CS208: Applied Privacy for Data Science
Implementing (Centralized) Differential Privacy: Programming Interfaces for DP

James Honaker & Salil Vadhan

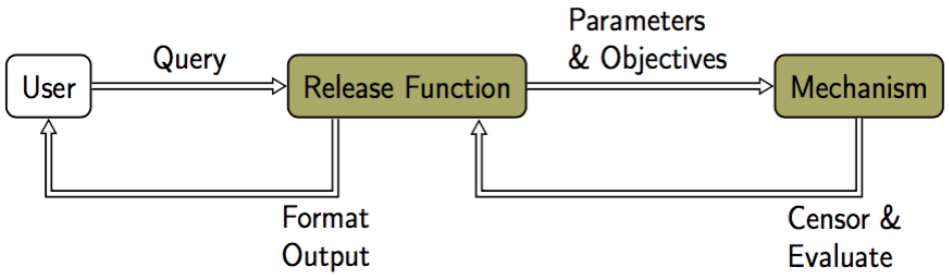
School of Engineering & Applied Sciences
Harvard University

March 15, 2019

CRCS Center for Research on Computation and Society
at Harvard John A. Paulson School of Engineering and Applied Sciences

PSI Library

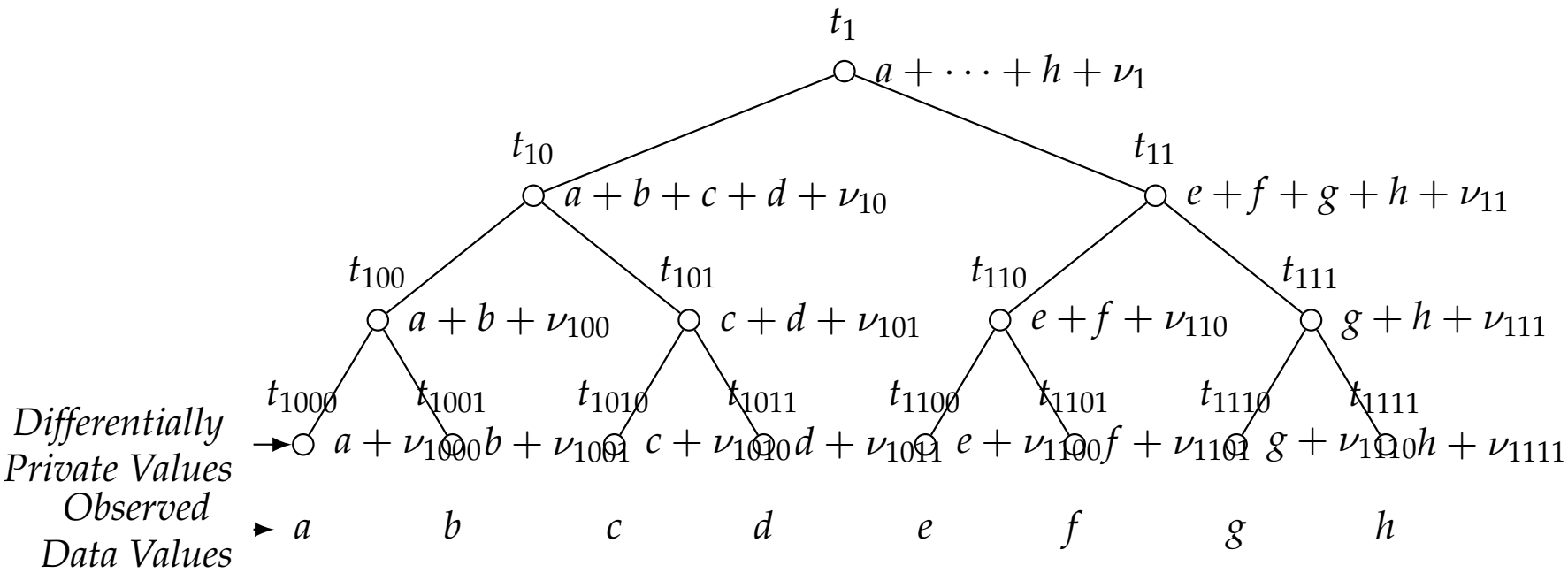
PSI Library



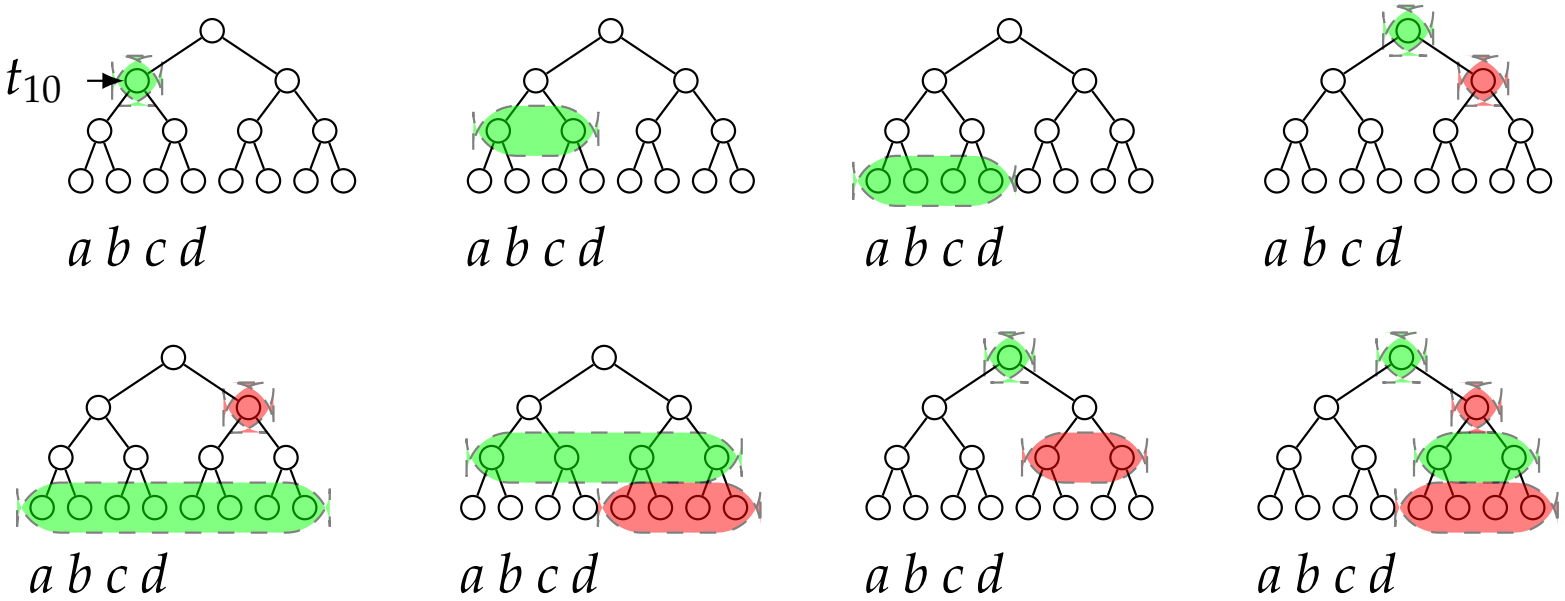
- Modular architecture
- Class based, using R
- Contains common exploratory statistics
  - Univariate descriptive statistics, such as means, quantiles, histograms, and approximate cumulative distribution functions.
  - Basic statistical estimators, matching algorithms and difference-of-means tests for causal inference, and low-dimensional linear, logit, probit and poisson regression.
- Reusable

