CS2080: Applied Privacy for Data Science
Interfaces for DP

School of Engineering & Applied Sciences
Harvard University

April 2, 2025

DP offers a rich suite of theoretical tools

- Privacy measures: Pure-DP, approximate-DP, zCDP, f-DP, etc.
- Basic mechanisms: Laplace, Gaussian, histograms, exponential, etc.
- Basic composition, advanced composition, etc.
- DP-SGD, output perturbation, objective perturbation, etc.
- etc.

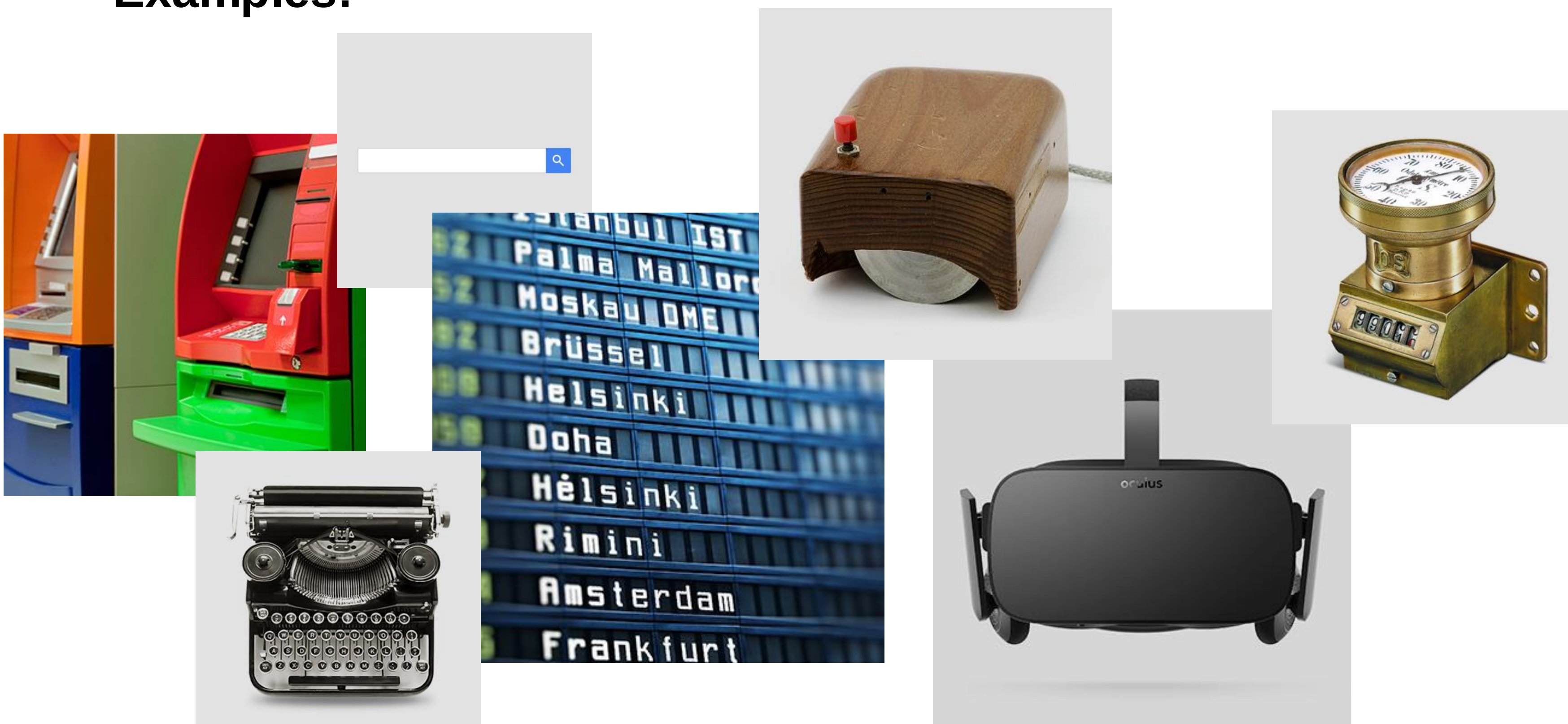
DP offers a rich suite of theoretical tools

- Privacy measures: Pure-DP, approximate-DP, zCDP, f-DP, etc.
- Basic mechanisms: Laplace, Gaussian, histograms, exponential, etc.
- Basic composition, advanced composition, etc.
- DP-SGD, output perturbation, objective perturbation, etc.
- etc.

How might people effectively interact with DP?

One answer: Interfaces!

- **Interface:** A medium through which a human interacts with a device, system, or concept
- **Examples:**



https://www.xerox.com/en-us/insights/user-interface-examples

A brief history of interfaces in HCI, as told by Card & Moran (1986)

User Technology: From Pointing to Pondering

Stuart K. Card and Thomas P. Moran
Xerox Palo Alto Research Center

From its beginning, the technology of personal workstations has been driven by visions of a future in which people would work in intimate partnership with computer systems on significant intellectual tasks. These visions have been expressed in various forms: Memex (Bush, 1945), Man-Machine Symbiosis (Licklider, 1960), NLS (Engelbart, 1963), Dynabook (Kay, 1977), and others.

The tight coupling between human and computer

1. The Vision of an Applied User Psychology

The opportunity to tackle a science of the user brought us to PARC in 1974 (collaborating with Allen Newell, as consultant). As other PARC researchers were beginning to pursue the vision of highly graphic, interactive, network-based personal workstations, we were following a vision of our own. The idea was to draw concepts from

How might we create interfaces for humans to interact with DP?

How might we create interfaces for humans to interact with DP?

How might we **design, build, and evaluate** interfaces for humans to **use** DP?

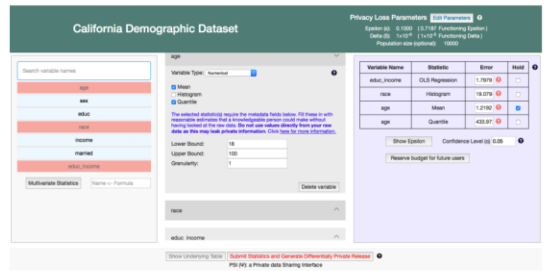
How might we create interfaces for humans to interact with DP?

How might we **design, build, and evaluate** interfaces for humans to **use** DP?

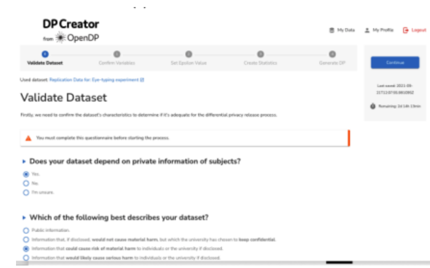
How might we **design, build, and evaluate** interfaces for **data curators and data analysts** to **use** DP?

Several interfaces have been proposed for DP

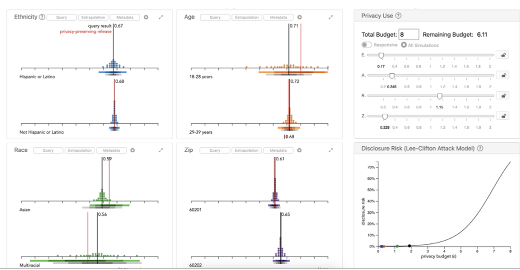
DP Comp (Hay et al. 2016)



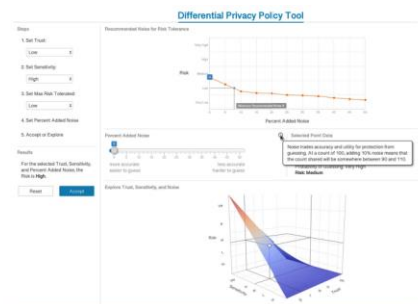
PSI (Gaboardi et al. 2018)



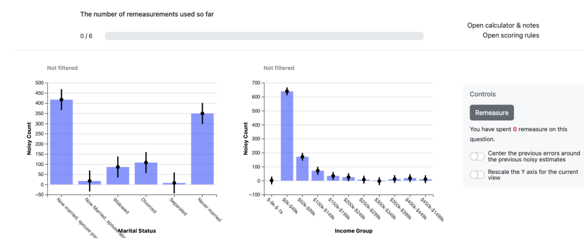
DP Creator from OpenDP (Sarathy et al. 2023)



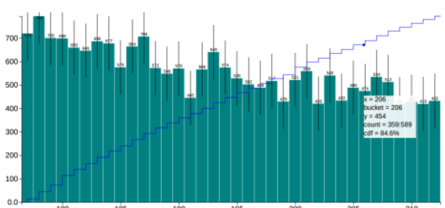
ViP (Nanayakkara et al. 2022)



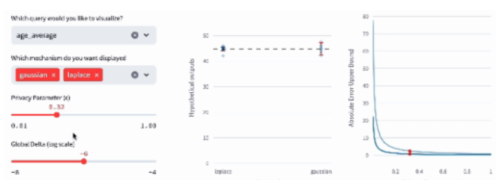
DPP (St. John et al. 2021)



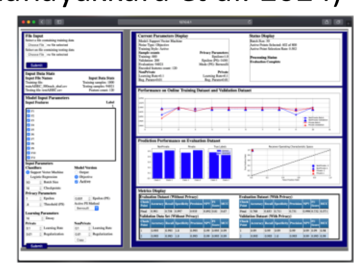
Measure-Observe-Remeasure (Nanayakkara et al. 2024)



Overlook (Thaker et al. 2022)

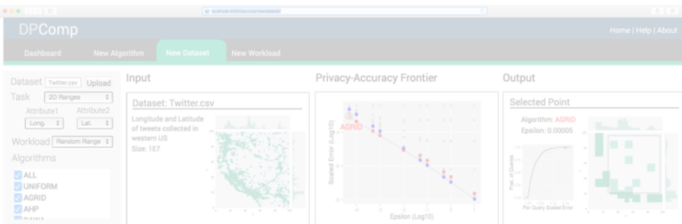


Panavas et al. 2024

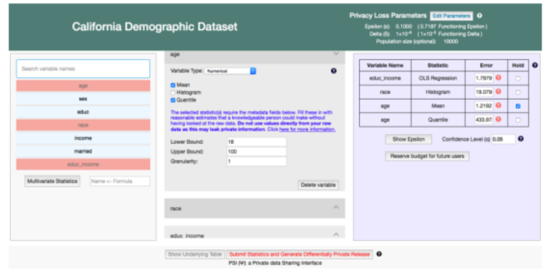


Bittner et al. 2020

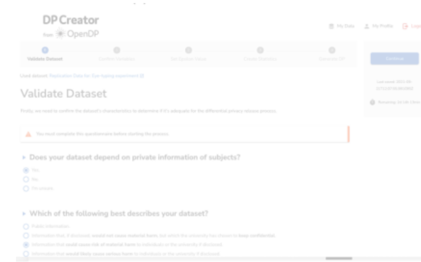
Several interfaces have been proposed for DP



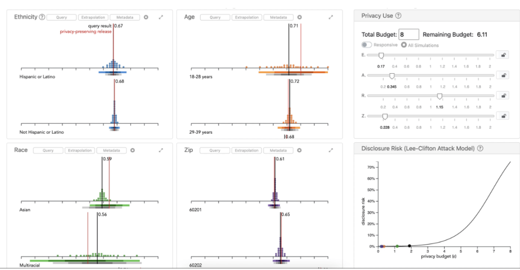
DP Comp (Hay et al. 2016)



PSI (Gaboardi et al. 2018)



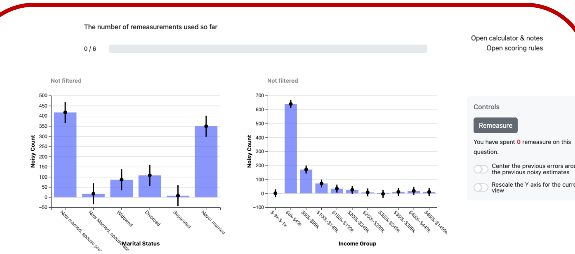
DP Creator from OpenDP (Sarathy et al. 2023)



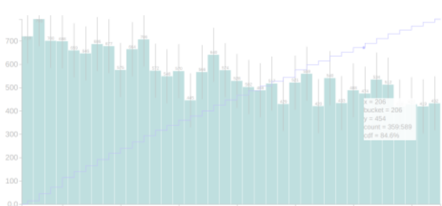
ViP (Nanayakkara et al. 2022)

today

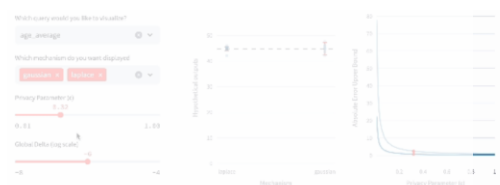
DPP (St. John et al. 2021)



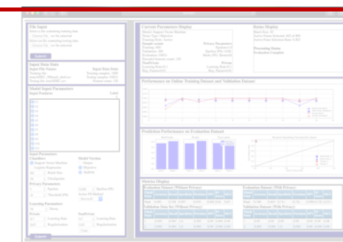
Measure-Observe-Remeasure (Nanayakkara et al. 2024)



Overlook (Thaker et al. 2022)

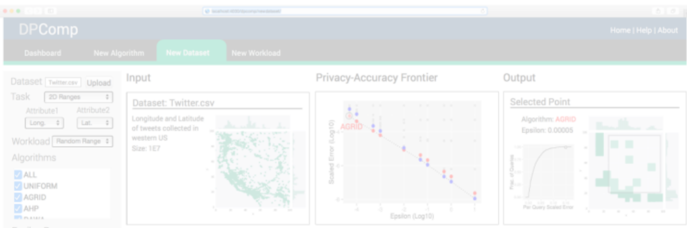


Panavas et al. 2024

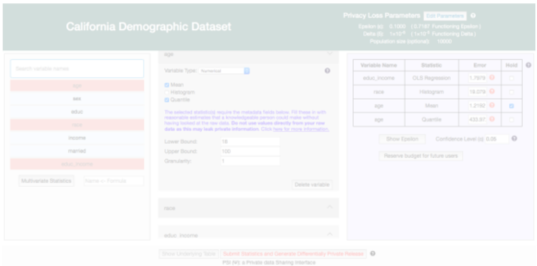


Bittner et al. 2020

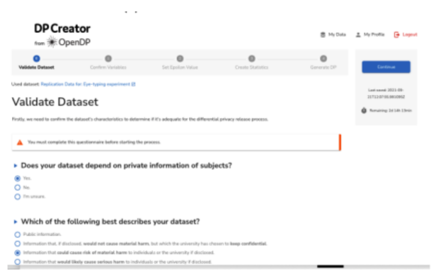
Several interfaces have been proposed for DP



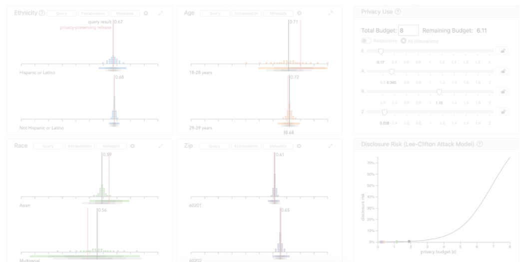
DP Comp (Hay et al. 2016)



PSI (Gaboardi et al. 2018)



DP Creator from OpenDP
(Sarathy et al. 2023)

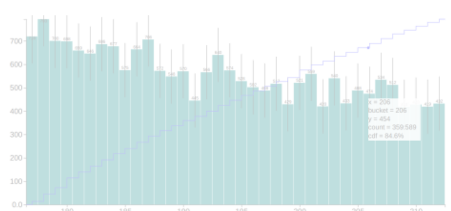


ViP (Nanayakkara et al. 2022)

DPP (St. John et al. 2021)

today

Measure-Observe-Remeasure
(Nanayakkara et al. 2024)



Overlook (Thaker et al. 2022)

Panavas et al. 2024

Bittner et al. 2020

Today’s goals

- Become familiar with the space of DP interfaces for data analysts & data curators/depositors
- Learn multiple styles of user studies and the types of questions they enable us to answer

PSI (Ψ): a Private data Sharing Interface*

(WORKING PAPER)

Marco Gaboardi† James Honaker ‡ Gary King § Jack Murtagh¶

Kobbi Nissim‖ Jonathan Ullman** Salil Vadhan††

with contributions from

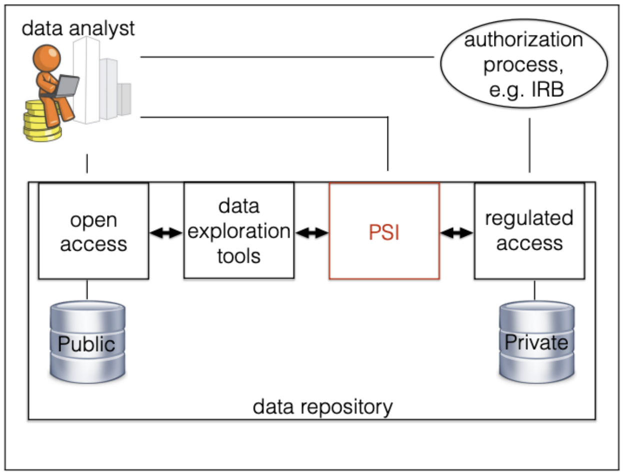
Nabib Ahmed, Andreea Antuca, Brendan Avent, Jordan Awan, Christian Baehr, Connor Bain, Victor Balcer, Thomas Brawner, Jessica Bu, Mark Bun, Stephen Chong, Fanny Chow, Katie Clayton, Holly Cunningham, Vito D'Orazio, Gian Pietro Farina, Anna Gavrilman, Benjamin Glass, Caper Gooden, Paul Handorff, Raquel Hill, Alyssa Hu, Jason Huang, Justin Kaashoek, Allyson Kaminsky, Chan Kang, Murat Kuntarcioglu, Vishesh Karwa, George Kellaris, Michael Lackner, Jack Landry, Hyun Woo Lim, Giovanni Malloy, Michael Lopiccolo, Nathan Manohar, Ross Mawhorter, Dan Muise, Marcelo Novaes, Ana Luisa Oaxaca, Raman Prasad, Sofya Raskhodnikova, Grace Rehaut, Ryan Rogers, Or Sheffet, Adam D. Smith, Thomas Steinke, Kathryn Taylor, Julia Vasile, Clara Wang, Haoqing Wang, Remy Wang, Lancelot Wathieu, David Xiao, Anton Xue, and Joy Zheng

