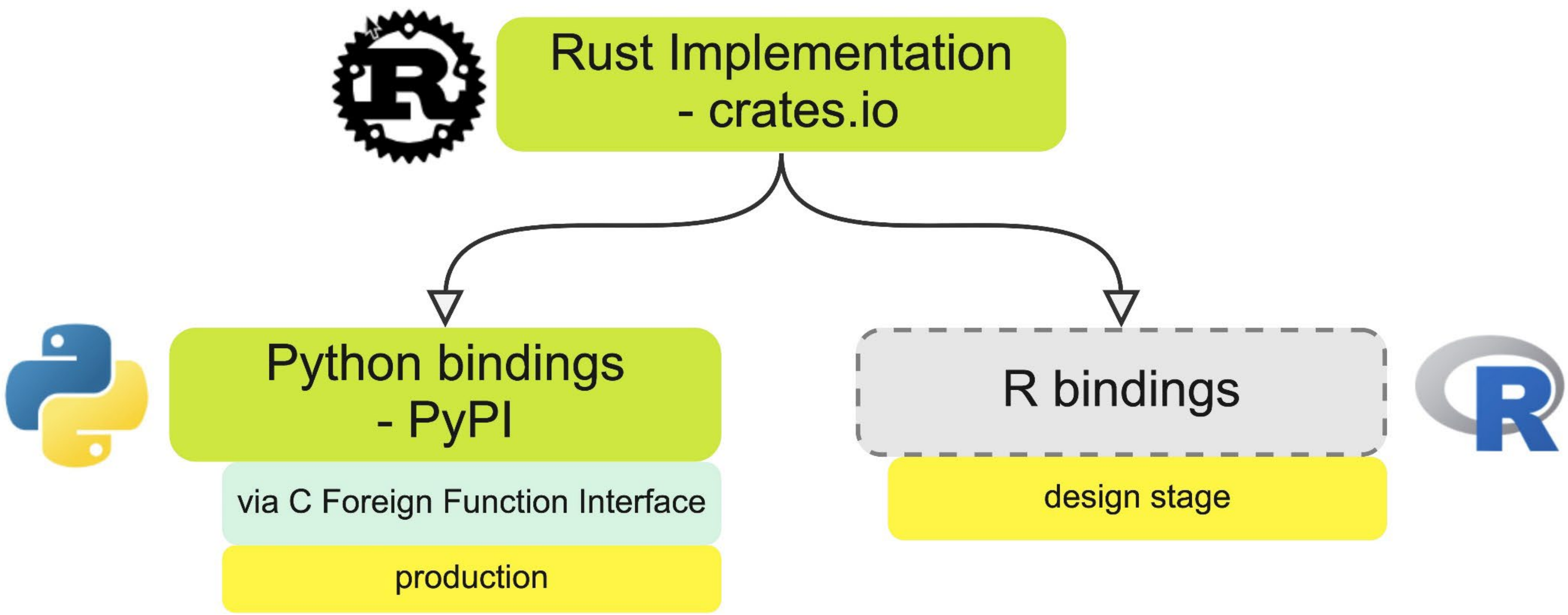
HARVARD UNIVERSITY

BOSTON UNIVERSITY

ALFRED P. SLOAN FOUNDATION

SIMONS FOUNDATION

The OpenDP Library



Integration of Proofs and Code

Struct opendp::measures::ZeroConcentratedDivergence

source · [−]

```
pub struct ZeroConcentratedDivergence<Q>(PhantomData<Q>);
```

[−] ρ-zero concentrated differential privacy.

The greatest zero-concentrated divergence between any randomly selected subset of the support.

Proof Definition

`d`-closeness

For any two vectors $u, v \in \texttt{D}$ and any d of generic type $\texttt{Q}$, define P and Q to be the distributions of $M(u)$ and $M(v)$. We say that u, v are d-close under the alpha-Renyi divergence measure (abbreviated as D_α) whenever

$$D_\alpha(P \| Q) = \frac{1}{1-\alpha} \mathbb{E}_{x \sim Q}\left[\ln\left(\frac{P(x)}{Q(x)}\right)^\alpha\right] \leq d\alpha.$$

for all possible choices of $\alpha \in (1, \infty)$.

A Sample Proof

fn make_count

Sílvia Casacuberta, Grace Tian, Connor Wagaman

This proof resides in **"contrib"** because it has not completed the vetting process.

Proves soundness of `make_count` in `mod.rs` at commit f5bb719 (outdated[1]).