Example from R Library

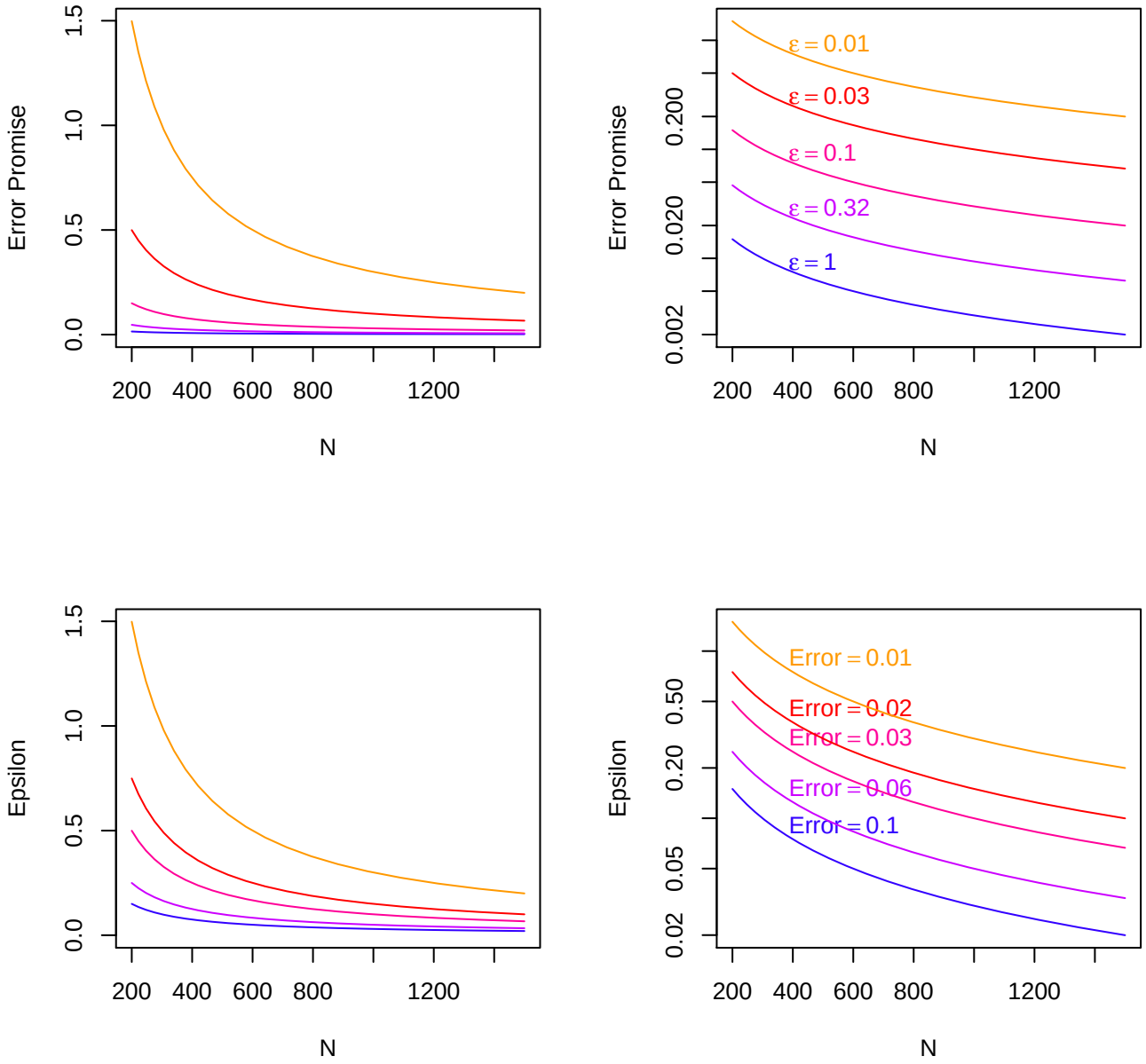
```
# Generate and print dp release of mean of age

dp.release <- dpMean$new(mechanism='mechanismLaplace',
    var.type='numeric',
    variable="age",
    n=10000,
    epsilon=0.1,
    delta=10^-6,
    rng=c(0,110))
dp.release$release(my_data)
```



Example tree, showing leaf values a through h below, node labels above, and differentially private values to the right with noise ν.



Eight of the possible ways of using the nodes on the tree to estimate $a + b + c + d$, the sum of the left four leaves. Nodes in green are added, and nodes in red subtracted to create an estimate.


Error Promise
0.0
0.5
1.0
1.5
200
400
600
800
1200
N
Error Promise
0.002
0.020
0.200
ε = 0.01
ε = 0.03
ε = 0.1
ε = 0.32
ε = 1
200
400
600
800
1200
N
Epsilon
0.0
0.5
1.0
1.5
200
400
600
800
1200
N
Epsilon
0.02
0.05
0.20
0.50
Error = 0.01
Error = 0.02
Error = 0.03
Error = 0.06
Error = 0.1
200
400
600
800
1200
N

PSI Budgeter

Open Science: Discovery, Replication, Reuse

"Accessible and reusable data are fundamental to science in order to continuously validate and build upon previous research. Progressive expansive scientific advance rests upon access to data accompanied with sufficient information for reproducible results, a scientific ethic to maximize the utility of data to the research community, and a foundational norm that scientific communication is built on attribution."

- Crosas et al. 2015

Open Science: Discovery, Replication, Reuse

Tl;dr: **Replication is the foundation of science; reuse is the building block.**

"Accessible and reusable data are fundamental to science in order to continuously validate and build upon previous research. Progressive expansive scientific advance rests upon access to data accompanied with sufficient information for reproducible results, a scientific ethic to maximize the utility of data to the research community, and a foundational norm that scientific communication is built on attribution."