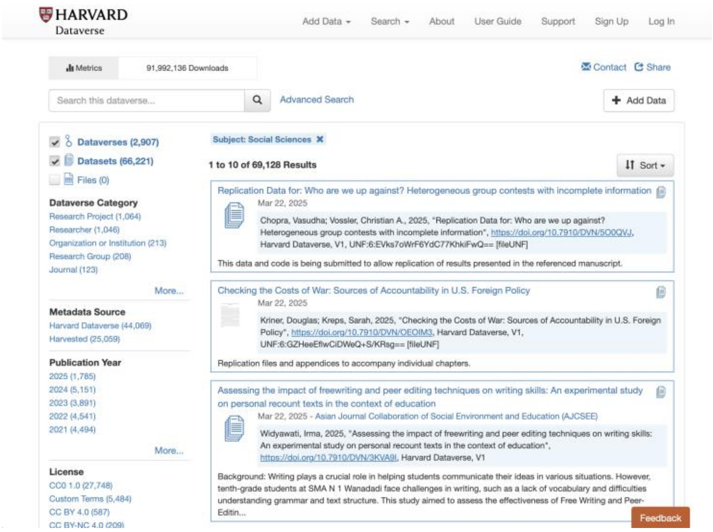
August 7, 2018

- System for sharing and exploring sensitive datasets for research use
- We'll focus on the interface for data analysts and (briefly) a usability study

Private data Sharing Interface (PSI)

- System to enable social science researchers to access sensitive datasets under DP
- Intended to reduce monetary and time costs of applying for regulated access to sensitive datasets
- Integrate with dataset repositories, i.e., Dataverse

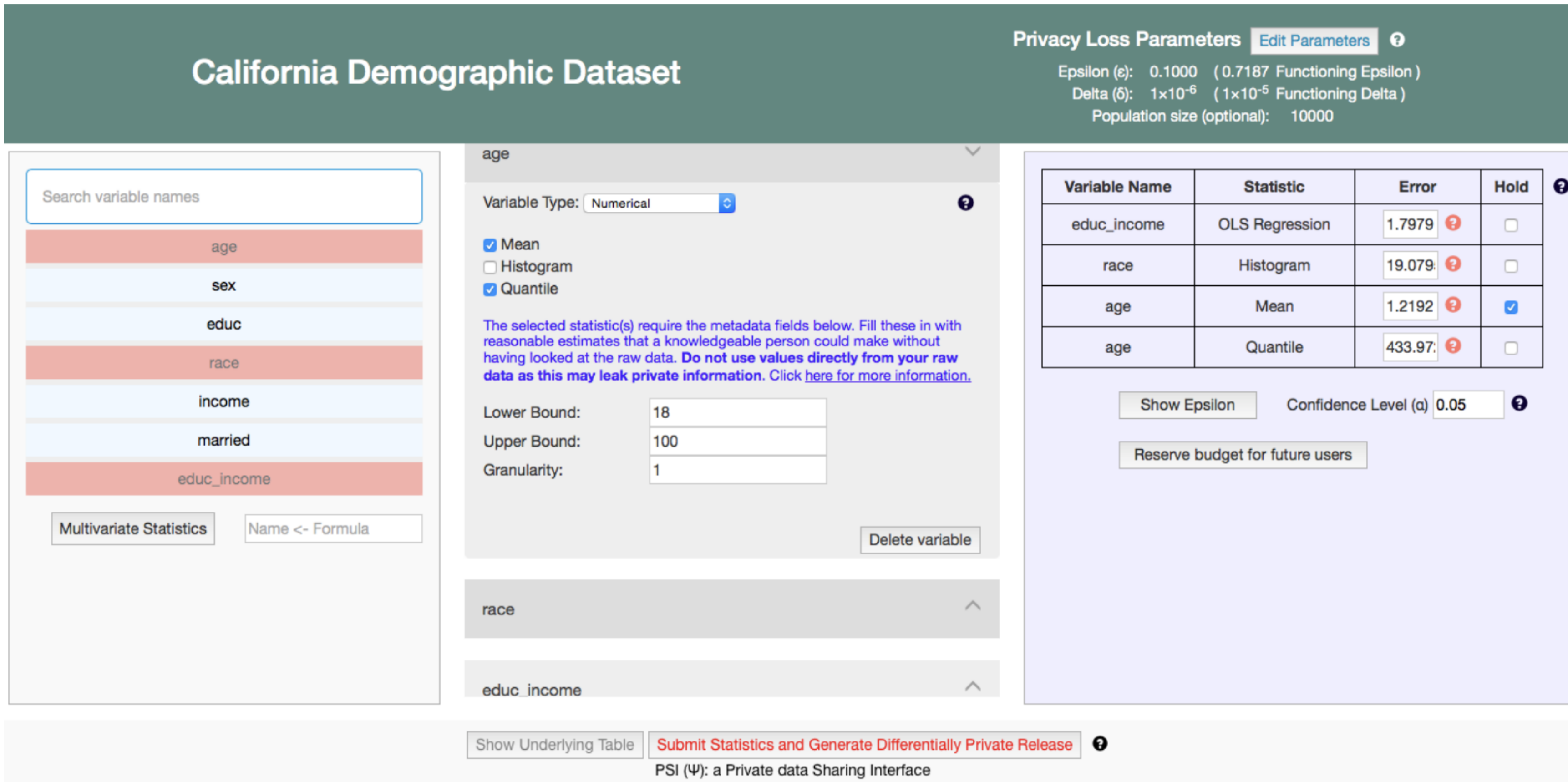




PSI design goals

- **Accessibility by non-experts**
  - System should be usable by social science researchers, without help from data privacy, CS, stats experts
- **Generality**
  - System should be applicable to and effective on heterogenous datasets
- **Workflow-compatibility**
  - System should fit naturally into the workflow of its user and offer more clear benefits (e.g., access to sensitive data, protect privacy) than impediments

PSI prototype



Usability study

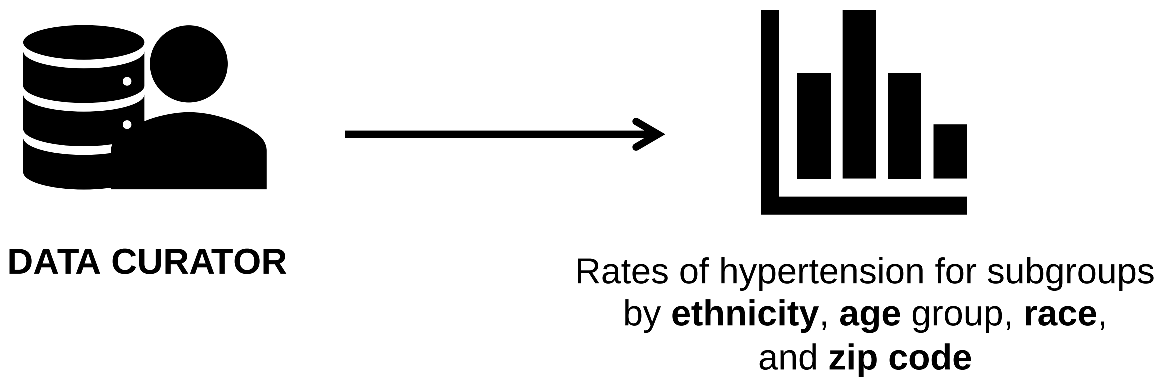
- 20 participants with some data analysis experience
  - 10% some college, 25% bachelor’s degree, 50% master’s degree, 15% PhD
  - 85% unfamiliar or somewhat familiar with DP
- Given a toy dataset with demographic information of 1,000 people and told to “advertise” the dataset to social scientists interested in the relationship between race and income across ages
  - First, asked to set privacy loss parameters
  - Second, completed 11 tasks varying in terms of task generality:
    - “You no longer wish to include a quantile for income. Delete this statistic.”
    - “Make it so that the released mean age is off from its true mean by at most one year. Is this more or less accurate than what you had before?”

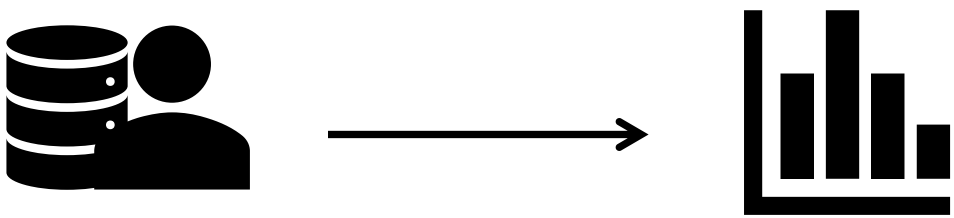
sciendo Forthcoming in Proceedings on Privacy Enhancing Technologies 2022; 2022 (2):1–18

Priyanka Nanayakkara*, Johes Bater, Xi He, Jessica Hullman, and Jennie Rogers

Visualizing Privacy-Utility Trade-Offs in Differentially Private Data Releases

- Interactive visualization interface (ViP) for data curators setting privacy loss budgets
- User study evaluating practitioners' ability to complete tasks related to setting privacy loss budgets according to their values with ViP vs. a control spreadsheet

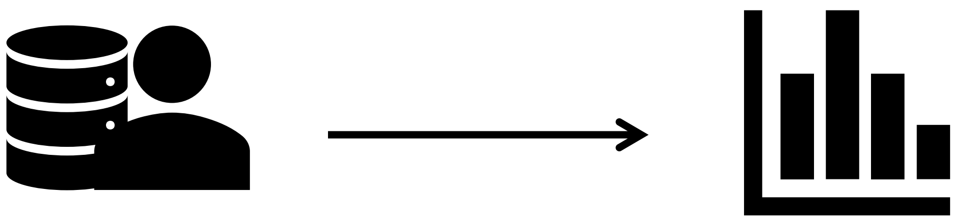

DATA CURATOR
Rates of hypertension for subgroups
by **ethnicity**, **age** group, **race**,
and **zip code**



DATA CURATOR

Rates of hypertension for subgroups
by **ethnicity**, **age** group, **race**,
and **zip code**

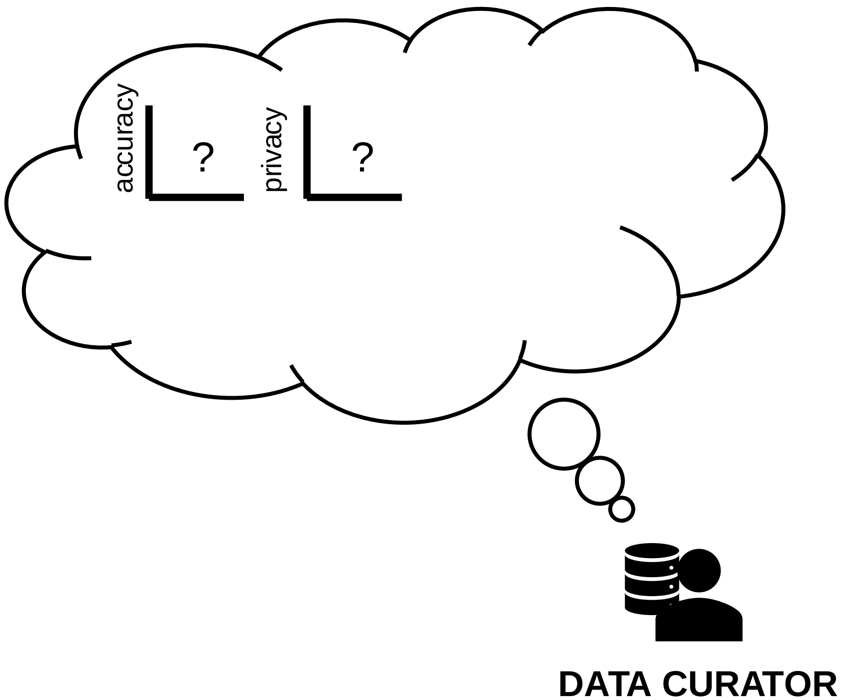
CIs for the population proportions

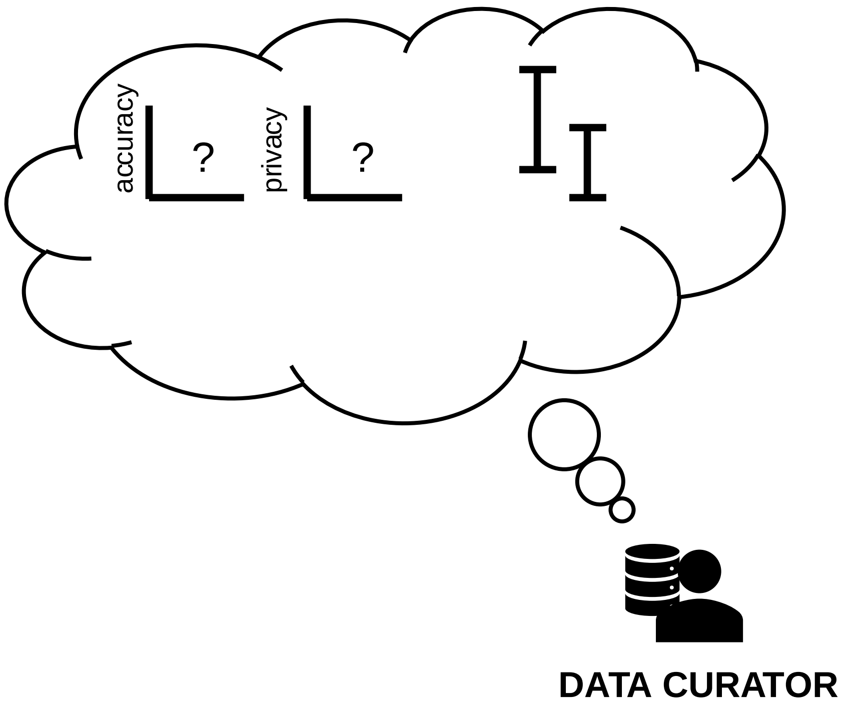


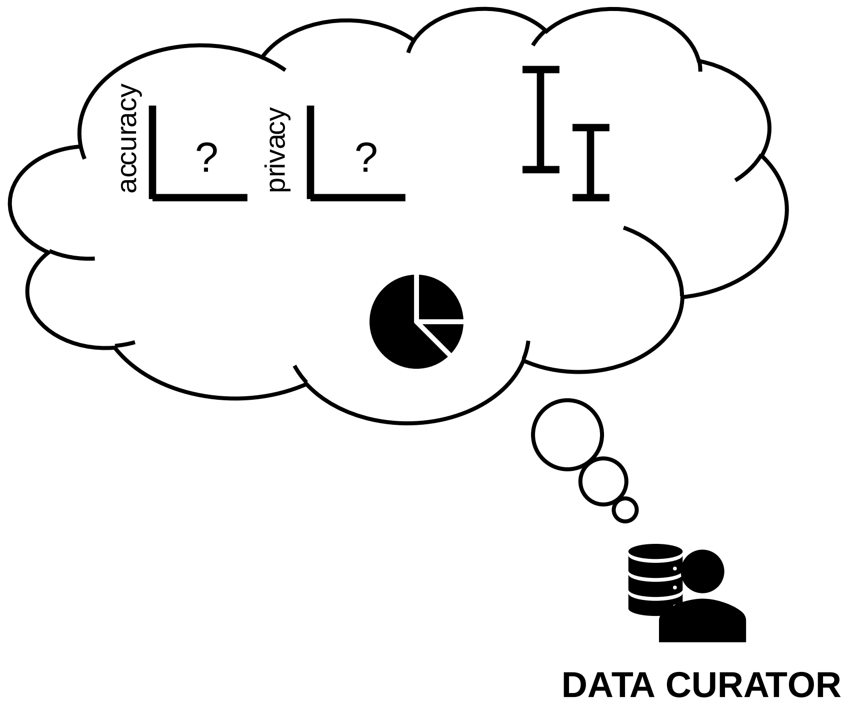
DATA CURATOR

Noisy estimates of rates of hypertension for subgroups by **ethnicity**, **age** group, **race**, and **zip code**