`make_count` returns a Transformation that computes a count of the number of records in a vector. The length of the vector, of type `usize`, is exactly casted to a user specified output type `TO`. If the length is too large to be represented exactly by `TO`, the cast saturates at the maximum value of type `TO`.

Vetting History

- Pull Request #513

1 Hoare Triple

Precondition

- `TIA` (atomic input type) is a type with trait `Primitive`. `Primitive` implies `TIA` has the trait bound:
  - `CheckNull` so that `TIA` is a valid atomic type for `AllDomain`
- `TO` (output type) is a type with trait `Number`. `Number` further implies `TO` has the trait bounds:
  - `InfSub` so that the output domain is compatible with the output metric
  - `CheckNull` so that `TO` is a valid atomic type for `AllDomain`
  - `ExactIntCast` for casting a vector length index of type `usize` to `TO`. `ExactIntCast` further implies `TO` has the trait bound:
    - `ExactIntBounds`, which gives the `MAX_CONSECUTIVE` value of type `TO`
  - `One` provides a way to retrieve `TO`'s representation of 1
  - `DistanceConstant` to satisfy the preconditions of `new_stability_map_from_constant`

Pseudocode

```
def make_count():
    input_domain = VectorDomain(AllDomain(TIA))
    output_domain = AllDomain(TO)

    def function(data: Vec[TIA]) -> TO:
        size = input_domain.size(data)
        try:
            return TO.exact_int_cast(size)
        except FailedCast:
            return TO.MAX_CONSECUTIVE
```

[1] See new changes with `git diff f5bb719..c81deb8e rust/src/transformations/count/mod.rs`

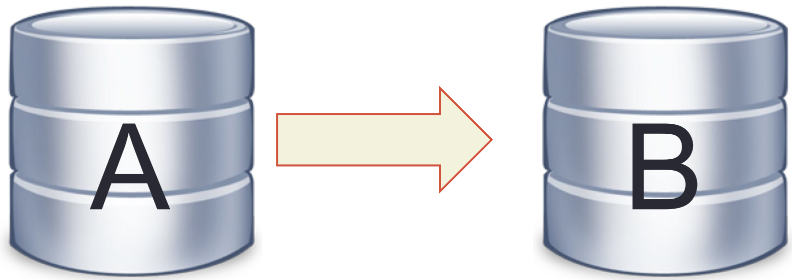
The OpenDP Programming Framework
[Gaboardi-Hay-V. `20]

- Generality in privacy definitions & algorithms
  - Pure DP, approximate DP, concentrated DP, f-DP, etc.
  - Node-level privacy in graphs, user-level privacy in streams, etc.
- Generality in privacy calculus
  - Composition, amplification by subsampling, group privacy, etc.
- Safe extensions of framework with vetted contributions
  - Clear spec for each component's privacy-relevant properties
- Interactive DP algorithms as first-class citizens
  - Adaptive composition, sparse vector, etc.
  - Still in implementation!
- Implementation in Rust w/Python bindings

Transformations and Measurements

Transformations:

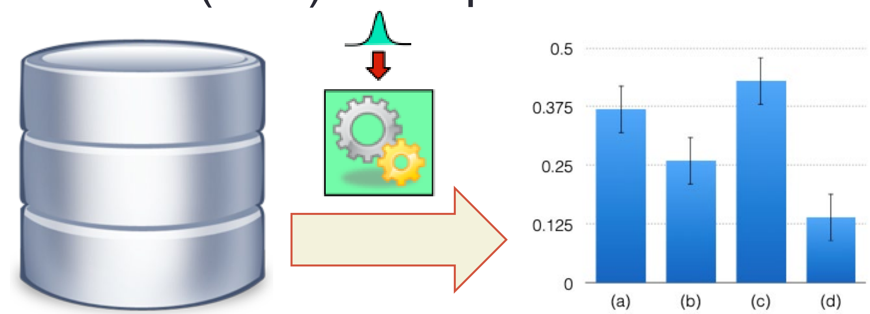
Function from data(sets) to data(sets).