- Crosas et al. 2015

Overview

Our goal is Privacy Protection that is:

1. **Accessible:** no DP expert optimizing algorithms for a particular dataset or application
2. **General:** works across the spectrum of social science datasets
3. **Workflow-Embracing:** fits into workflow of practicing social scientists, using familiar concepts and tools

Overview

Our goal is Privacy Protection that is:

1. **Accessible:** no DP expert optimizing algorithms for a particular dataset or application
2. **General:** works across the spectrum of social science datasets
3. **Workflow-Embracing:** fits into workflow of practicing social scientists, using familiar concepts and tools
4. **Additive:** strong privacy protection *increases* the ability to explore and share and open data

Overview

Our goal is Privacy Protection that is:

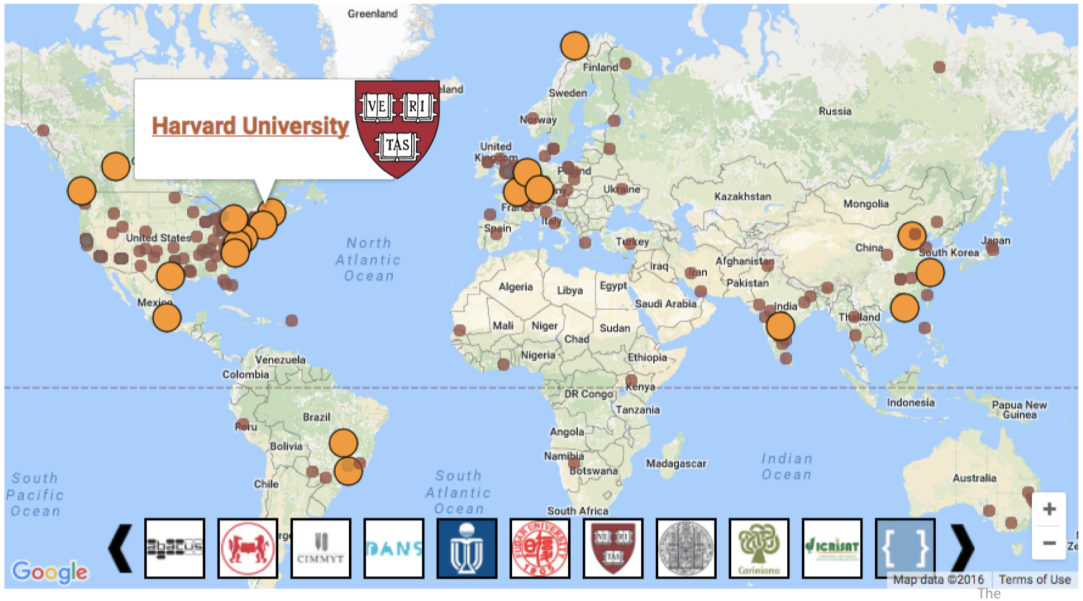
1. **Accessible:** no DP expert optimizing algorithms for a particular dataset or application
2. **General:** works across the spectrum of social science datasets
3. **Workflow-Embracing:** fits into workflow of practicing social scientists, using familiar concepts and tools
4. **Additive:** strong privacy protection *increases* the ability to explore and share and open data
5. **Tiered Access:** DP interface for wide access to rough statistical information; users can still apply for raw data (cf. Census PUMS vs RDC's)

27 Instances around the world

Used by researchers from > 500 institutions

75,000 datasets in Harvard Dataverse repository

`http://dataverse.org`



The Dataverse Project

Harvard Dataverse > Murray Research Archive Dataverse > **Two Generations of College-Educated Women: The Post-parental Phase of the Life Cycle, 1957-1979**

View Dataset Versions | Metrics | 60 Downloads

Two Generations of College-Educated Women: The Post-parental Phase of the Life Cycle, 1957-1979

Ida Davidoff; Marjorie Platt, 2007, "Two Generations of College-Educated Women: The Post-parental Phase of the Life Cycle, 1957-1979", http://hdl.handle.net/1902.1/00511, Harvard Dataverse, V2

Download Citation

If you use these data, please add this citation to your scholarly resources. Learn about Data Citation Standards.

Download | Request Access

7 Files

00511Davidoff-Platt-BoxCoverSheets.pdf
Adobe PDF - 406.7 KB - Nov 27, 2007 - 12 Downloads
Describes contents of each box of a paper data set
3. Supplementary Documentation
Download

00511Davidoff-Platt-Codebook.pdf
Adobe PDF - 27.7 MB - Nov 27, 2007 - 12 Downlo
Description of coded data variables
1. Documentation
Download

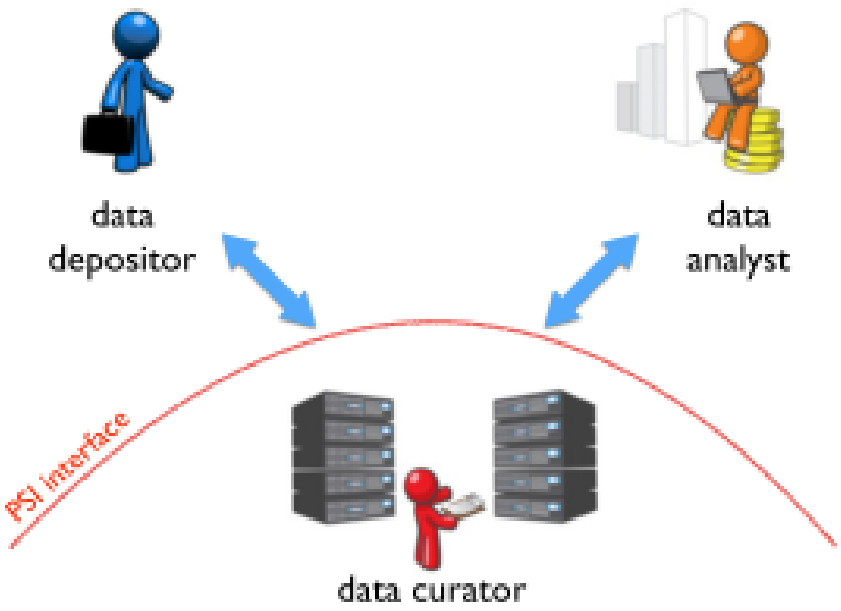
00511Davidoff-Platt-Data.por
SPSS Portable - 221.2 KB - Nov 27, 2007 - 0 Downloads
Data for Study in SPSS Portable Format
2. Data