CIs for the population proportions


accuracy
?
privacy
?
DATA CURATOR


accuracy
?
privacy
?
DATA CURATOR


accuracy
?
privacy
?
DATA CURATOR

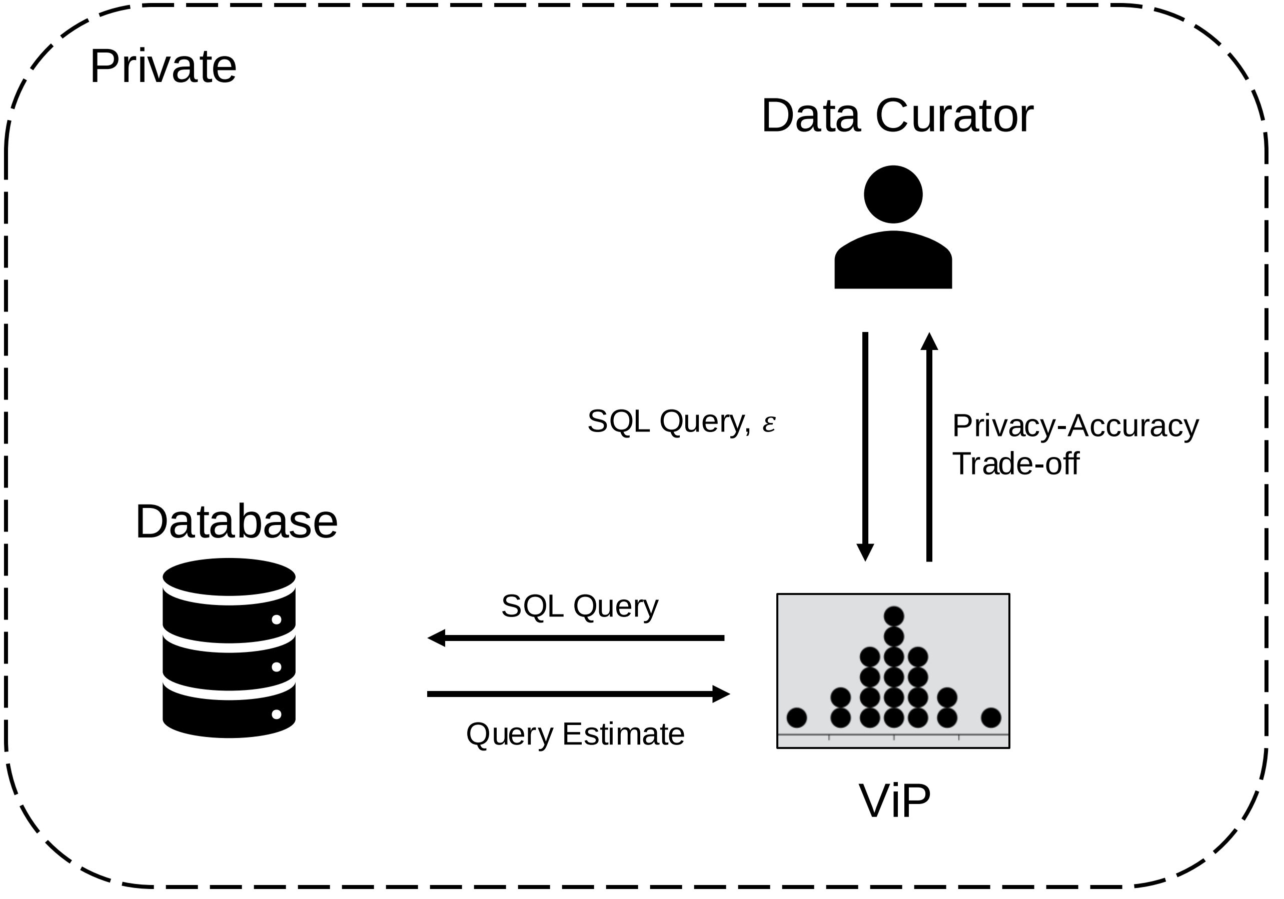
Design goals

Help a data curator understand

- the ε-accuracy relationship
- the ε-privacy relationship
- statistical inference under DP
- budget splitting across queries

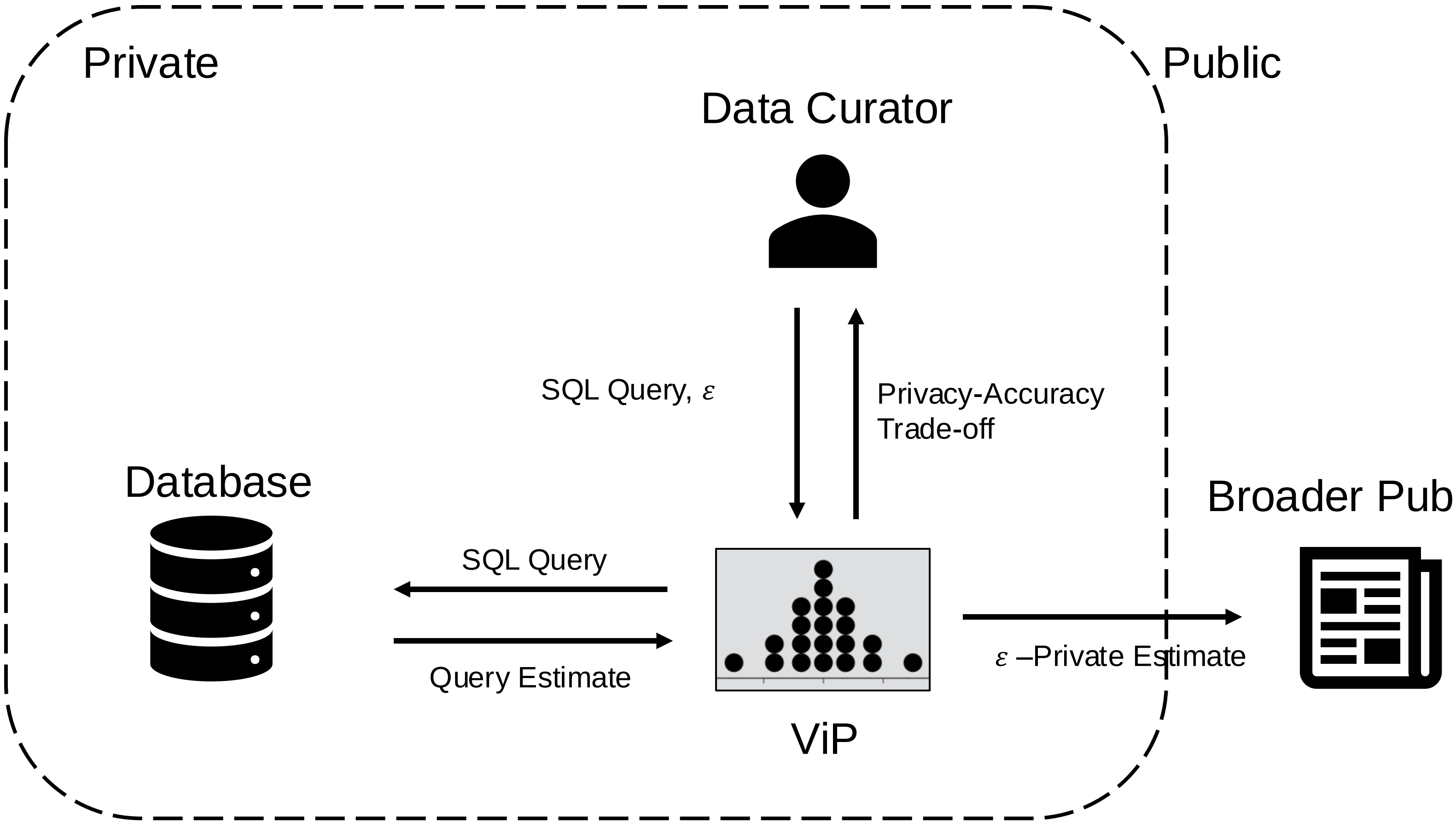
so that they can make more informed ε decisions

Workflow



Private
Data Curator
SQL Query, ε
Privacy-Accuracy
Trade-off
Database
SQL Query
Query Estimate
ViP

Workflow

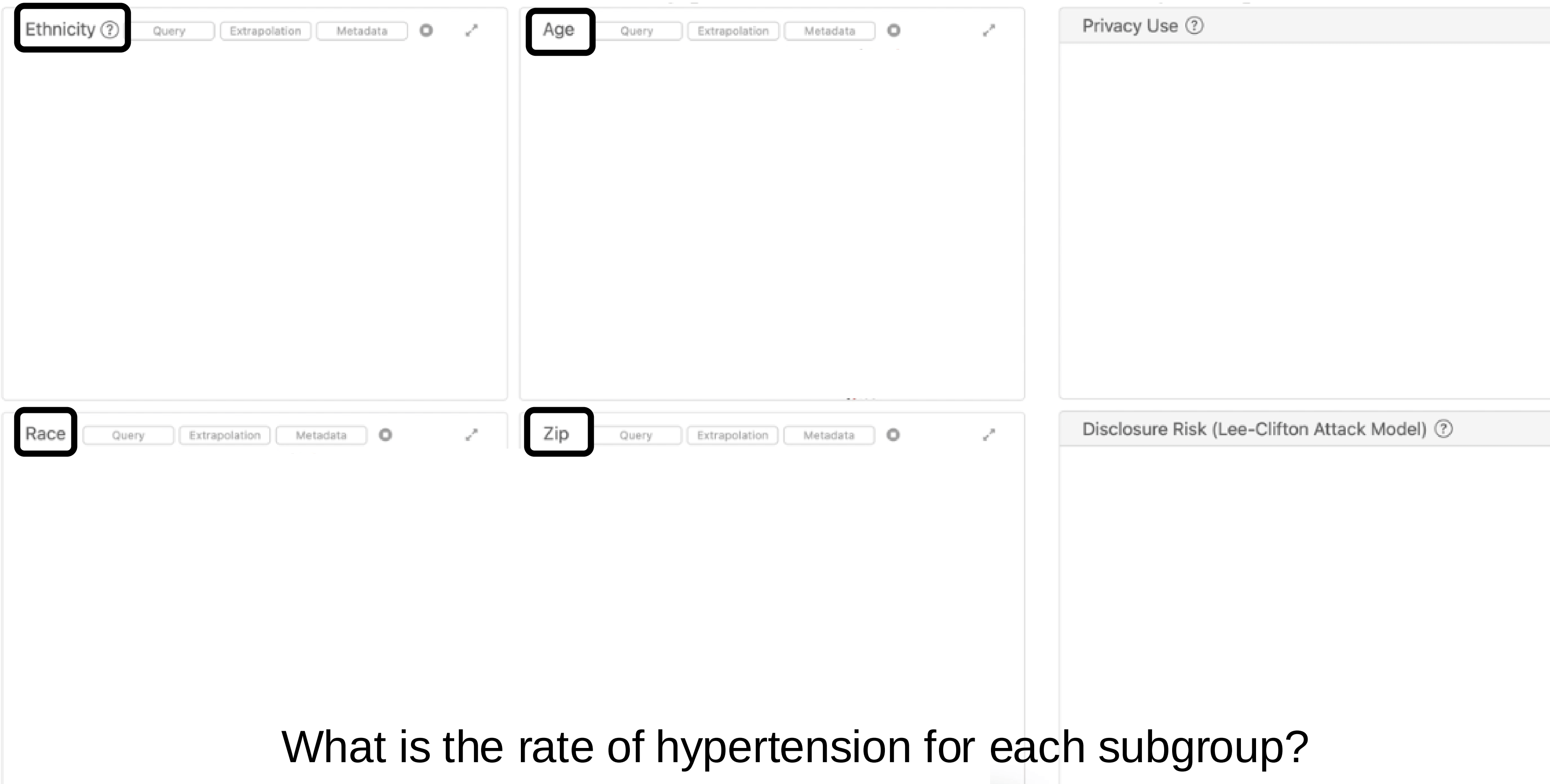


Private
Public
Data Curator
SQL Query, ε
Privacy-Accuracy
Trade-off
Database
SQL Query
Query Estimate
ViP
Broader Public
ε –Private Estimate

What about privacy loss from testing multiple ε's?

- Instead consider having data curators set parameters based on publicly available data that we expect to be distributionally similar to the sensitive dataset of interest
- Alternatively, consider adapting the interface based on private selection from private candidates (see HoDP pp 228 - last week's reading)

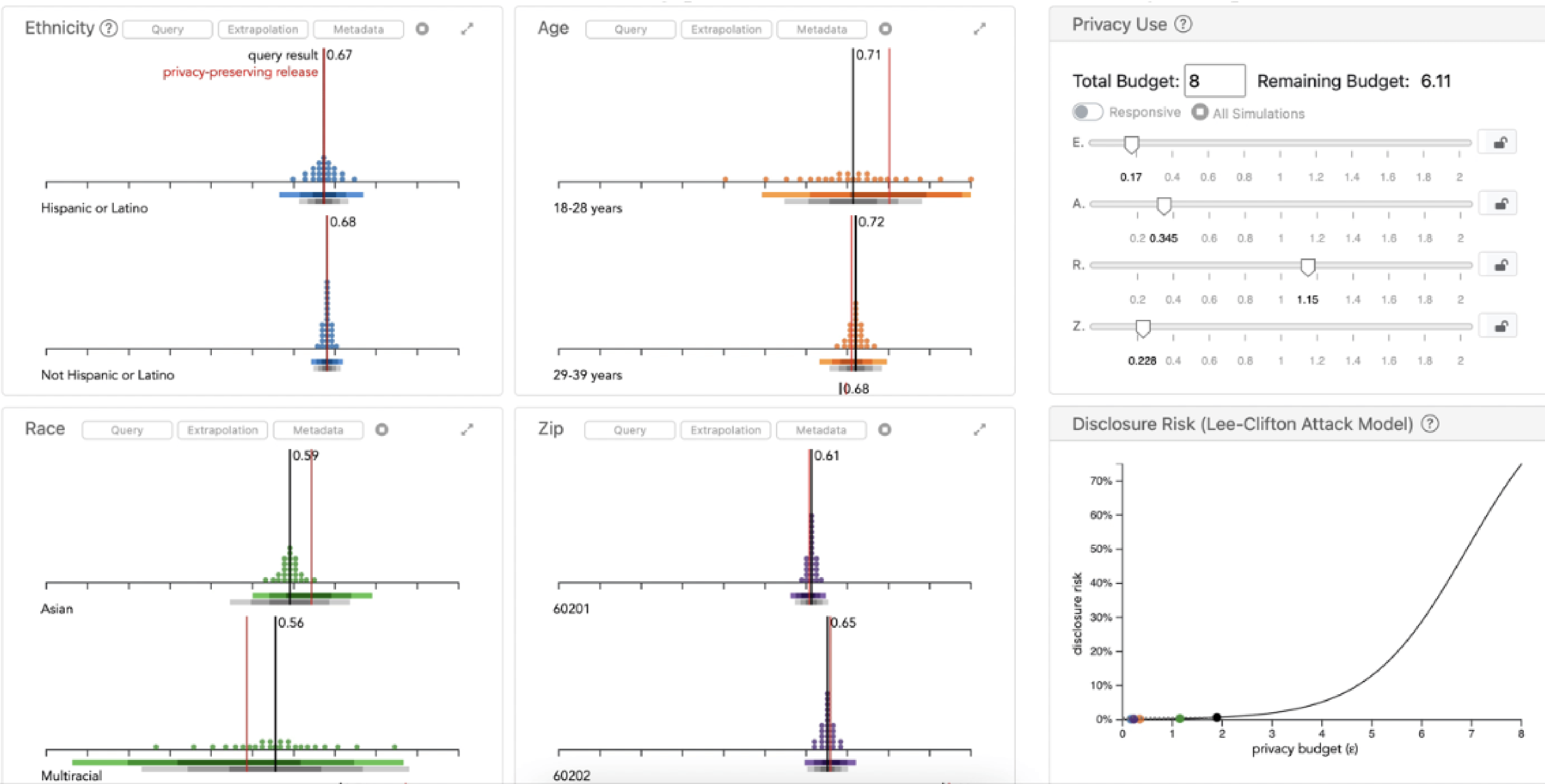
Visualizing Privacy (ViP)



What is the rate of hypertension for each subgroup?

https://priyakalot.github.io/ViP-demo

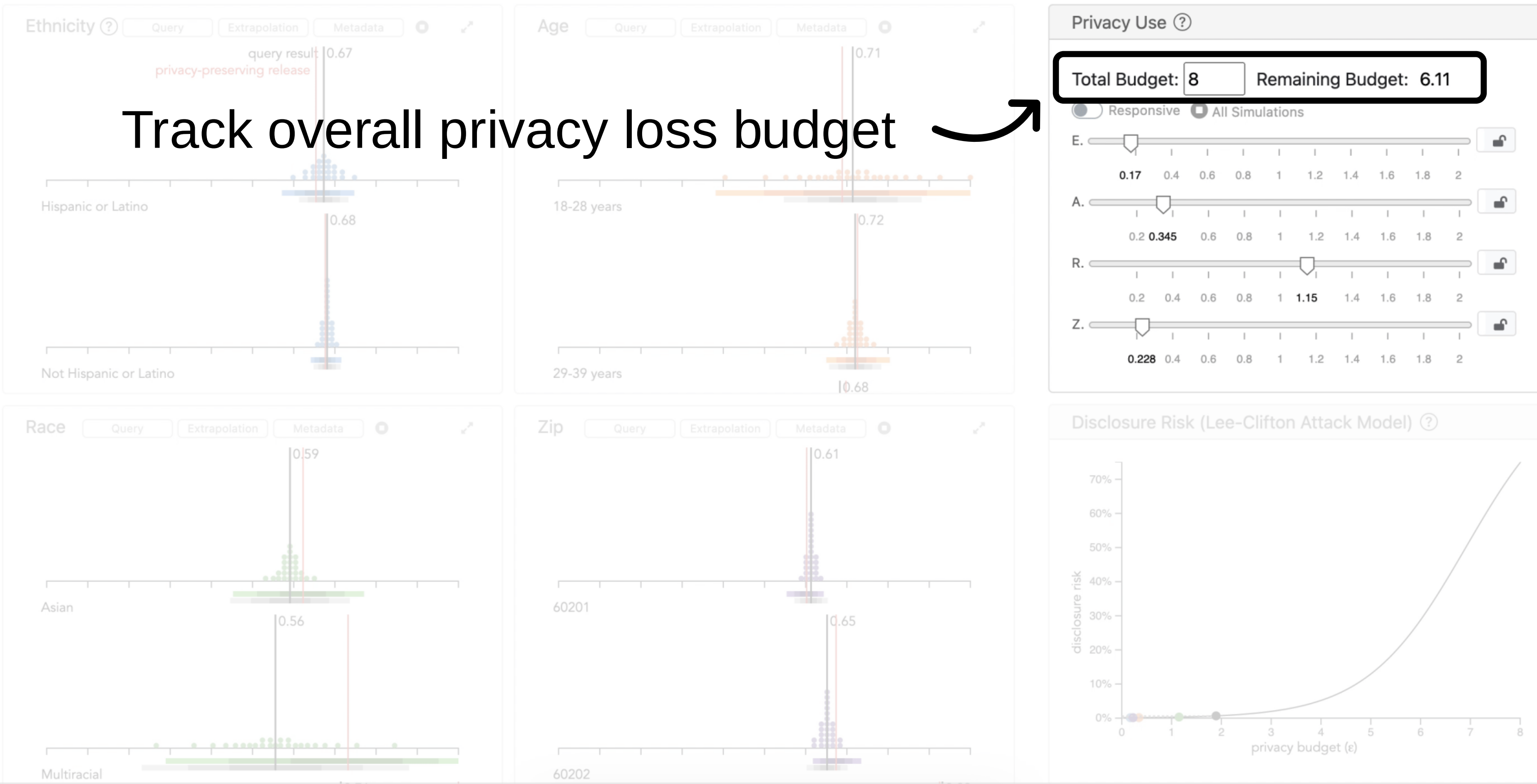
Visualizing Privacy (ViP)