Transformation Attributes
- Input domain
- Input metric
- Output domain
- Output metric
- Function
- Stability map

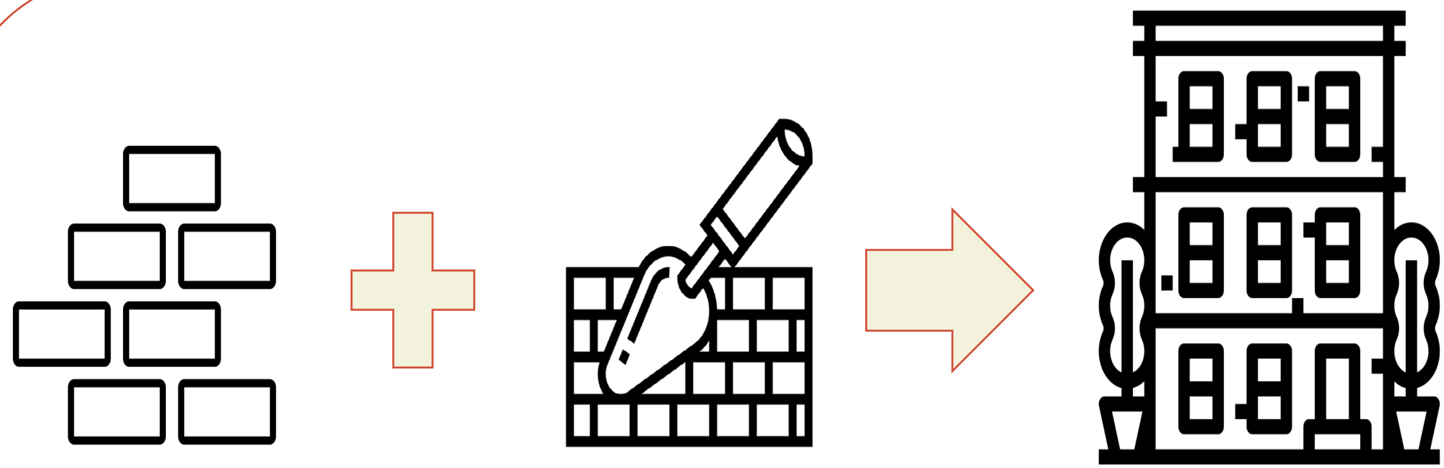
Measurements:

Randomized functions from data(sets) to outputs.



Measurement Attributes
- Input domain
- Input metric
- Output measure
- Function
- Privacy map

Combinators



Measurements

&

Transformations

Combinators,

e.g. Chaining,

Composition,

Post-processing

Compl

Privacy proof automatically derived

Privacy Calculus

To implement a privacy calculus based on the idea of stability we have:

- privacy maps in measurements to capture several notions of privacy. E.g. DP, approx. DP, Renyi DP, zCDP, f-DP.
- stability maps in transformations to capture general aggregate operations. E.g. sums, bounded joins.
- combination of these relations by means of combinators such as chaining and composition.

d_{out}=map(d_{in}) should imply:
if two inputs are "d_{in}-close",
then the corresponding outputs (or distributions) are "d_{out}-close".

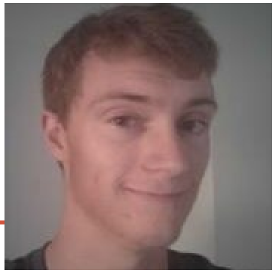
Measurement attributes

- Input domain
- Input metric
- ***Output measure***
- Function
- ***Privacy map***

Transformation attributes

- Input domain
- Input metric
- Output domain
- Output metric
- Function
- ***Stability map***

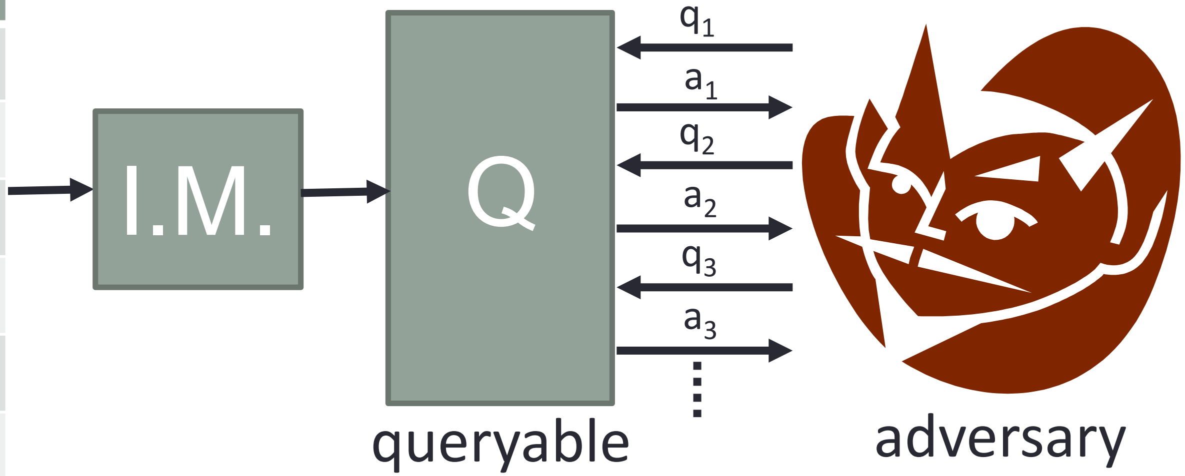
Let’s see how this works in code!