00511Davidoff-Platt-Data.tab

Datasets are restricted due to privacy concerns

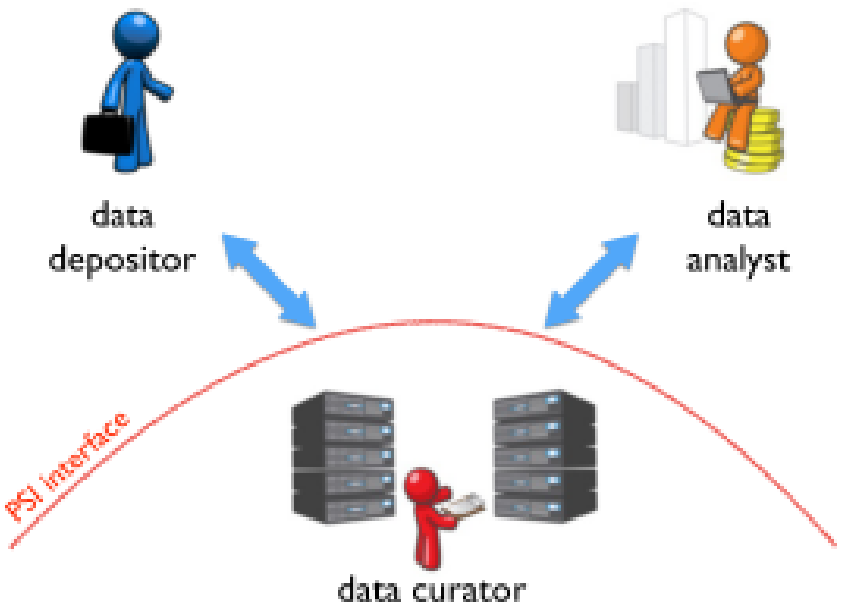
Goal: enable wider sharing while protecting privacy

Actors



data depositors: Come to deposit their sensitive dataset in a repository,
and may wish to make DP access available.

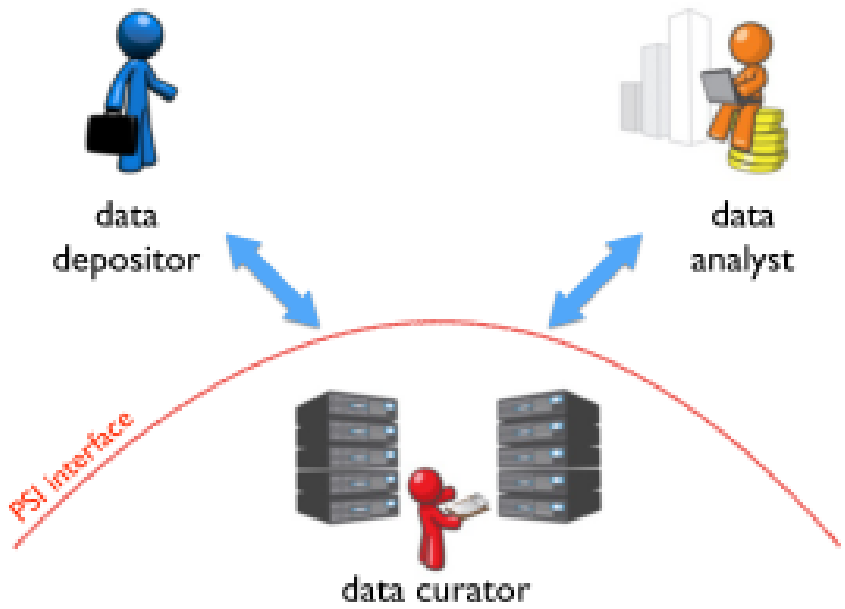
Actors



data depositors: Come to deposit their sensitive dataset in a repository,
and may wish to make DP access available.

data curators: Maintain the hardware and software on which PSI runs and the accompanying repository infrastructure

Actors

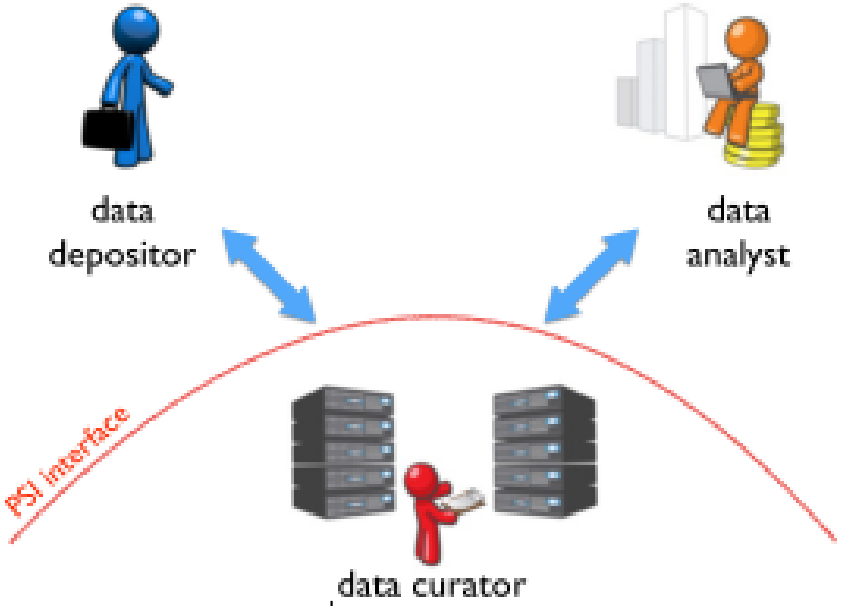


data depositors: Come to deposit their sensitive dataset in a repository,
and may wish to make DP access available.

data curators: Maintain the hardware and software on which PSI runs and the accompanying repository infrastructure

data analysts: Come to access sensitive datasets in the repository, often with the goal of data exploration

Actors



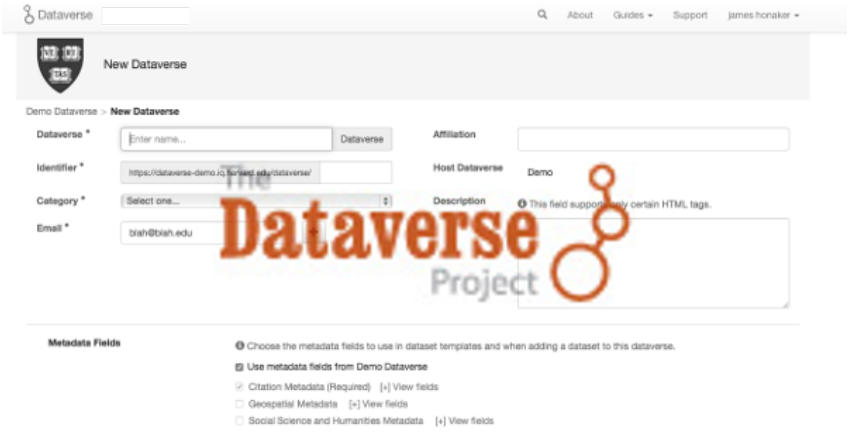
	Level of Trust	Responsibilities
data depositors:	Trusted	Define the Appropriate Privacy Protection
data curators:	Trusted	Maintain Integrity of Software System
data analysts:	**Untrusted** **Semi-Trusted**	None Abide by DUA