https://priyakalot.github.io/ViP-demo

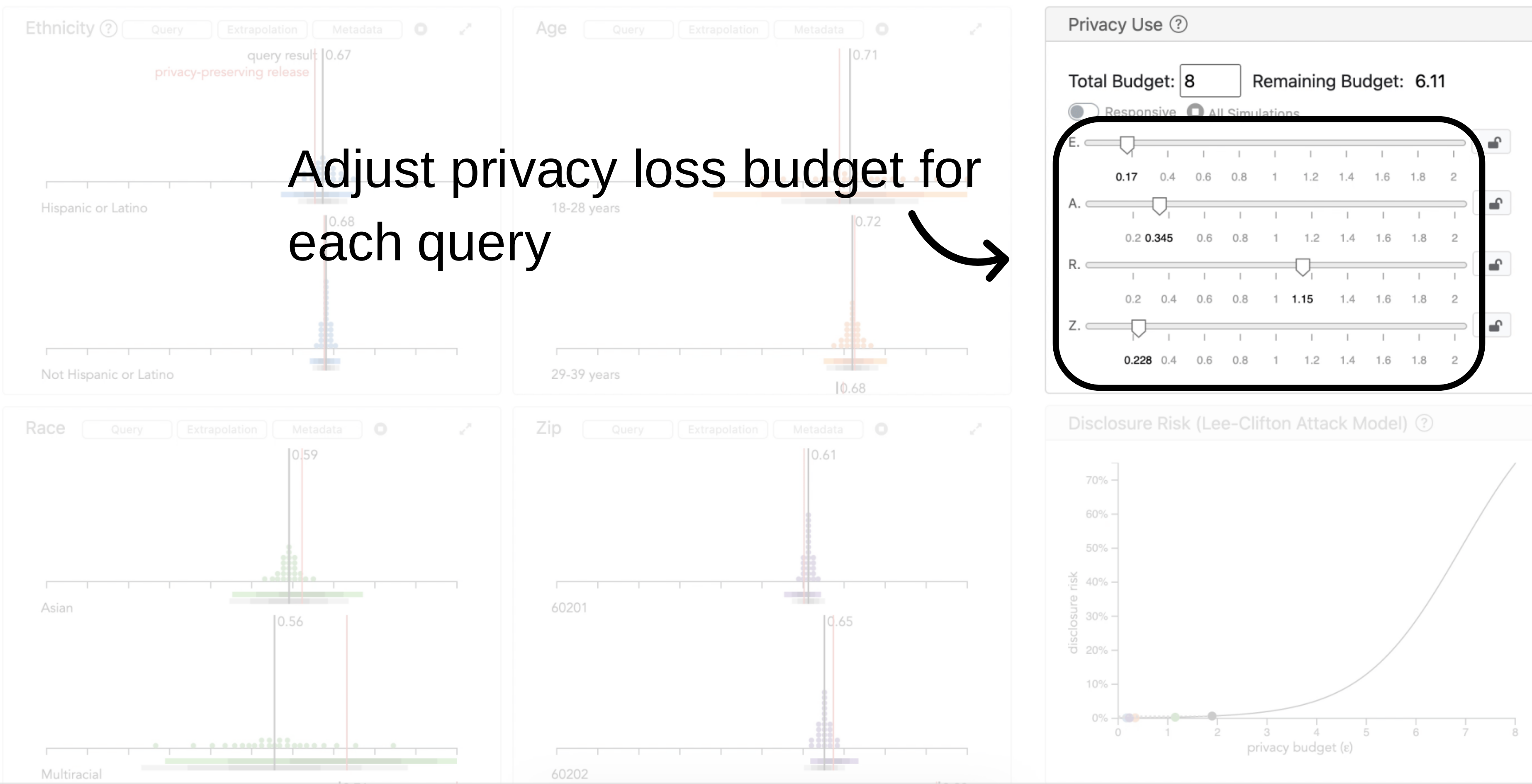
Visualizing Privacy (ViP)

Track overall privacy loss budget



https://priyakalot.github.io/ViP-demo

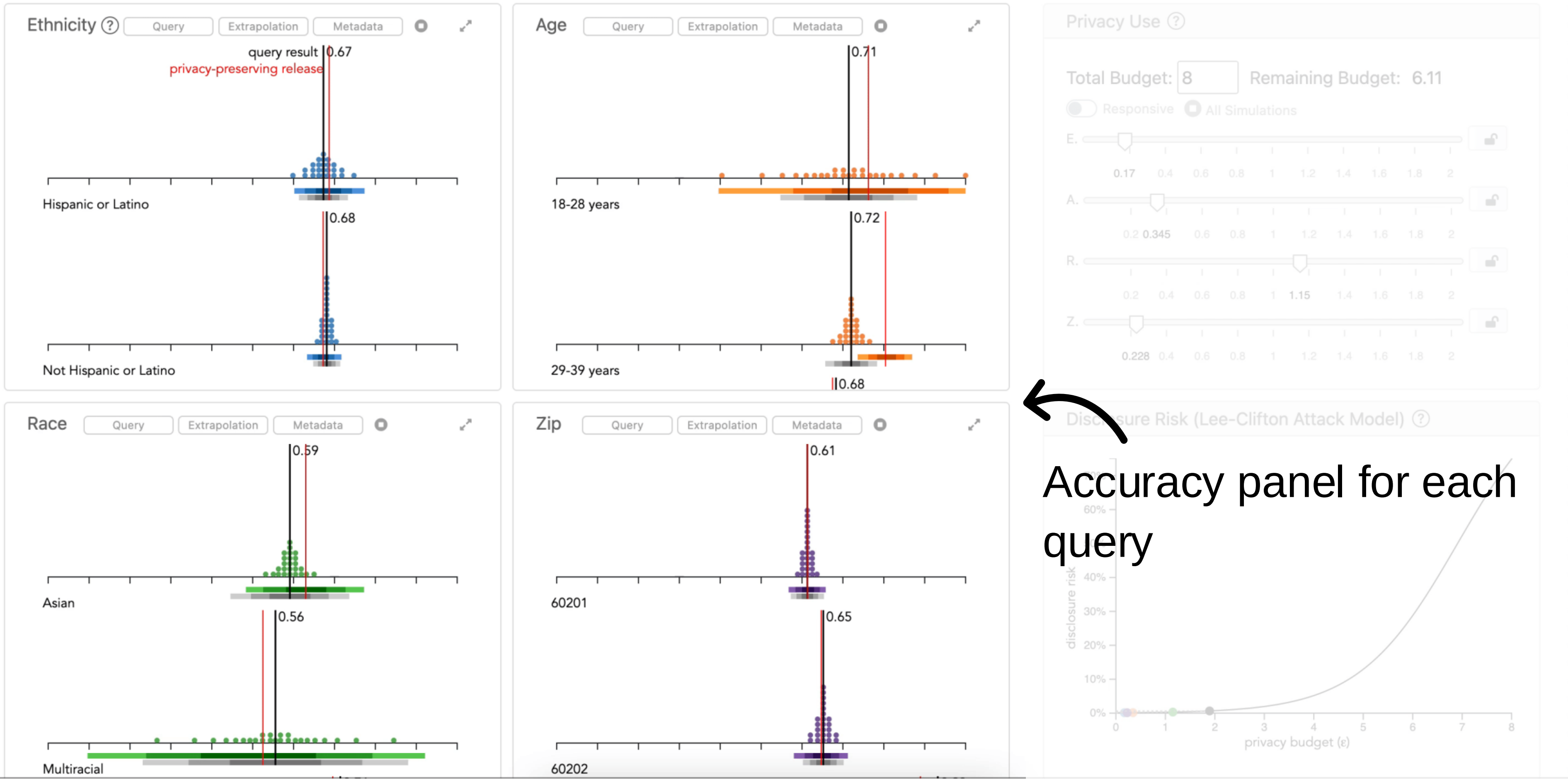
Visualizing Privacy (ViP)



Adjust privacy loss budget for each query

https://priyakalot.github.io/ViP-demo

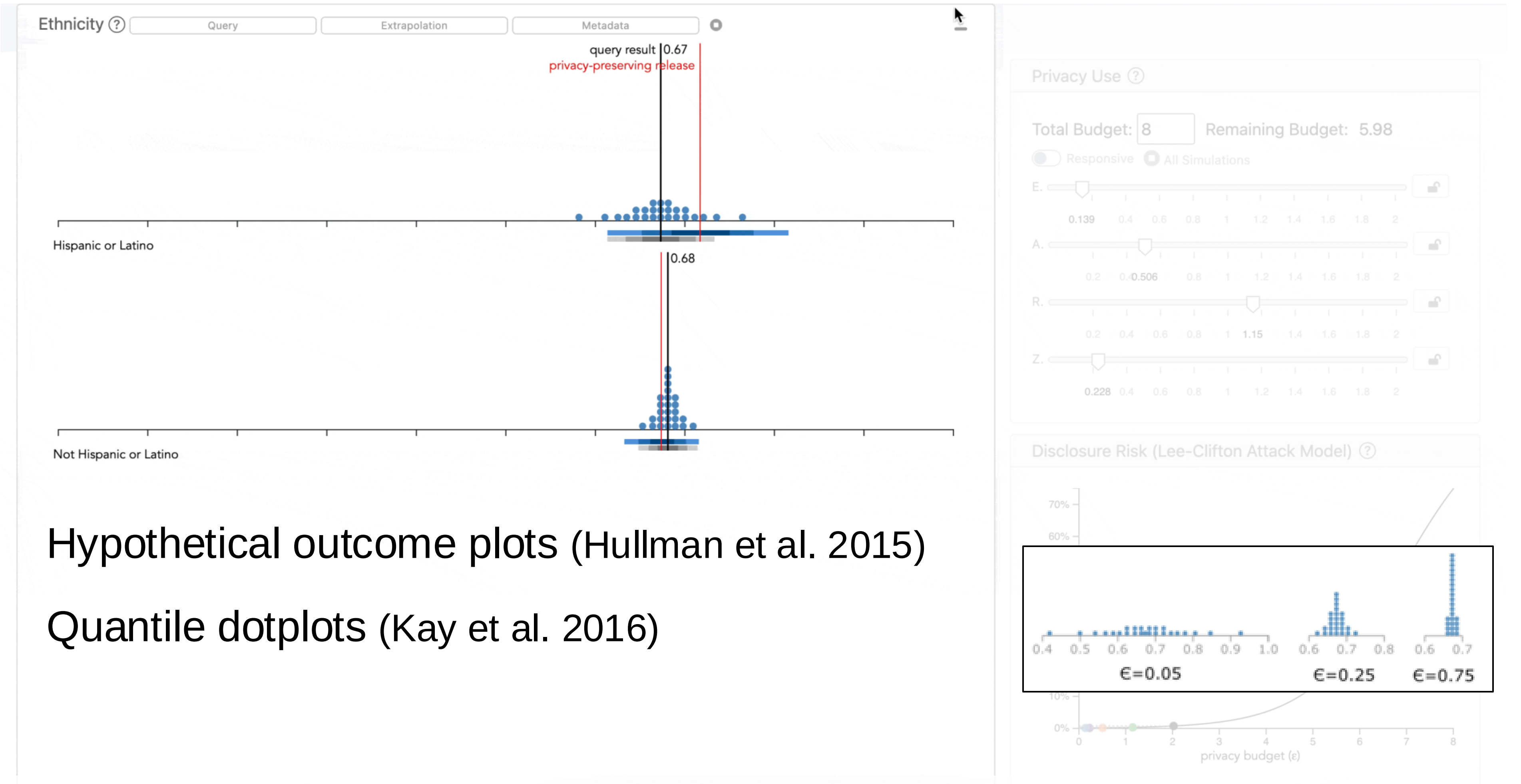
Visualizing Privacy (ViP)



Accuracy panel for each query

https://priyakalot.github.io/ViP-demo

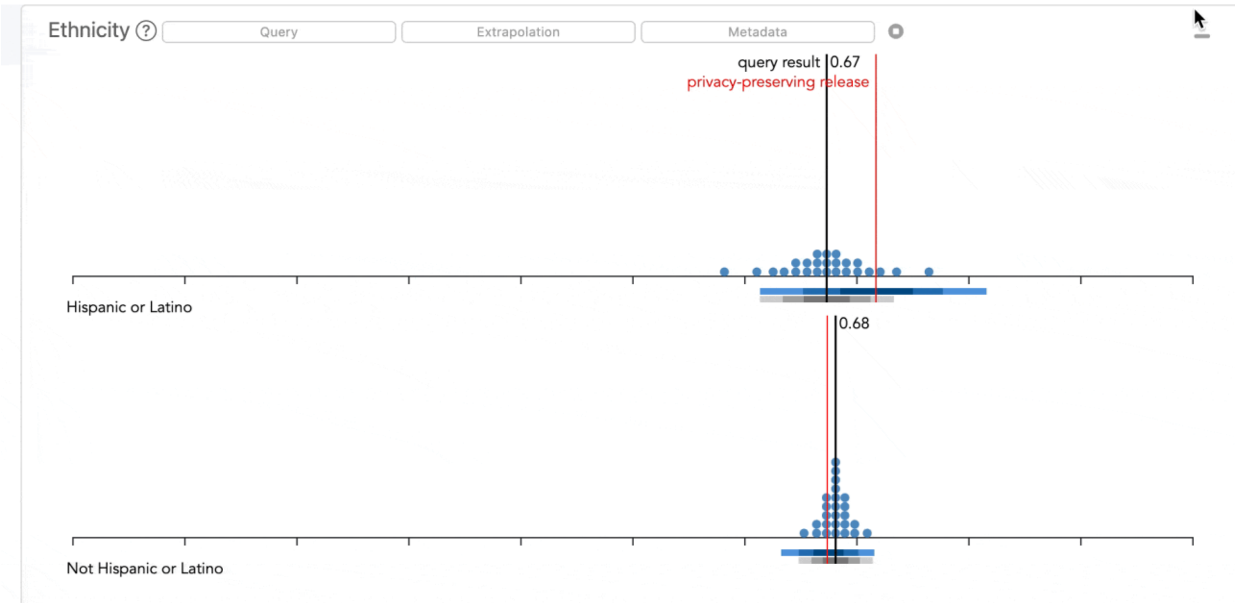
Visualizing Privacy (ViP)

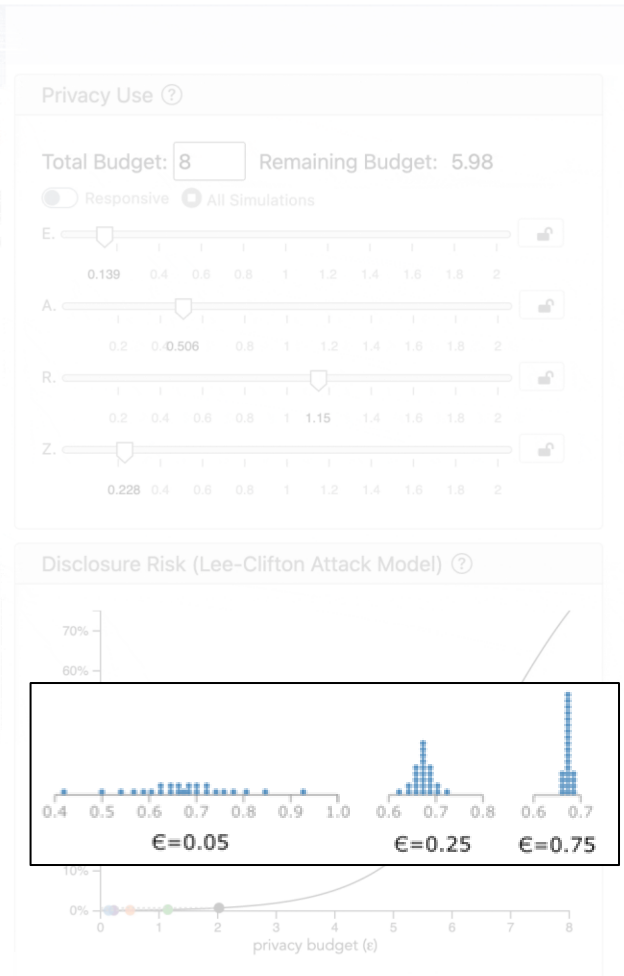


Hypothetical outcome plots (Hullman et al. 2015)

Quantile dotplots (Kay et al. 2016)

Visualizing Privacy (ViP)