1. Open Jupyter notebook: shorturl.at/cimp8
   - Can use in your browser
   - Or File->Download to use local Python installation
2. Fill out audience poll: strawpoll.com/continent

Interactive Measurements

Sex	Blood	…	HIV?
F	B	…	Y
F	A		Y
M	O	…	N
M	O	…	Y
F	A	…	N
M	B	…	Y



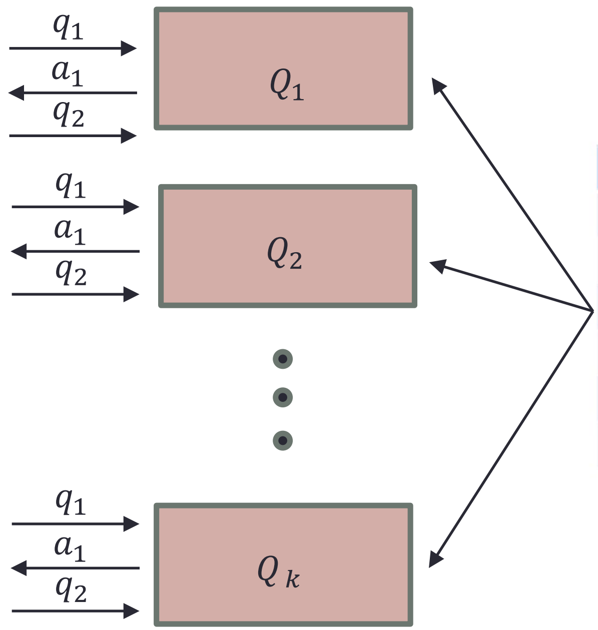
Requirement: for all neighboring u, v

$$\mathrm{View}_A\big(A \leftrightarrow Q(u)\big) \approx_{\varepsilon,\delta} \mathrm{View}_A\big(A \leftrightarrow Q(v)\big)$$

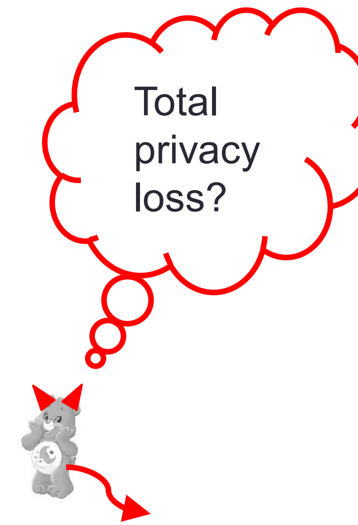
- Models adaptive composition, privacy filters, sparse vector, etc.
- First-class citizen in OpenDP framework
- Currently in implementation!

Concurrent Composition

New challenge: an adversary can arbitrarily interleave its queries to the different queryables



Sex	Blood	⋯	HIV
F	B	⋯	Y
M	A	⋯	N
M	O	⋯	N
M	O	⋯	Y
F	A	⋯	N
M	B	⋯	Y



[V.-Wang `21, Lyu `22, V.-Zhang `22]: Most standard composition theorems extend to concurrent composition.

Library Roadmap

<u>Near term (few months):</u>

- Ramp up external contributions
- More algorithms!
- Interactive Measurements
- R bindings

<u>Longer term:</u>

- Data interchange (Apache Arrow)
- Large-scale/external compute
- Federated learning/distributed models
- Beyond tabular data
- Benchmarking suite

Project Goals for 2023

- Scale up Use Cases & Contributions
- Development and functionality driven by use cases
- We're hiring! Community manager, research scientist, technical writer, support engineer, interns, …
- 3rd OpenDP Community Meeting
- Community Working Groups
  - Educational Materials
  - Statistical Uncertainty Measures
  - Best Practices for Using DP

Join the Community!

opendp.org

OpenDP

About ▾ Opportunities ▾ Community ▾ Software ▾ People ▾ Events Blog

OpenDP Library v0.6 Released! | Application Open for the 2023 Fellows Program | We Are Hiring

Developing Open Source Tools for Differential Privacy

OpenDP is a community effort to build trustworthy, open-source software tools for statistical analysis of sensitive private data. These tools, which we call OpenDP, will offer the rigorous protections of differential privacy for the individuals who may be represented in confidential data and statistically valid methods of analysis for researchers who study the data.

Join Us on Slack, Github, Mailing List!

Learn more about us