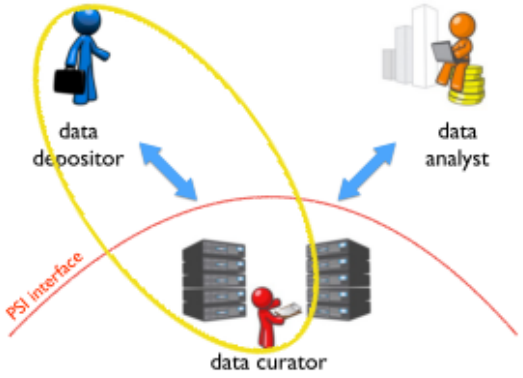
workflow

http://psiprivacy.org/about

workflow

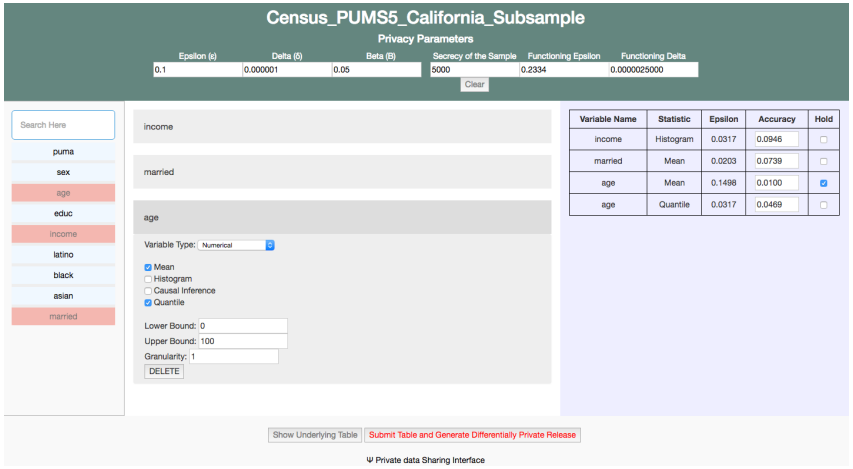

Upload
Budget
Release
Explore
Query

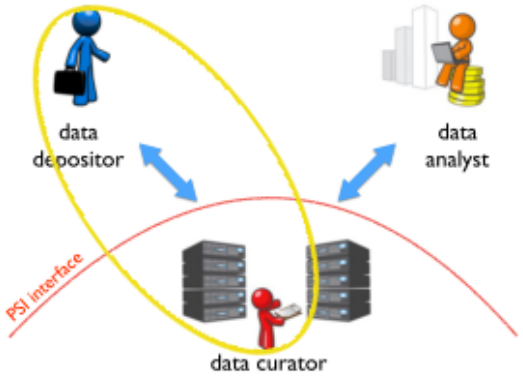

Dataverse
New Dataverse
Demo Dataverse > New Dataverse
Dataverse *
Identifier *
Category *
Email *
Affiliation
Host Dataverse
Demo
Description
Metadata Fields
Choose the metadata fields to use in dataset templates and when adding a dataset to this dataverse.
Use metadata fields from Demo Dataverse
Citation Metadata (Required)
Geospatial Metadata
Social Science and Humanities Metadata


data depositor
data analyst
data curator
PSI interface

workflow

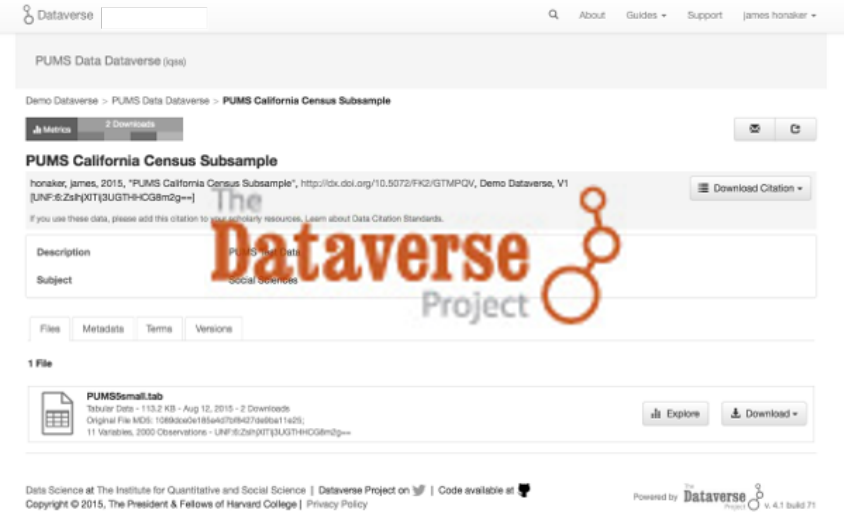

Upload
Budget
Release
Explore
Query

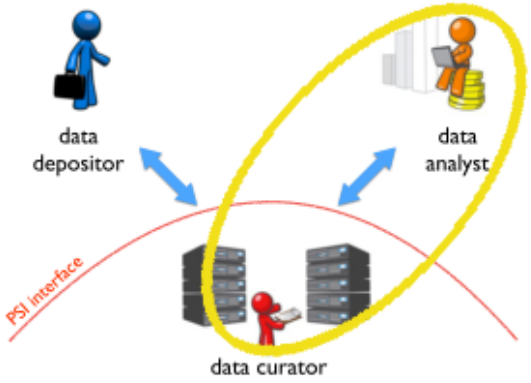

Census_PUMS5_California_Subsample
Privacy Parameters
Epsilon (ε)
Delta (δ)
Beta (β)
Secrecy of the Sample
Functioning Epsilon
Functioning Delta
0.1
0.000001
0.05
5000
0.2334
0.0000025000
Clear
Search Here
puma
sex
age
educ
income
latino
black
asian
married
income
married
age
Variable Type: Numerical
Mean
Histogram
Causal Inference
Quantile
Lower Bound: 0
Upper Bound: 100
Granularity: 1
DELETE
Variable Name | Statistic | Epsilon | Accuracy | Hold
income | Histogram | 0.0317 | 0.0946
married | Mean | 0.0203 | 0.0739
age | Mean | 0.1498 | 0.0100
age | Quantile | 0.0317 | 0.0469
Show Underlying Table
Submit Table and Generate Differentially Private Release
Private data Sharing Interface


data depositor
data analyst
PSI interface
data curator

workflow

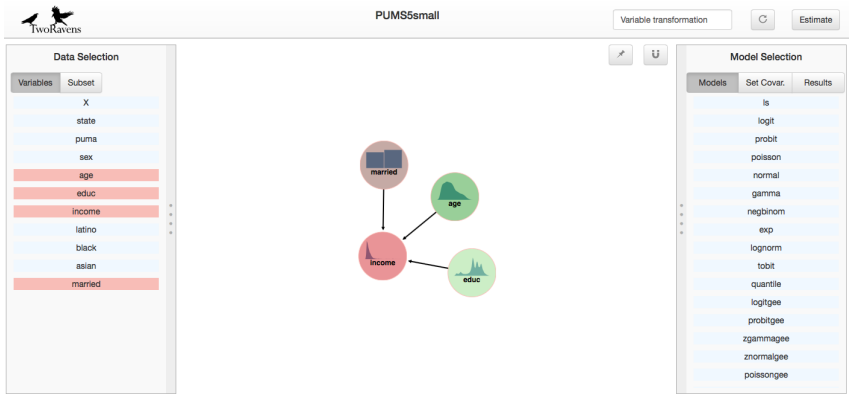
Upload → Budget → Release → Explore → Query