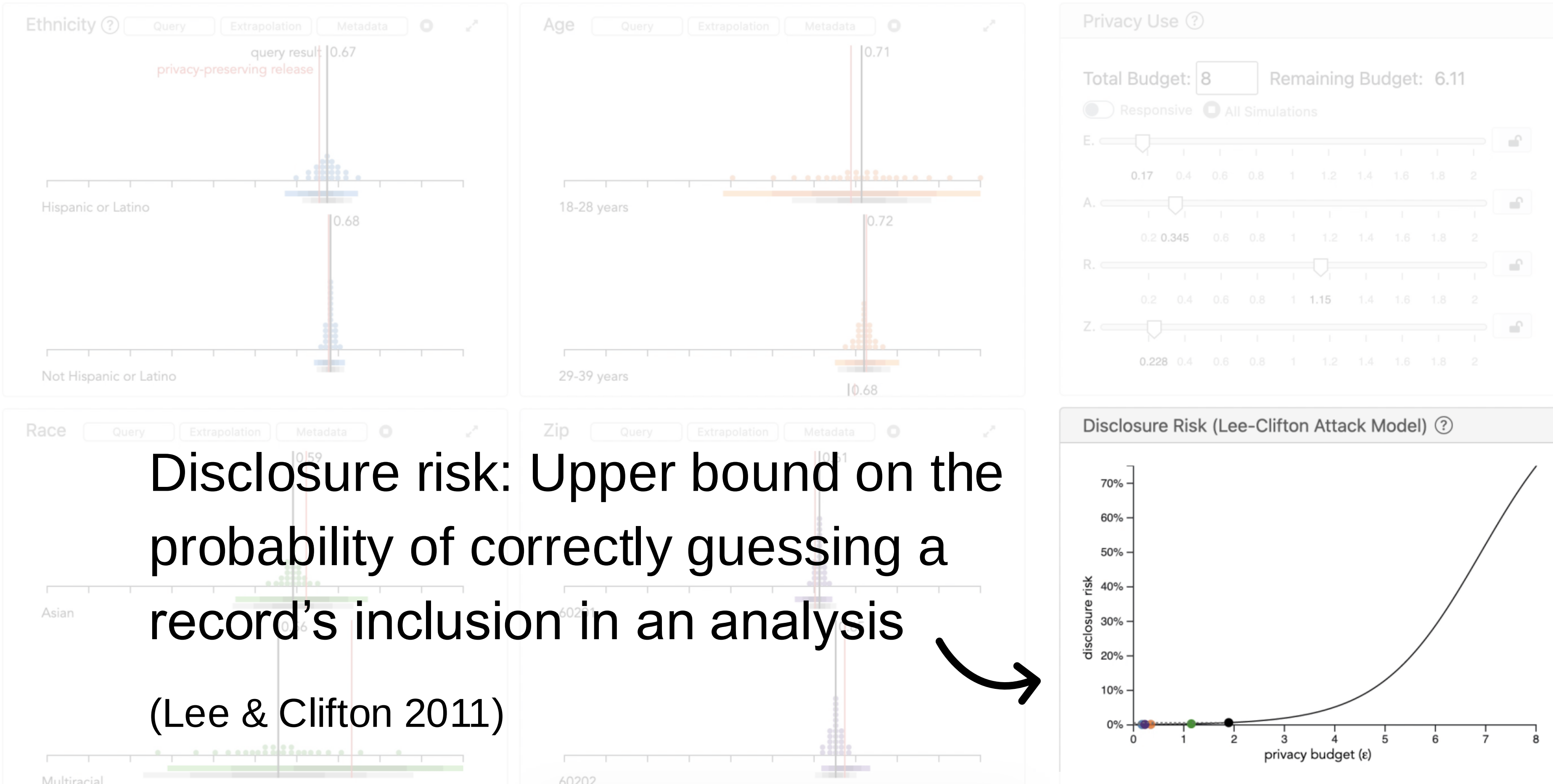


Hypothetical outcome plots (Hullman et al. 2015)

Quantile dotplots (Kay et al. 2016)

Differentially private confidence intervals (Ferrando et al. 2022)

Visualizing Privacy (ViP)



Disclosure risk: Upper bound on the probability of correctly guessing a record's inclusion in an analysis

(Lee & Clifton 2011)

https://priyakalot.github.io/ViP-demo

Evaluative user study

16 participants, experienced analyzing sensitive data but no differential privacy expertise

7 tasks using ViP and a control spreadsheet

Accuracy comparison

CDF judgment

Risk requirement

CI comparison

Equalize accuracy

Budget splitting

Prob. of superiority

Evaluative user study

At epsilon = 0.2, which subgroup in the Ethnicity query do we expect to have the most accurate privacy-preserving release?

Accuracy comparison | CDF judgment | Risk requirement | CI comparison | Equalize accuracy | Budget splitting | Prob. of superiority

Evaluative user study

At ε = *0.17* for the *Ethnicity* query, what is the probability that the privacy-preserving release for the *Hispanic or Latino* group will be greater than *0.7*?

Accuracy comparison | **CDF judgment** | Risk requirement | CI comparison | Equalize accuracy | Budget splitting | Prob. of superiority

Evaluative user study

For the Race query, what is value of ε that corresponds to re-identification risk of 0.2%.

Accuracy comparison | CDF judgment | **Risk requirement** | CI comparison | Equalize accuracy | Budget splitting | Prob. of superiority

Evaluative user study

Set ε for the *Ethnicity* query to 0.11. For the *Hispanic or Latino* group, estimate how many times wider we expect the privacy-preserving 95% CI to be compared to the traditional 95% CI.

Accuracy comparison | CDF judgment | Risk requirement | **CI comparison** | Equalize accuracy | Budget splitting | Prob. of superiority

Evaluative user study

Find the smallest values for ε for each query where the privacy-preserving releases for the *Female*, *Not Hispanic or Latino*, Asian, and *60202* subgroups are within 0.1 of the query result.

Accuracy comparison | CDF judgment | Risk requirement | CI comparison | **Equalize accuracy** | Budget splitting | Prob. of superiority

Evaluative user study

Suppose you have a total budget of $\varepsilon = 1.25$. Allocate your budget across queries such that the risk corresponding to each query is no more than 0.3% and the release is guaranteed to be within 0.1 of the un-noised query result for the *Male*, *Hispanic or Latino*, *Native Hawaiian or Other Pacific Islander*, and *60201 zip code* subgroups with roughly 90% probability.

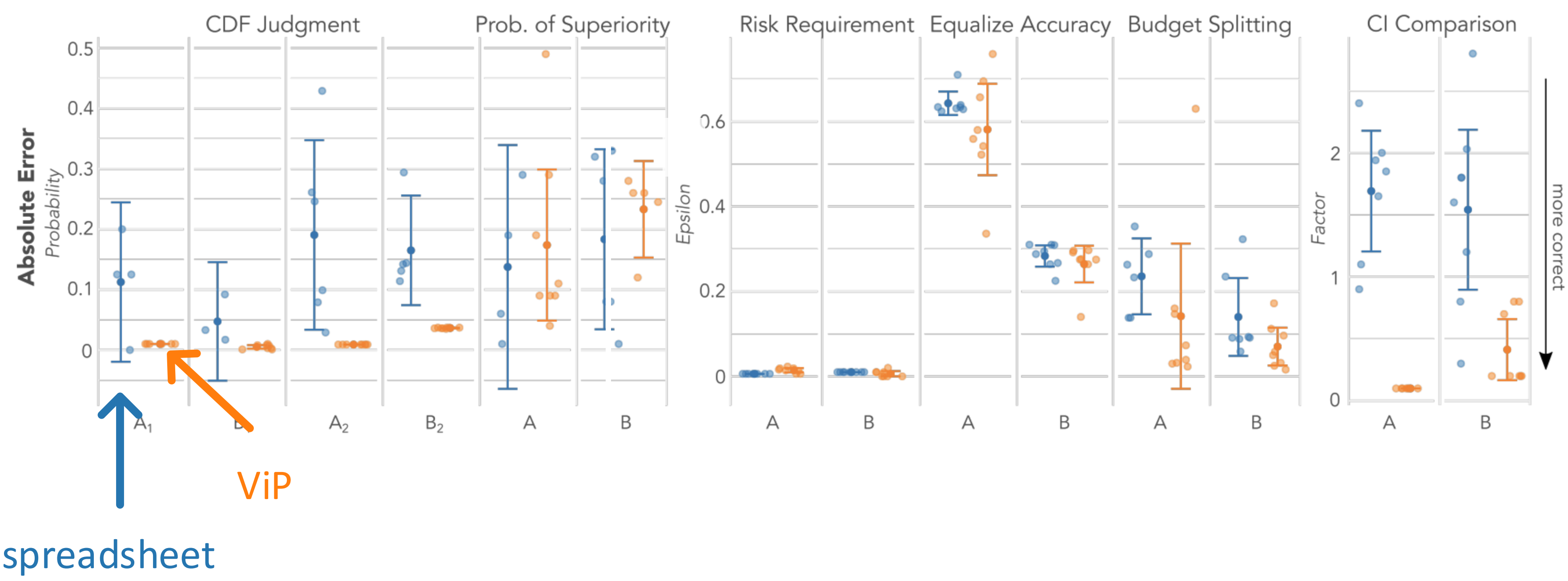
Accuracy comparison | CDF judgment | Risk requirement | CI comparison | Equalize accuracy | **Budget splitting** | Prob. of superiority

Evaluative user study

Set ε = 0.4 for the *Ethnicity* query. Estimate the probability that the privacy-preserving release for the *Hispanic or Latino* group will be greater than the privacy-preserving release for the *Not Hispanic or Latino* group.

Accuracy comparison | CDF judgment | Risk requirement | CI comparison | Equalize accuracy | Budget splitting | **Prob. of superiority**

Results



CDF Judgment
Prob. of Superiority
Risk Requirement
Equalize Accuracy
Budget Splitting
CI Comparison
Absolute Error
Probability
Epsilon
Factor
more correct
spreadsheet
ViP

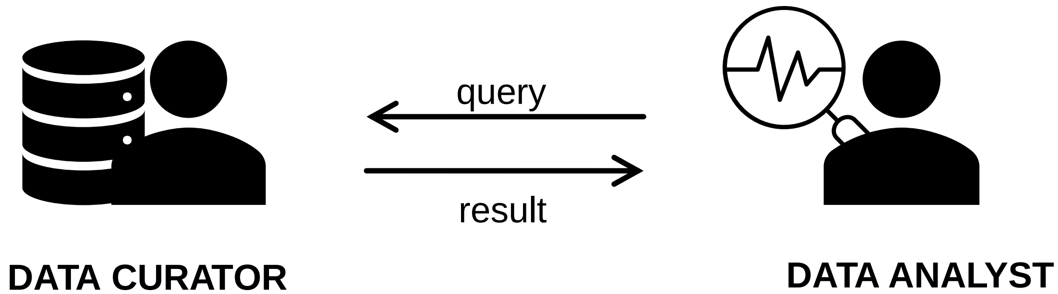
Conferences > 2024 IEEE Symposium on Securi...

Measure-Observe-Remeasure: An Interactive Paradigm for Differentially-Private Exploratory Analysis

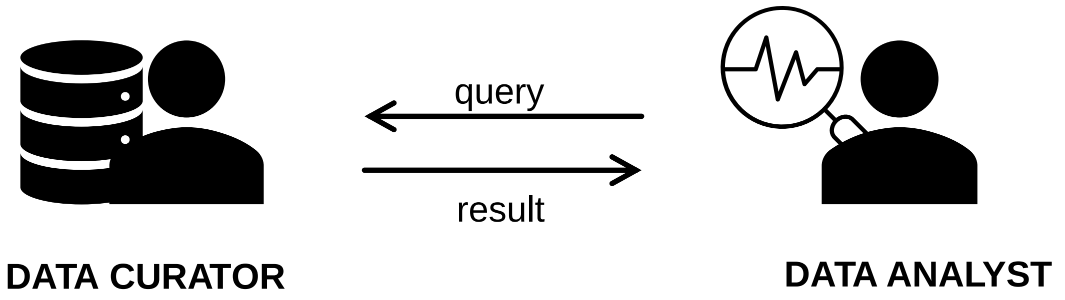
Publisher: IEEE Cite This PDF

Priyanka Nanayakkara ; Hyeok Kim ; Yifan Wu ; Ali Sarvghad ; Narges Mahyar ; Gerome Miklau **All Authors**

- Interactive paradigm, instantiated in an interactive visualization interface, for exploratory data analysis under DP
- Exploratory user study, within a decision-theoretic framework, on how analysts interact with the paradigm

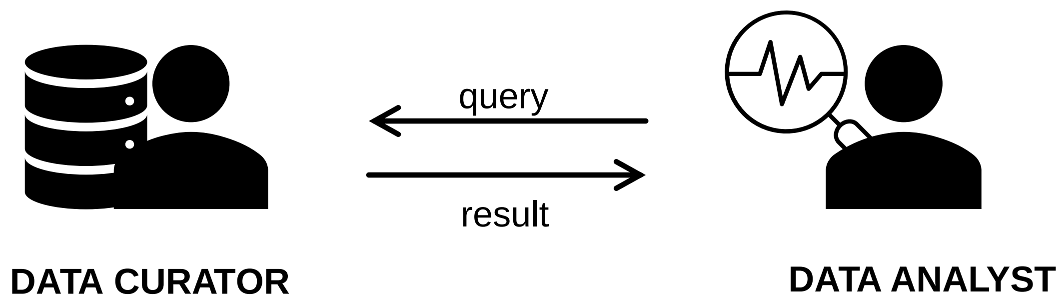


Analyst conducting an **exploratory** data analysis



Analyst conducting an **exploratory** data analysis

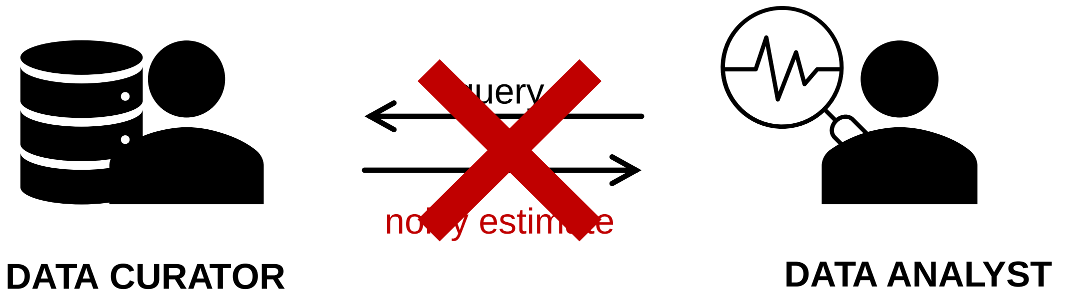
High-level analysis goals, but queries are developed and refined along the way



Analyst conducting an **exploratory** data analysis

High-level analysis goals, but queries are developed and refined along the way

Queries have varying levels of importance

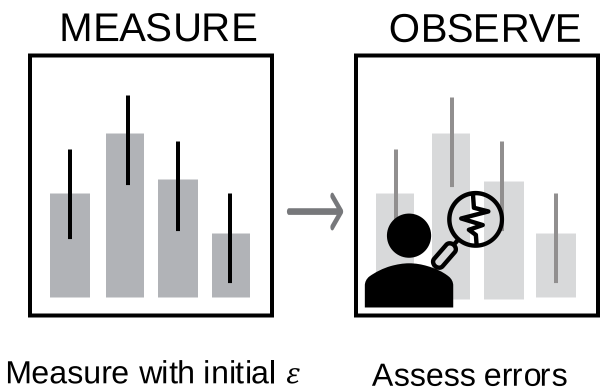


Analyst conducting an **exploratory** data analysis under DP

High-level analysis goals, but queries are developed and refined along the way

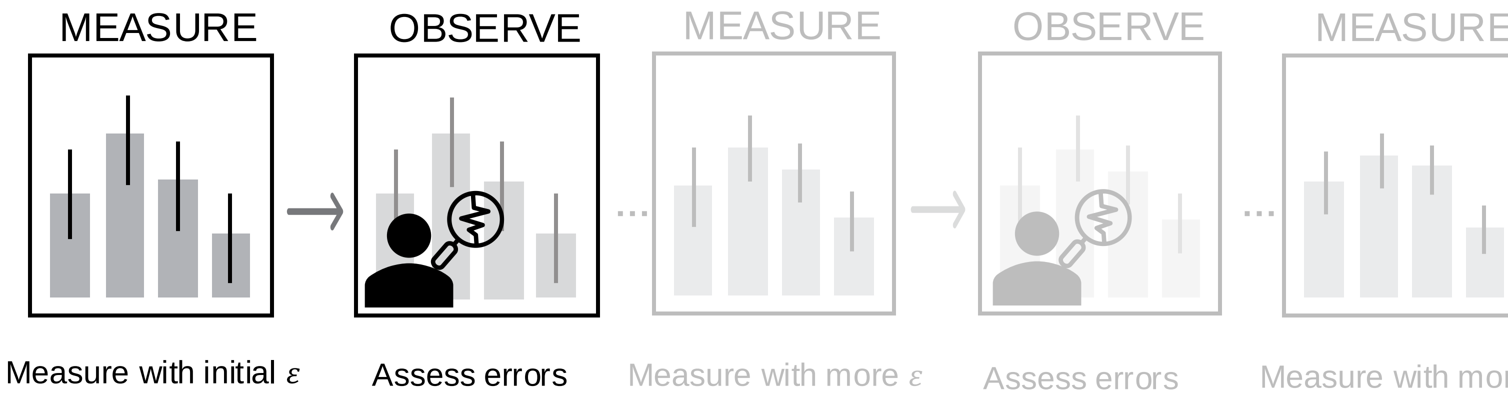
Queries have varying levels of importance

Measure-Observe
(Current Paradigm)