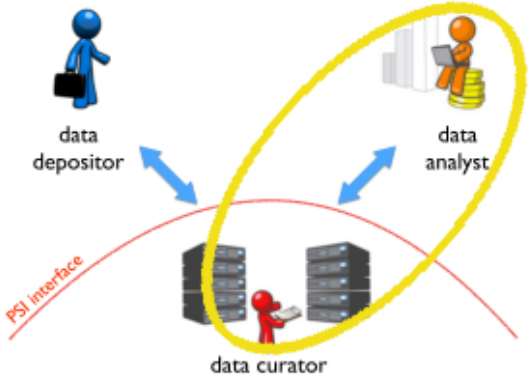




workflow

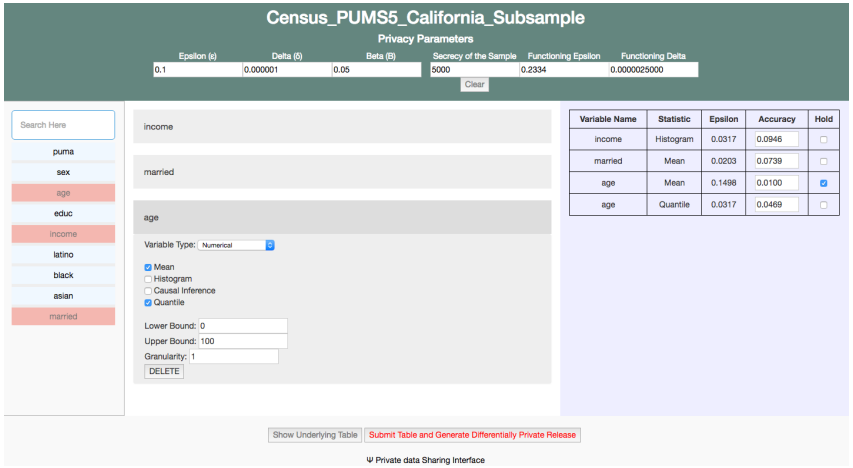
Upload → Budget → Release → Explore → Query

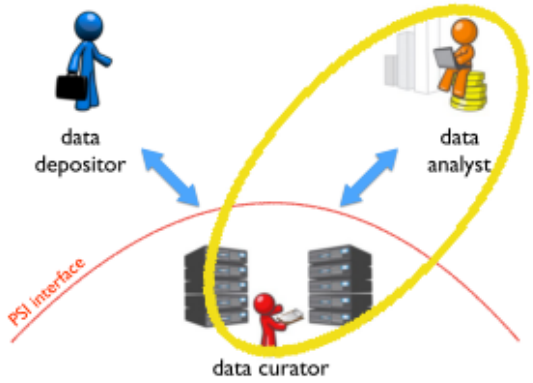




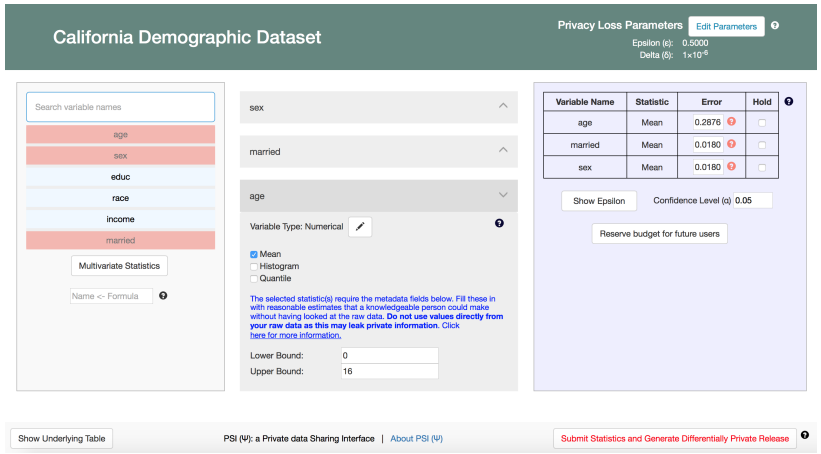
workflow


Upload
Budget
Release
Explore
Query


Census_PUMS5_California_Subsample
Privacy Parameters
Epsilon (ε)
Delta (δ)
Beta (β)
Secrecy of the Sample
Functioning Epsilon
Functioning Delta
0.1
0.000001
0.05
5000
0.2334
0.0000025000
Clear
Search Here
puma
sex
age
educ
income
latino
black
asian
married
income
married
age
Variable Type: Numerical
Mean
Histogram
Causal Inference
Quantile
Lower Bound: 0
Upper Bound: 100
Granularity: 1
DELETE
Variable Name | Statistic | Epsilon | Accuracy | Hold
income | Histogram | 0.0317 | 0.0946
married | Mean | 0.0203 | 0.0739
age | Mean | 0.1498 | 0.0100
age | Quantile | 0.0317 | 0.0469
Show Underlying Table
Submit Table and Generate Differentially Private Release
Private data Sharing Interface

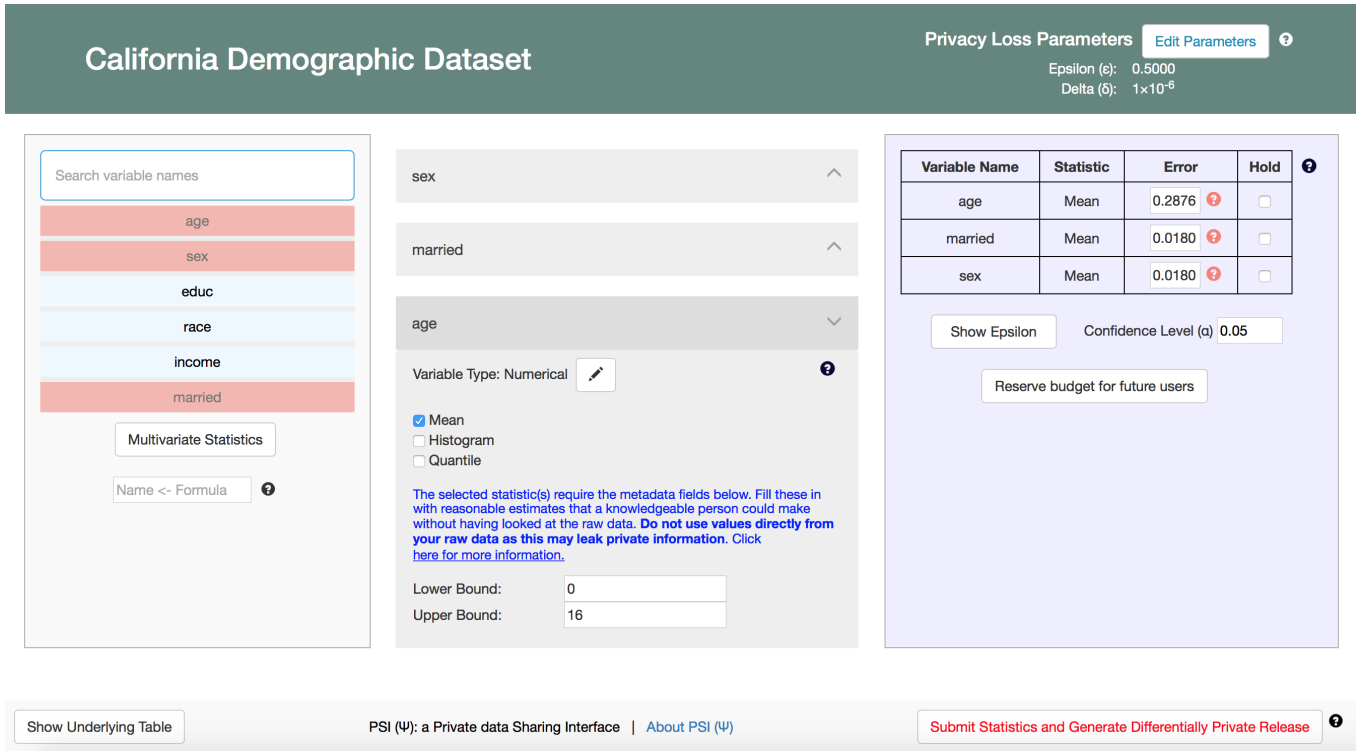

data depositor
data analyst
PSI interface
data curator

Demos

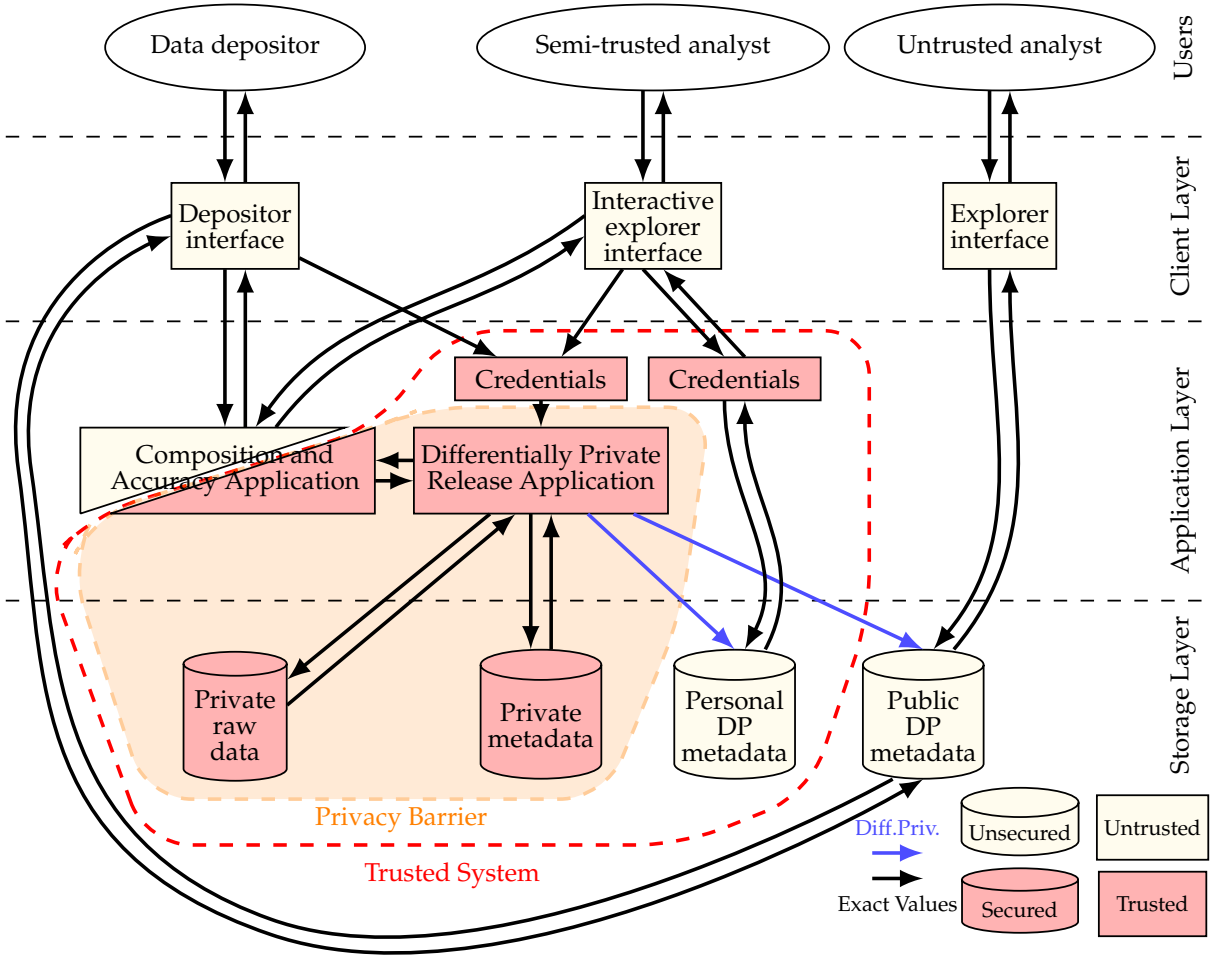


http://psiprivacy.org/

http://psiprivacy.org/about



http://2ra.vn/demos/PSIdemoCS208.mp4


Data depositor
Semi-trusted analyst
Untrusted analyst
Users
Depositor interface
Interactive explorer interface
Explorer interface
Client Layer
Credentials
Credentials
Composition and Accuracy Application
Differentially Private Release Application
Application Layer
Private raw data
Private metadata
Personal DP metadata
Public DP metadata
Storage Layer
Privacy Barrier
Trusted System
Diff.Priv.
Exact Values
Unsecured
Secured
Untrusted
Trusted