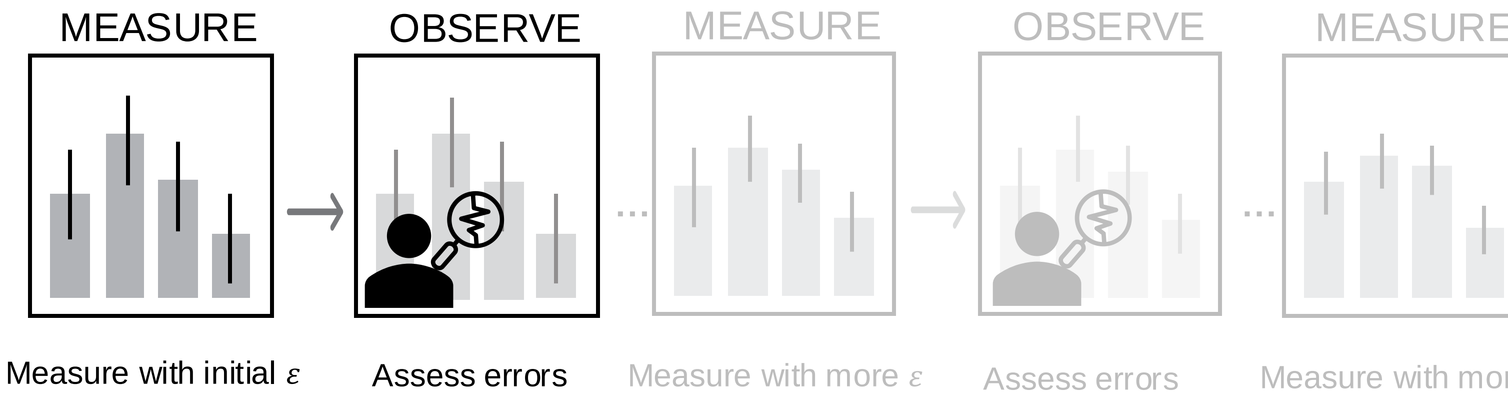
Measure-Observe
(Current Paradigm)

Measure-Observe
(Current Paradigm)



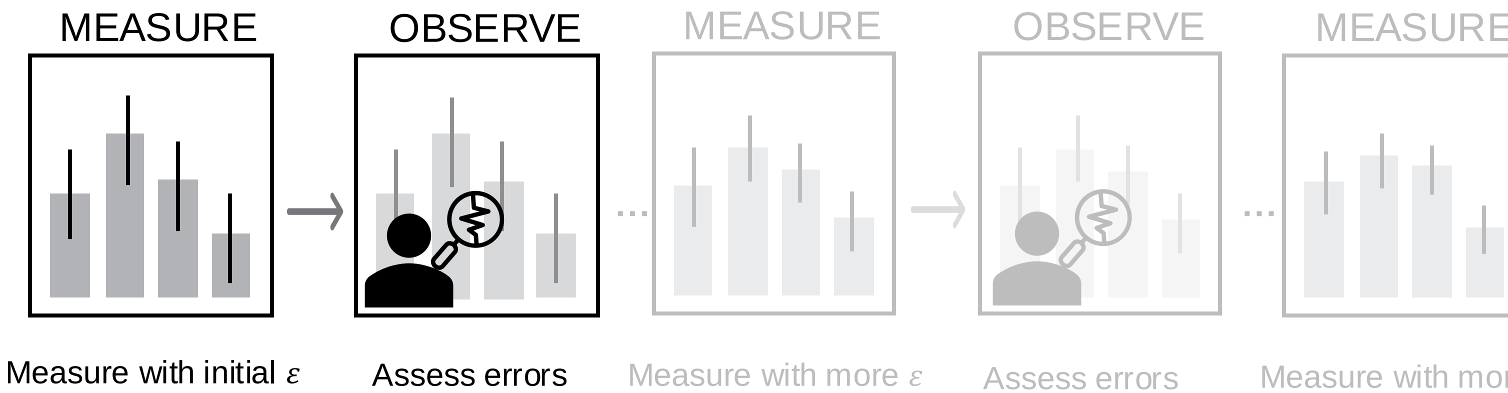
Measure-Observe

(Current Paradigm)

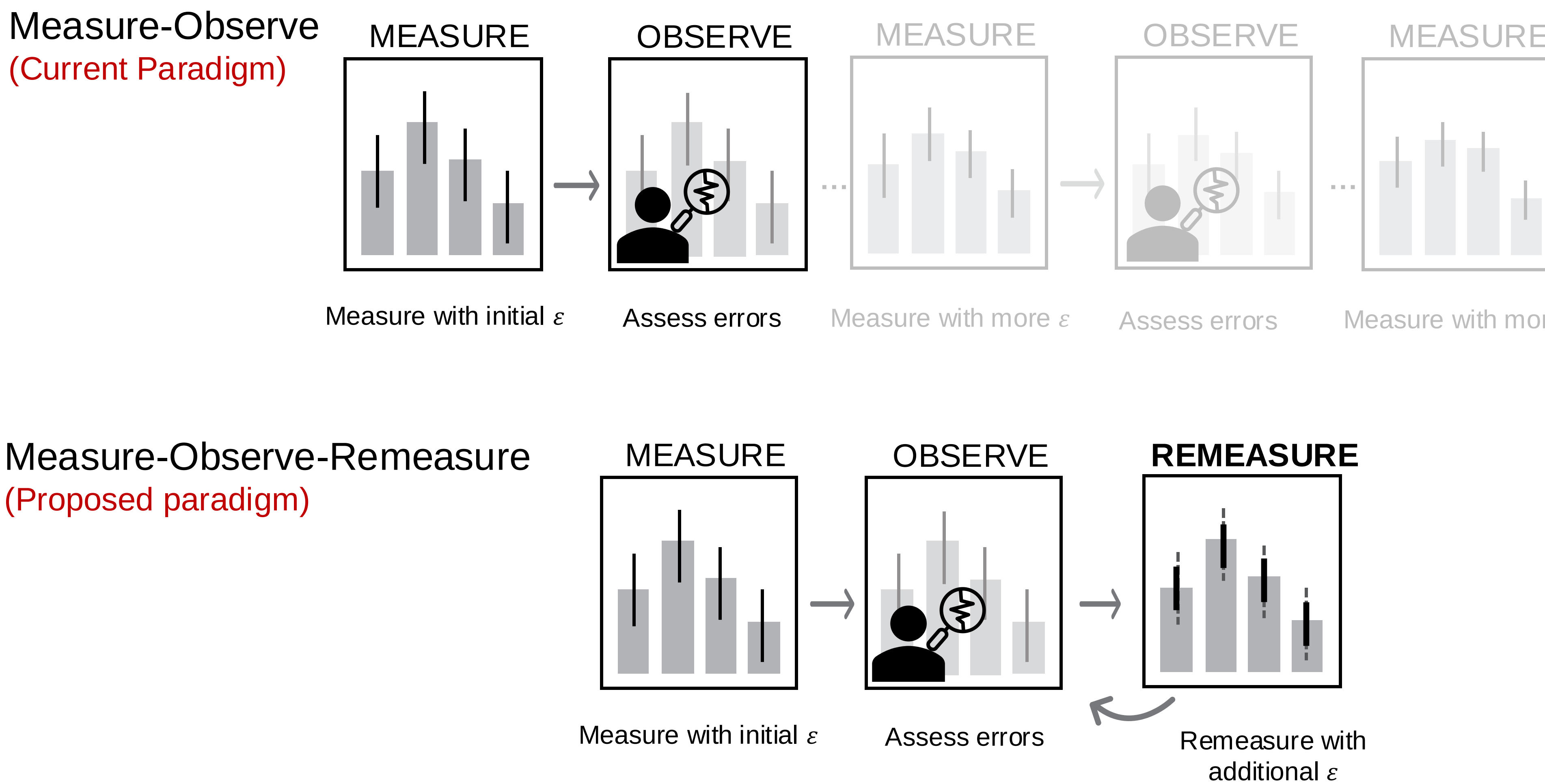


assumes queries are **known in advance**

Measure-Observe
(Current Paradigm)

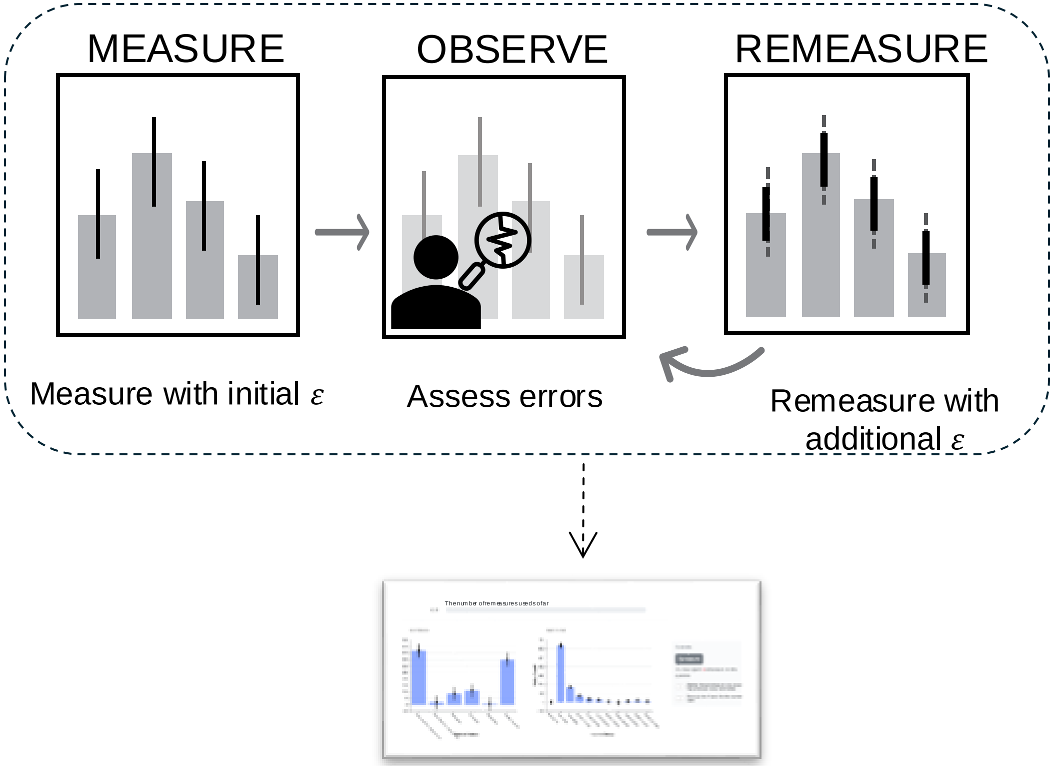


How might analysts efficiently spend **ε** when queries are **not known in advance**?

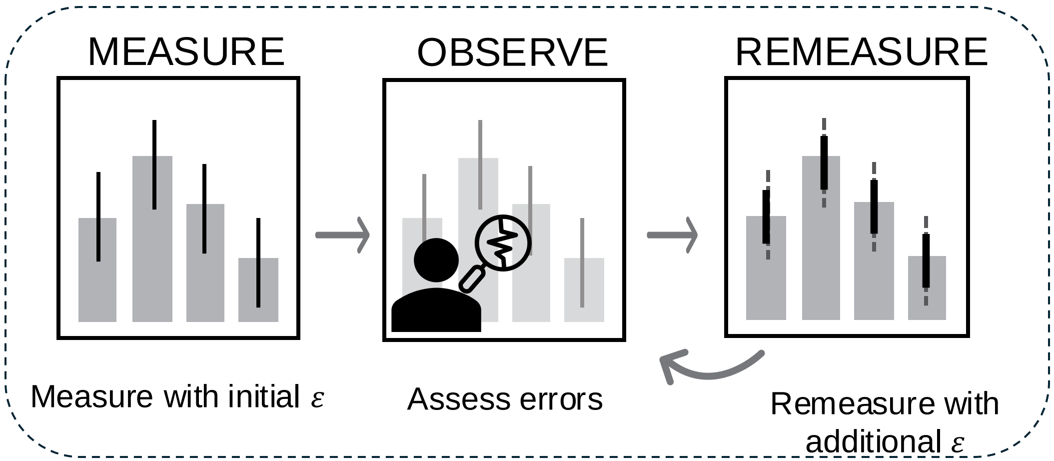


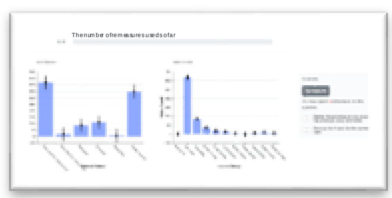
Measure-Observe
(Current Paradigm)
MEASURE
OBSERVE
MEASURE
OBSERVE
MEASURE
Measure with initial ε
Assess errors
Measure with more ε
Assess errors
Measure with mor
Measure-Observe-Remeasure
(Proposed paradigm)
MEASURE
OBSERVE
REMEASURE
Measure with initial ε
Assess errors
Remeasure with
additional ε

INSTANTIATING THE MEASURE-OBSERVE-REMEASURE PARADIGM



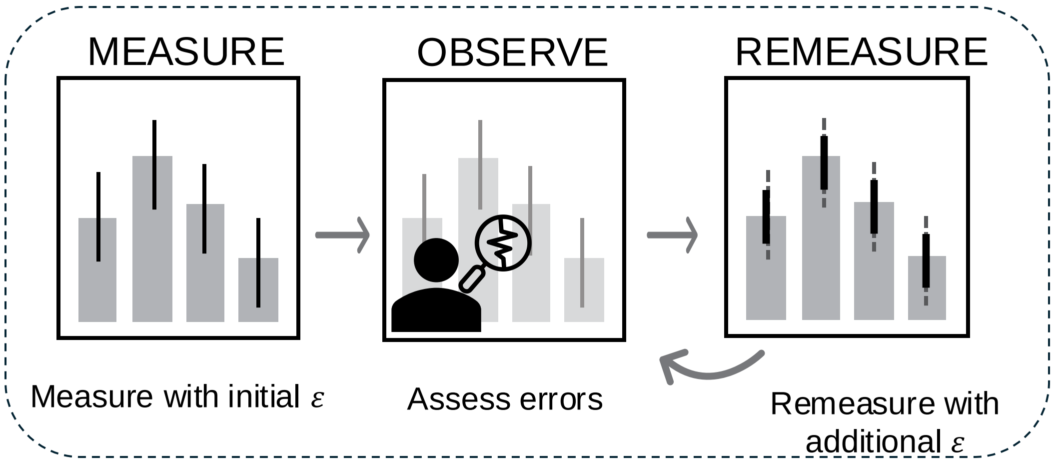
INSTANTIATING THE MEASURE-OBSERVE-REMEASURE PARADIGM

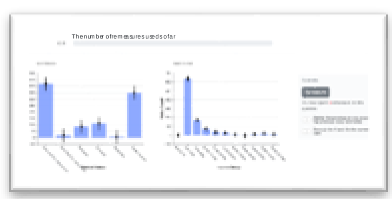




High Dimensional Matrix Mechanism (HDMM) (McKenna Miklau Hay Machanavajjhala 2018) using the Laplace Mechanism (Dwork McSherry Nissim Smith 06)

INSTANTIATING THE MEASURE-OBSERVE-REMEASURE PARADIGM

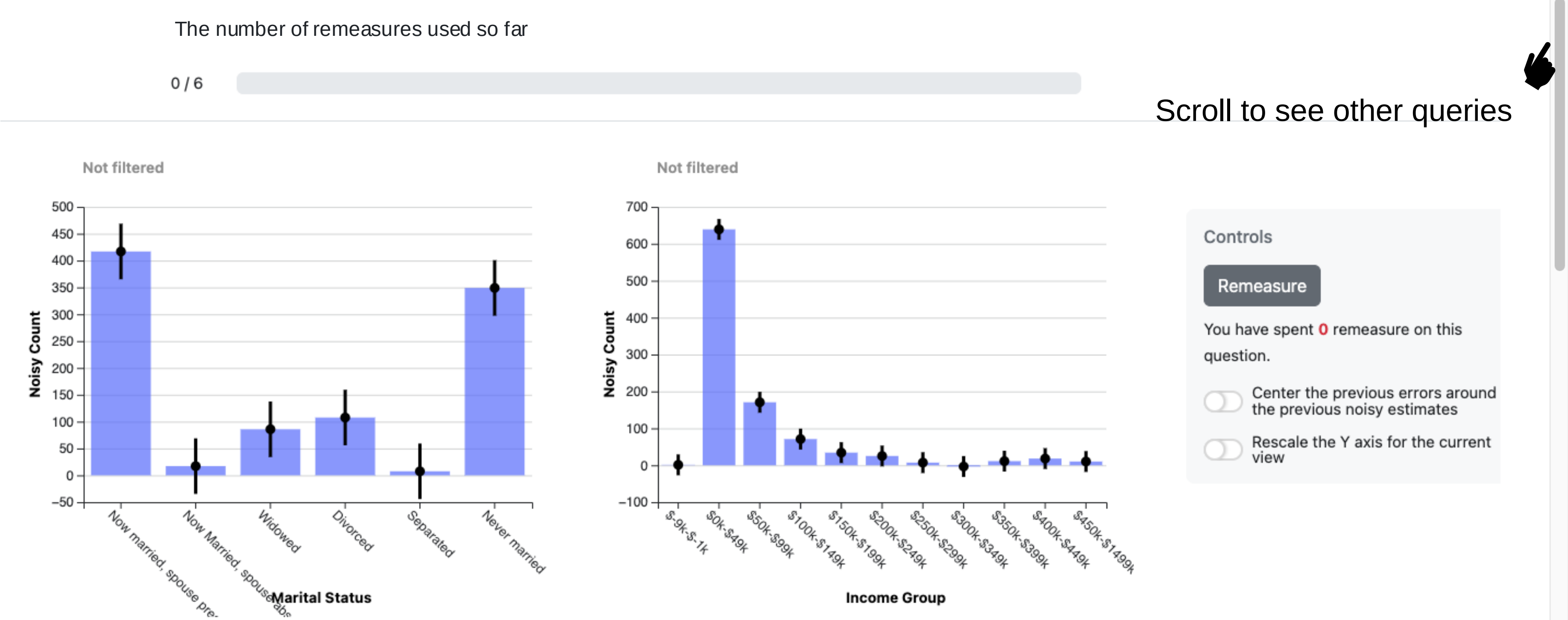




High Dimensional Matrix Mechanism (HDMM) (McKenna Miklau Hay Machanavajjhala 2018) using the Laplace Mechanism (Dwork McSherry Nissim Smith 06)