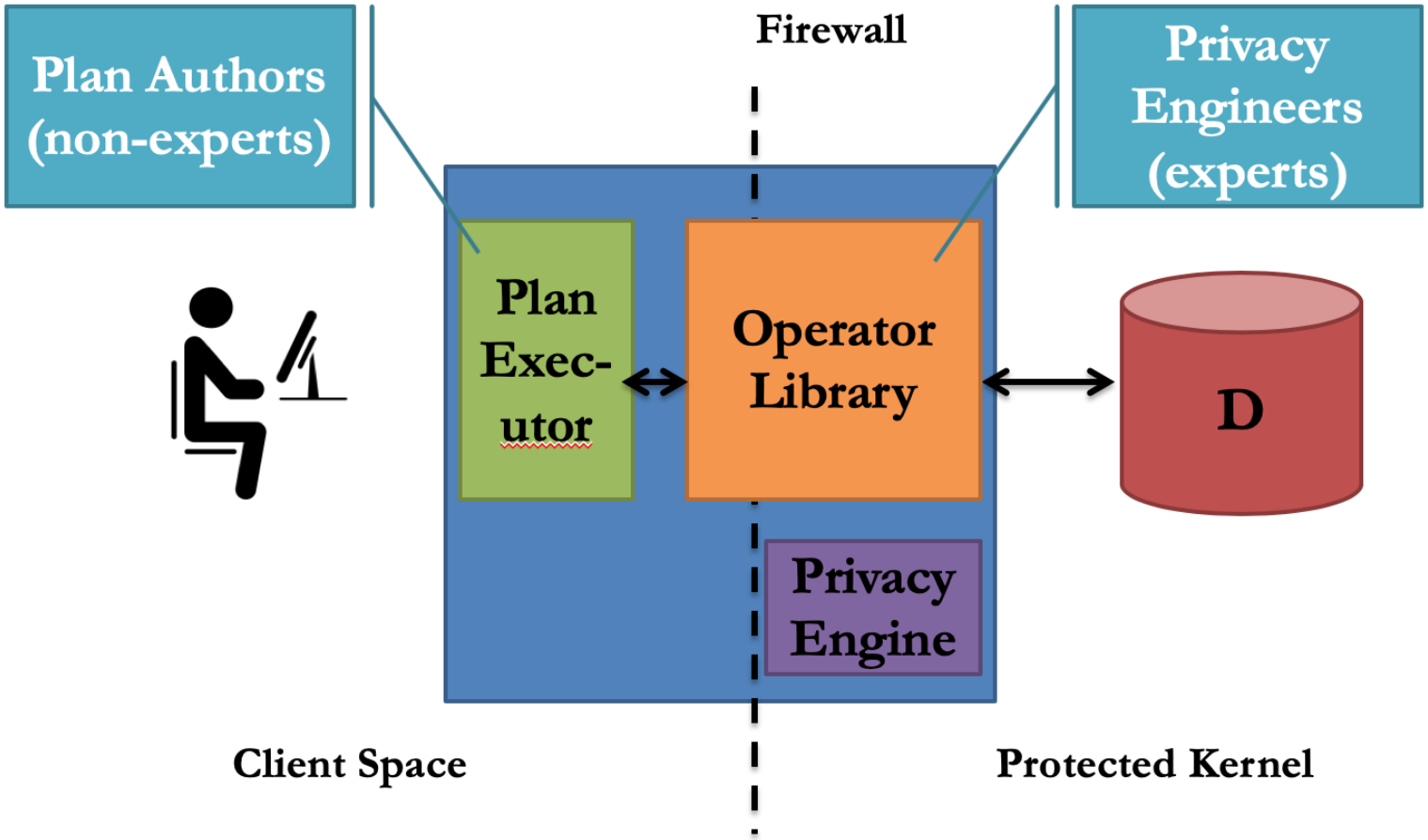
Ektelo

Ektelo

- Ektelo supports a library of vetted **operators**
- Operators encode best practices from literature
  - Allow sophisticated algorithms to be expressed
- Algorithm expressed as **plan***:* a sequence of **operators**
- Every plan automatically satisfies differential privacy.

[SIGMOD 2018] Zhang, McKenna, Kotsogiannis, Hay, M, Miklau
"Ektelo: A framework for describing differentially private algorithms"

Ektelo Implementation



Operator classes and types

Transform	
TV	T-Vectorize
TP	V-SplitByPartition
TR	V-ReduceByPartition

Inference	
LS	Least squares
NLS	Nneg Least squares
MW	Mult Weights
HR	Thresholding

Partition selection	
PA	AHPpartition
PG	Grid
PD	Dawa
PW	Workload-based
PS	Stripe(attr)
PM	Marginal(attr)

Query	
LM	Vector Laplace

Query selection	
SI	Identity
ST	Total
SP	Privelet
SH2	H2
SHB	HB
SG	Greedy-H
SU	UniformGrid
SA	AdaptiveGrids
SQ	Quadtree
SW	Worst-approx
SPB	PrivBayes select

Runs in the protected kernel; Affects sensitivity

Client side post-processing; does not affect privacy

Releases noisy answers; consumes privacy

Theorem: If red and orange operators are vetted, then any Ektelo plan satisfies DP

Code Reuse: Existing algorithms implemented with reusable operators.
Unified 16 implementations of Laplace mechanism

ID	Cite	Algorithm name	Plan signature
1	[8]	Identity	SI LM
2	[39]	Privelet	SP LM LS
3	[17]	Hierarchical (H2)	SH2 LM LS
4	[34]	Hierarchical Opt (HB)	SHB LM LS
5	[22]	Greedy-H	SG LM LS
6	-	Uniform	ST LM LS
7	[15]	MWEM	I:(SW LM MW)
8	[42]	AHP	PA TR SI LM LS
9	[22]	DAWA	PD TR SG LM LS
10	[6]	Quadtree	SQ LM LS
11	[33]	UniformGrid	SU LM LS
12	[33]	AdaptiveGrid	SU LM LS TP[SA LM] LS
13	NEW	DAWA-Striped	PS TP[PD TR SG LM] LS
14	NEW	HB-Striped	PS TP[SHB LM] LS
15	NEW	PrivBayesLS	SPB LM LS
16	NEW	MWEM variant b	I:(SW SH2 LM MW)
17	NEW	MWEM variant c	I:(SW LM NLS)
18	NEW	MWEM variant d	I:(SW SH2 LM NLS)

New variants of algorithms are easy to construct

Covert Channels

Other Issues in Programming DP

- Multi-relational databases
  - Standard joins have unbounded Lipschitz constant, so need to truncate
- Side-channel attacks
  - Info can be leaked through timing, approx. of real numbers, global state, exceptions, etc.
  - Constrain language & implementation to match model better.
- Verifying DP building blocks for more complex DP algs.
  - Specialized programming languages
  - Annotate programs with types to assist automated verification of DP
  - Trade-off between usability and expressiveness
  - Now can even synthesize DP algorithms from examples!
- Guidance on Privacy Budgeting
- Choice of Programming Model (eg. SQL vs. MapReduce vs. R)

Covert Channels/Side Channels

- Timing
- Floating Point
- Randomization
- Globals