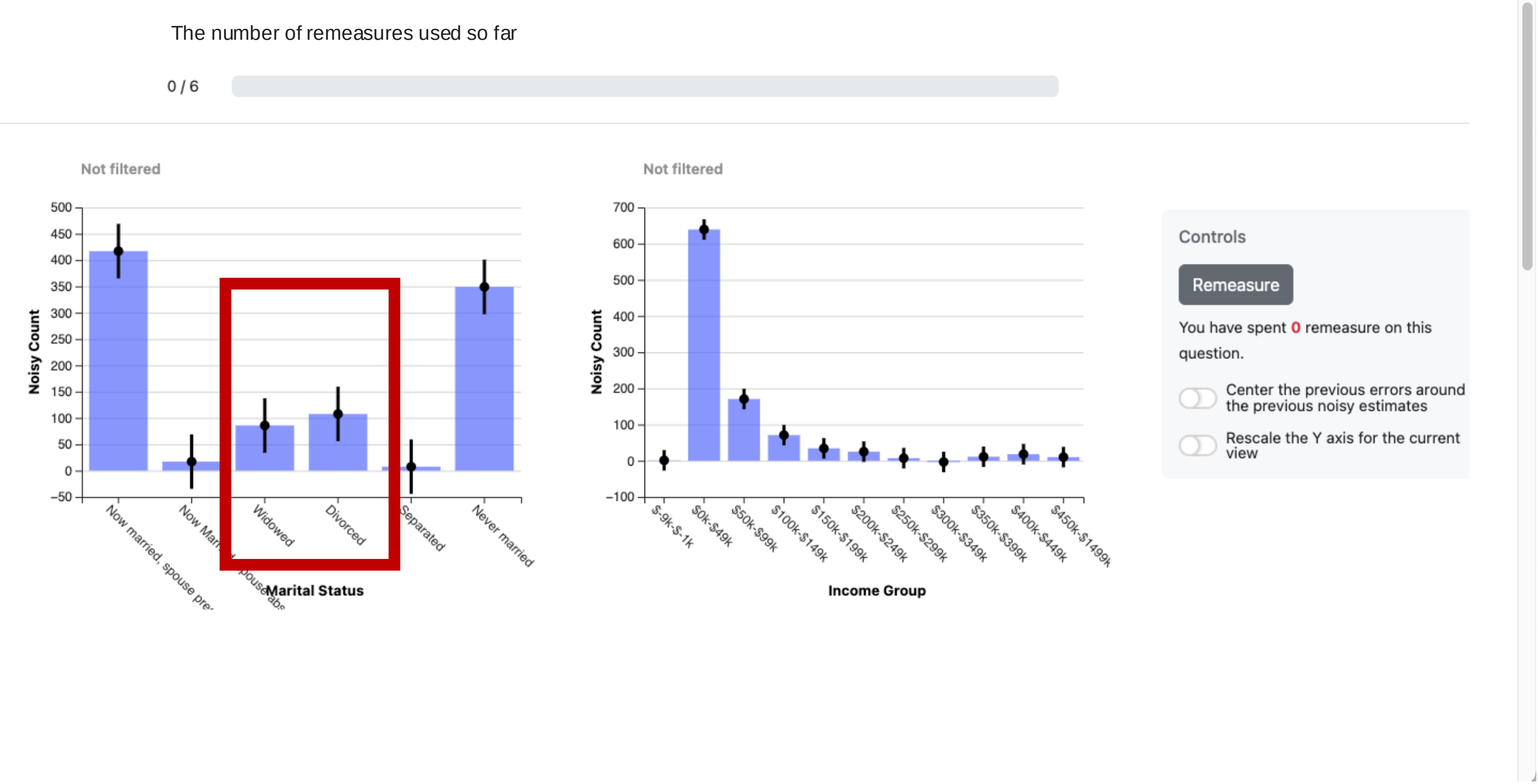
“Inference” step
(Li Miklau Hay McGregor Rastogi 15, McKenna et al. 2018)

MEASURE-OBSERVE-REMEASURE PARADIGM → INTERFACE

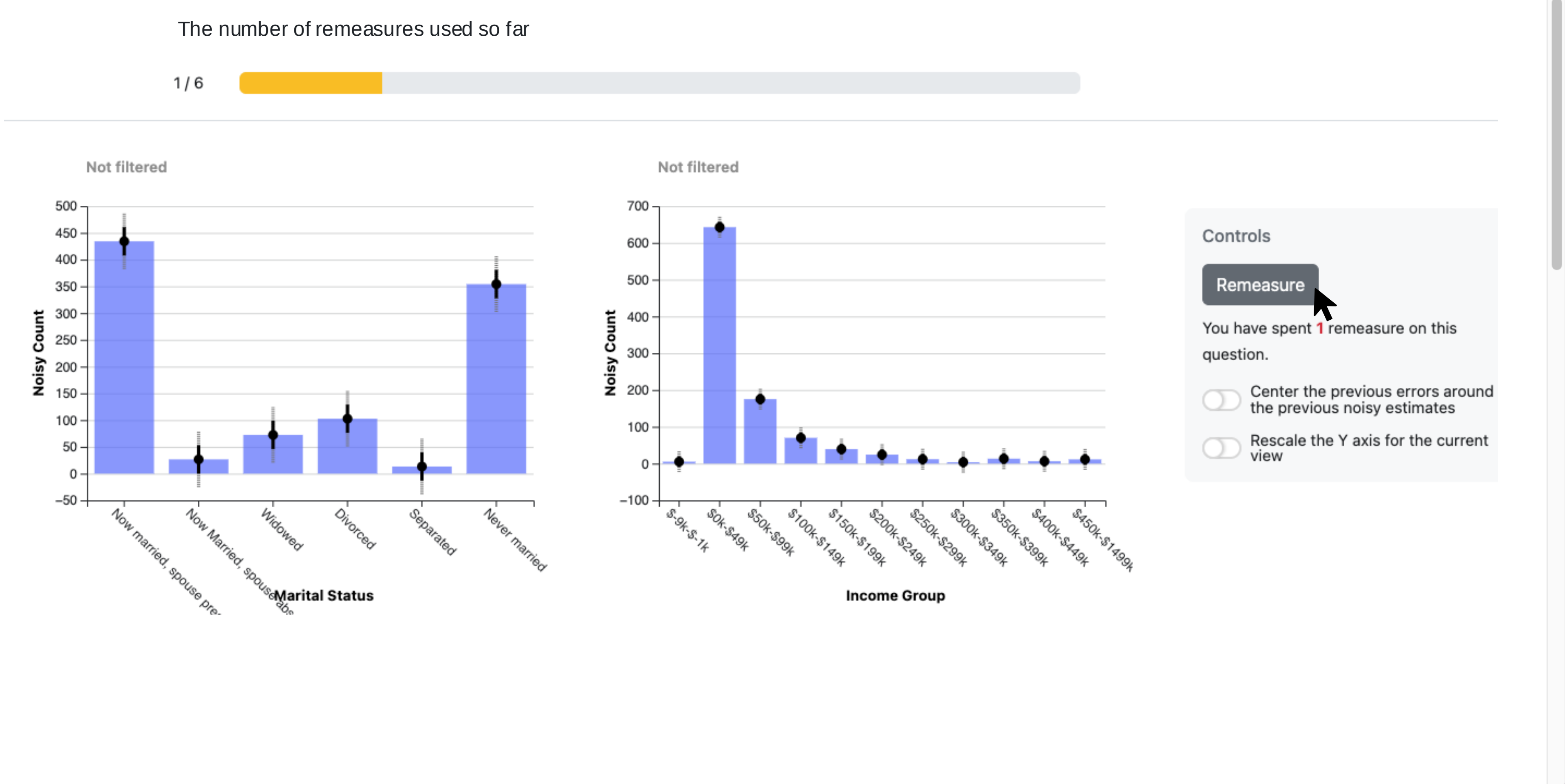


Scroll to see other queries

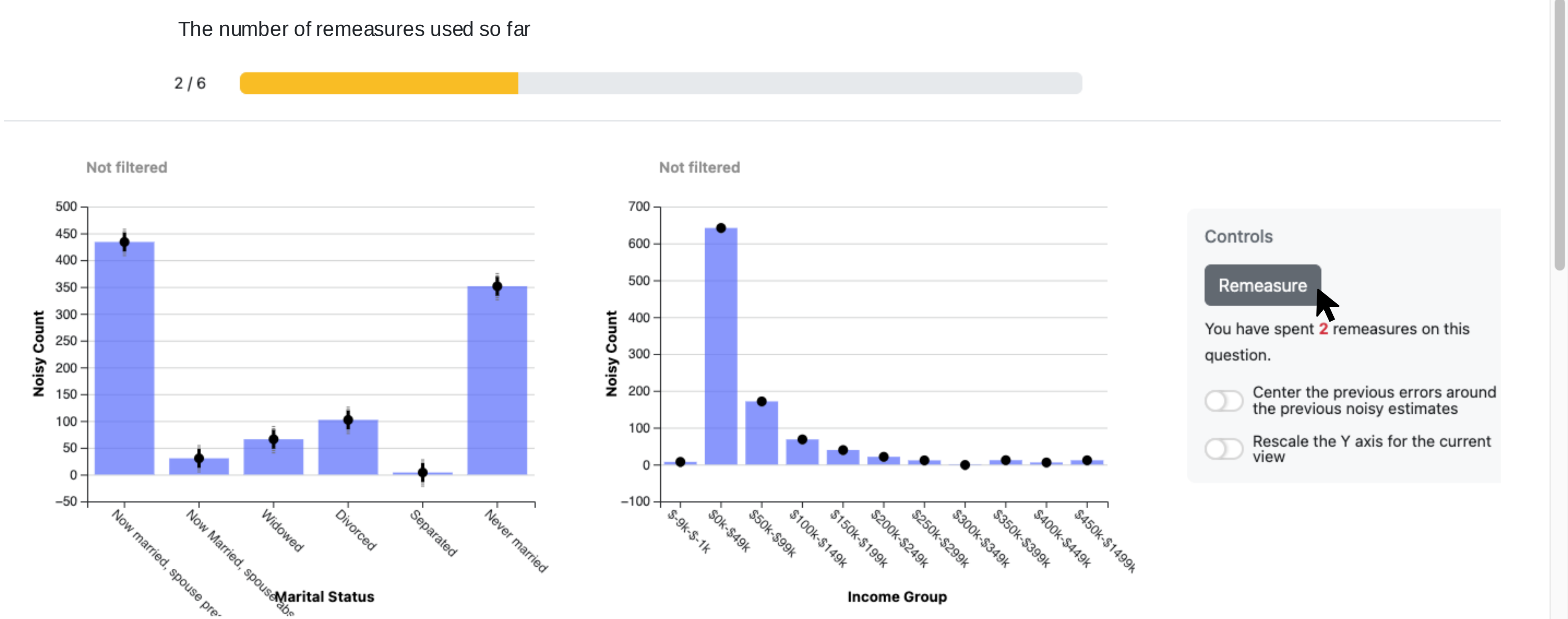
MEASURE-OBSERVE-REMEASURE PARADIGM → INTERFACE



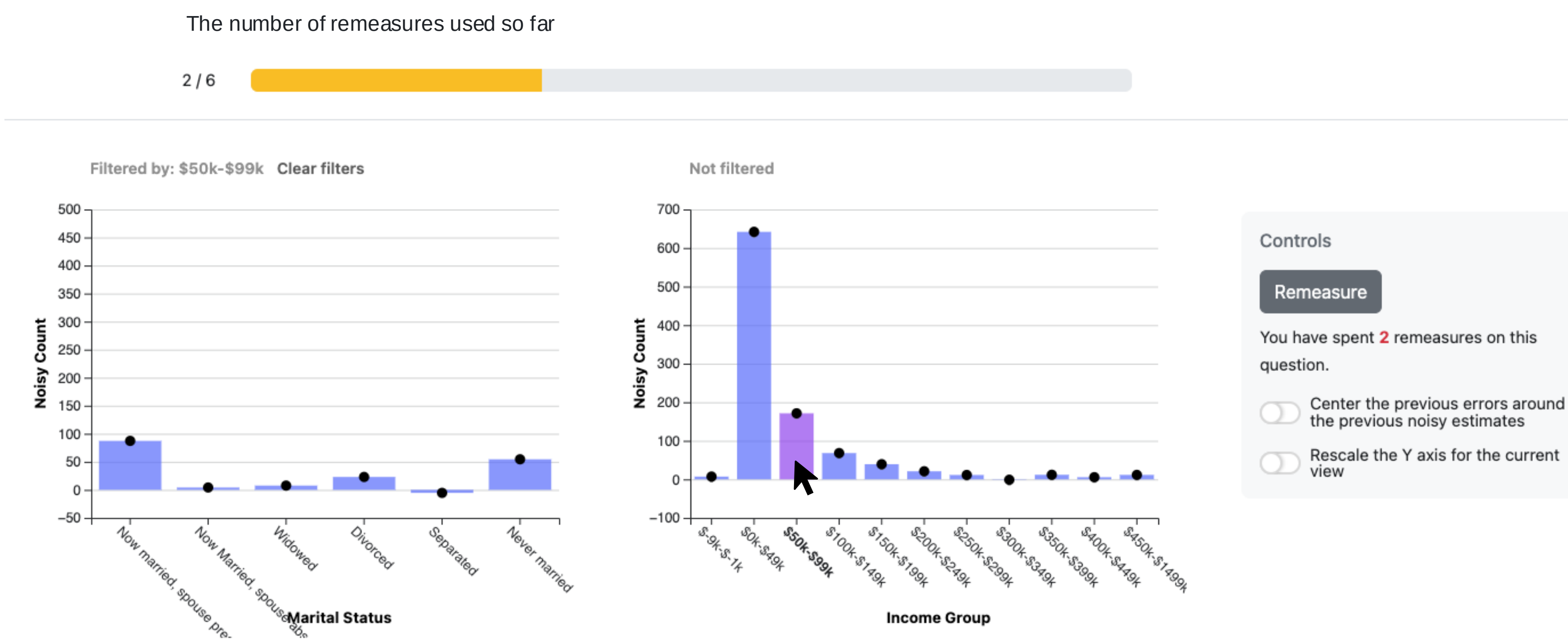
MEASURE-OBSERVE-REMEASURE PARADIGM → INTERFACE



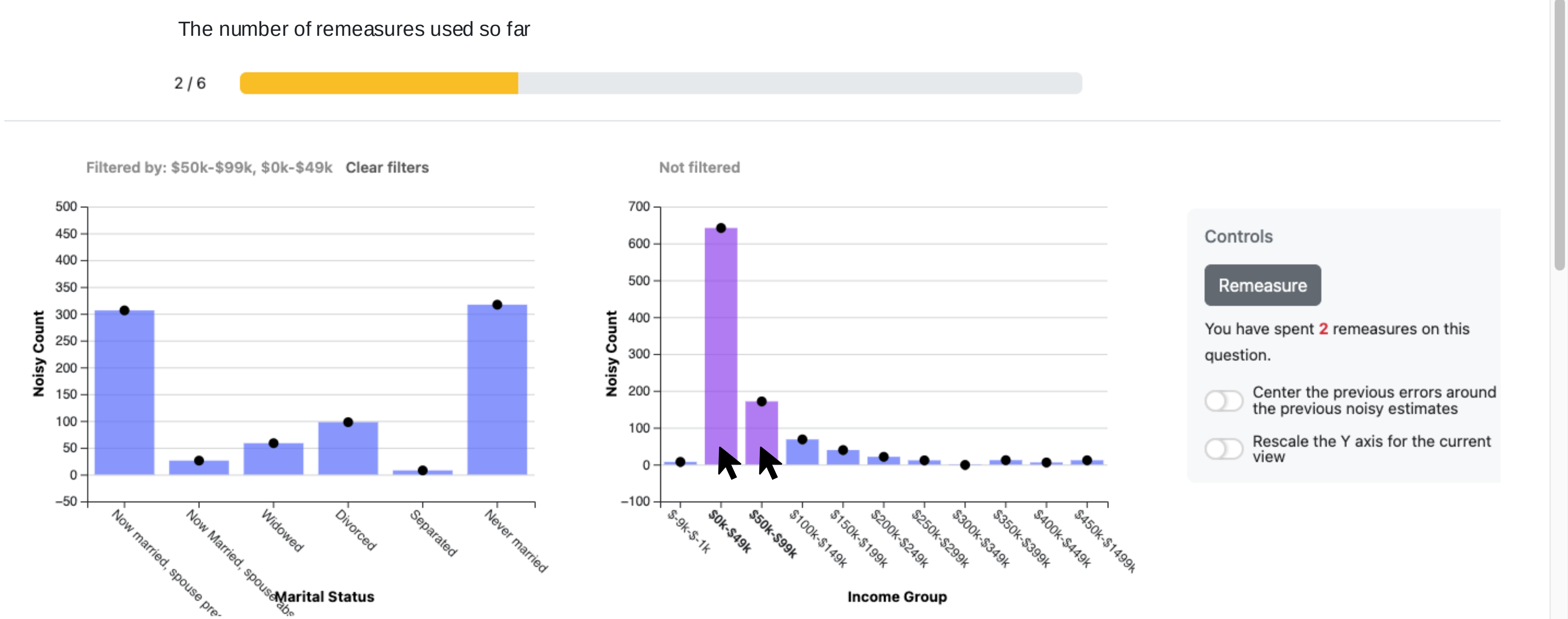
MEASURE-OBSERVE-REMEASURE PARADIGM → INTERFACE



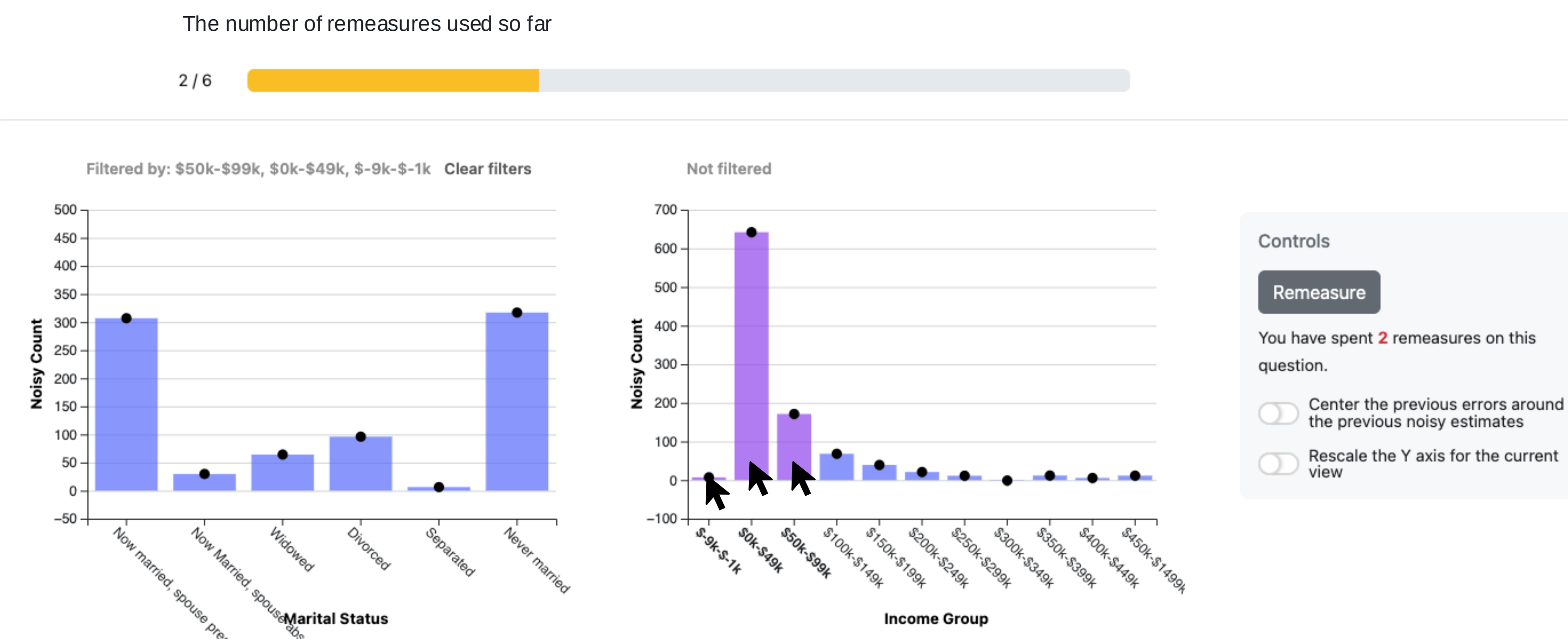
MEASURE-OBSERVE-REMEASURE PARADIGM → INTERFACE

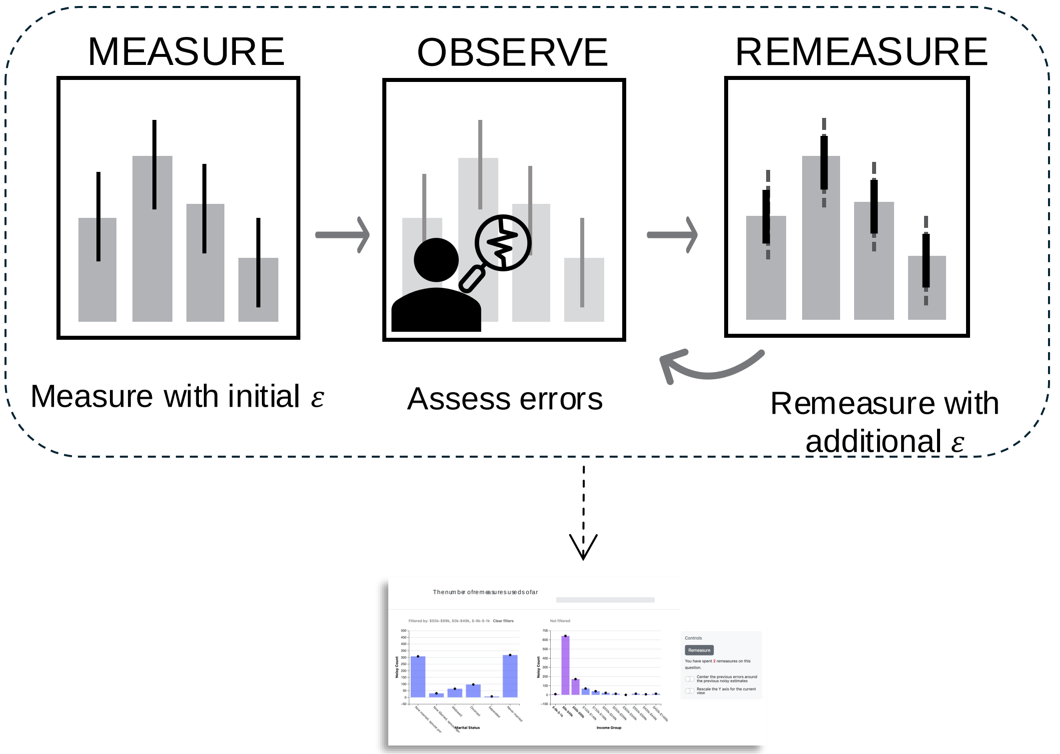


MEASURE-OBSERVE-REMEASURE PARADIGM → INTERFACE



MEASURE-OBSERVE-REMEASURE PARADIGM → INTERFACE




MEASURE
OBSERVE
REMEASURE
Measure with initial ε
Assess errors
Remeasure with additional ε

Exploratory user study

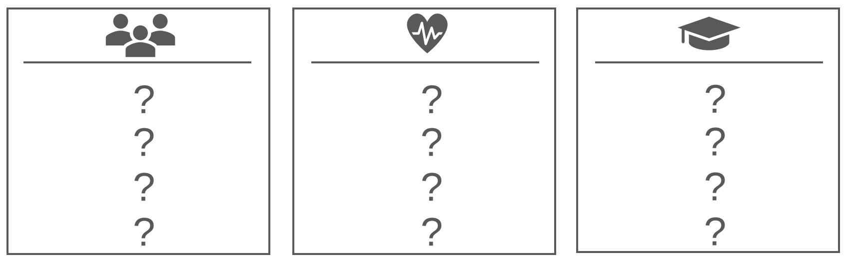
14 participants, experienced with quantitative data analysis, but <u>not</u> DP experts

Exploratory user study

14 participants, experienced with quantitative data analysis, but <u>not</u> DP experts

Exploratory user study

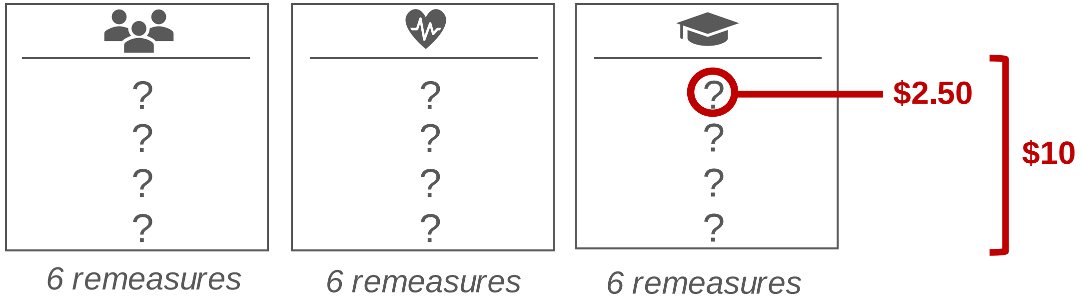
14 participants, experienced with quantitative data analysis, but not DP experts



6 remeasures *6 remeasures* *6 remeasures*

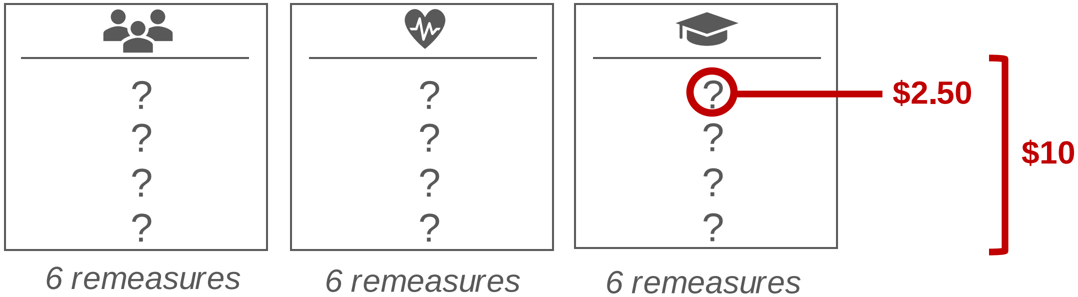
Exploratory user study

14 participants, experienced with quantitative data analysis, but <u>not</u> DP experts



Exploratory user study

14 participants, experienced with quantitative data analysis, but <u>not</u> DP experts

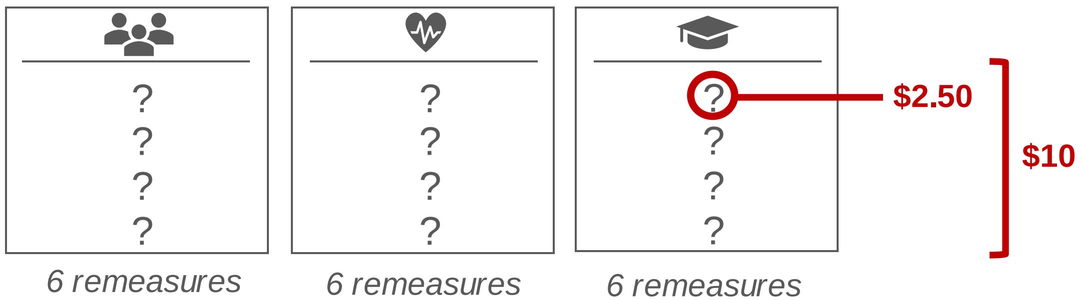


QUANTITATIVE

BINARY

Exploratory user study

14 participants, experienced with quantitative data analysis, but <u>not</u> DP experts



QUANTITATIVE

How many people over 65 years old are widowed or divorced?

Interval scoring rule (Gneiting & Raftery 2007)

BINARY

Are there more than 327 people who have never been married and make less than $100,000?

Brier/quadratic scoring rule (Brier 1950)

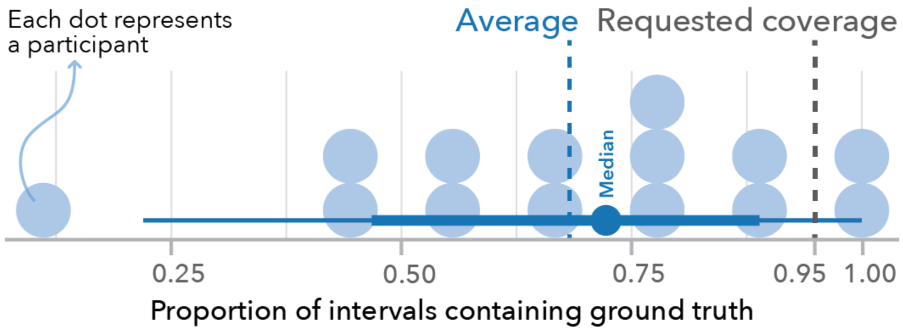
Exploratory user study

QUANTITATIVE QUESTIONS

~70% of interval responses contained the ground truth answer

11 (out of 14) participants provided intervals containing ground truth in over half their reported intervals

2 participants always provided intervals containing ground truth



Exploratory user study

BINARY QUESTIONS

Mean absolute error in “yes” probability responses was 0.37

(95% CI: [0.26, 0.49])

Over 60% of probability responses were in the same direction as ground truth (i.e., rounding to a whole number yields ground truth yes/no [1,0] answer)

- Participants tended to provide “extreme answers,” where probability allocated to yes/no were close to 0 or 1

Exploratory user study

PAYOFF

Avg payoff per dataset: $6.06 (out of $10)

Avg total payoff: $18.17 (out of $30)

Self-rated confidence in responses: mean = 75 (out of 100); median = 78

Exploratory user study

Rational agent framework (Wu Guo Mamakos Hartline Hullman 23)

Ideal world

RATIONAL AGENT

vs.

Real world (experiment)

PARTICIPANT

Slide adapted from Yifan Wu

Exploratory user study

Rational agent framework (Wu Guo Mamakos Hartline Hullman 23)

Ideal world

vs.

Real world (experiment)

RATIONAL AGENT

Upper bound

Lower bound

PARTICIPANT

Slide adapted from Yifan Wu

Exploratory user study

Rational agent framework (Wu Guo Mamakos Hartline Hullman 23)

Ideal world

vs.

Real world (experiment)

RATIONAL AGENT

Upper bound

Lower bound

PARTICIPANT

Non-optimal decisions

Slide adapted from Yifan Wu

Exploratory user study

Rational agent framework (Wu Guo Mamakos Hartline Hullman 23)

Define a rational agent

- Go through the experiment
- **Optimally** process visualizations of noisy estimates (Bayesian posterior)
- Make **best-response** decision

Slide adapted from Yifan Wu

Exploratory user study

Rational agent framework (Wu Guo Mamakos Hartline Hullman 23)

Define a rational agent

- Go through the experiment
- **Optimally** process visualizations of noisy estimates (Bayesian posterior)
- Make **best-response** decision

Prior
↓
See vis of noisy estimates
↓
Posterior
↓
Action

Slide adapted from Yifan Wu

Exploratory user study

Rational agent framework (Wu Guo Mamakos Hartline Hullman 23)

Define a rational agent

- Go through the experiment
- **Optimally** process visualizations of noisy estimates (Bayesian posterior)
- Make **best-response** decision

Prior

↓

See vis of noisy estimates

↓

Posterior

↓ Best action? Expected payoff maximizer

Action

Slide adapted from Yifan Wu

Exploratory user study

BENCHMARKS

UPPERBOUND = upper bound on participant's expected payoff

LOWERBOUND = lower bound on participants' expected payoff

Exploratory user study

BENCHMARKS

UPPERBOUND = upper bound on participant's expected payoff

RPOSTERIOR$_{\textbf{Zero}}$ = rational agent's expected payoff spending no remeasures

RPOSTERIOR$_{\textbf{Rand}}$ = rational agent's expected payoff using a random allocation strategy

RPOSTERIOR$_{\textbf{Same}}$ = rational agent's expected payoff using the same allocation strategies as participants

LOWERBOUND = lower bound on participants' expected payoff

Exploratory user study

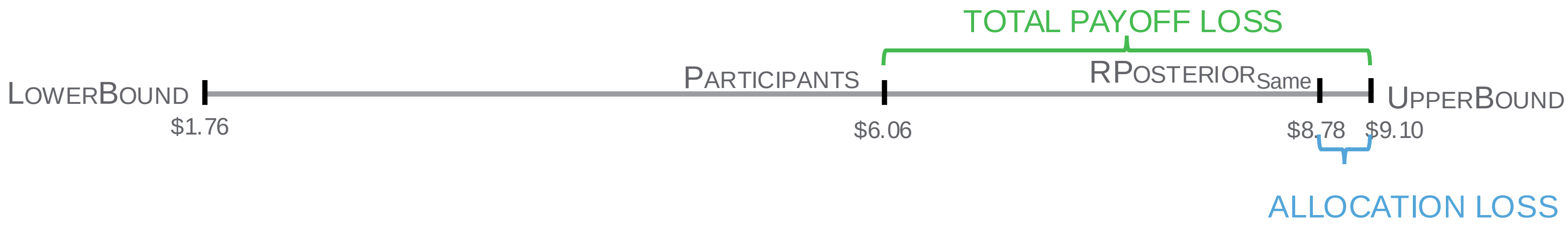
On average, **participants lost 41% of the possible payoff attainable**

Exploratory user study

On average, **participants lost 41% of the possible payoff attainable**

Exploratory user study

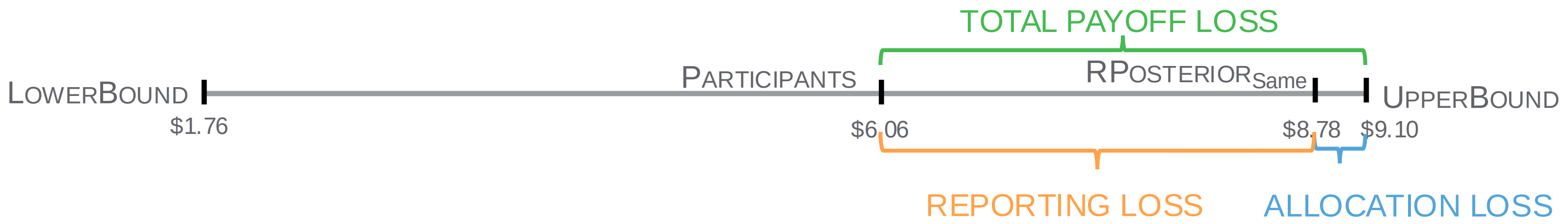
On average, **participants lost 41% of the possible payoff attainable**



4% due to allocation loss

Exploratory user study

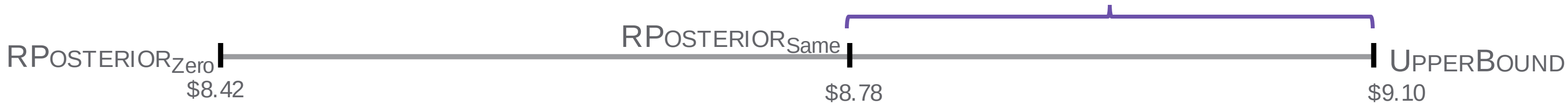
On average, **participants lost 41% of the possible payoff attainable**



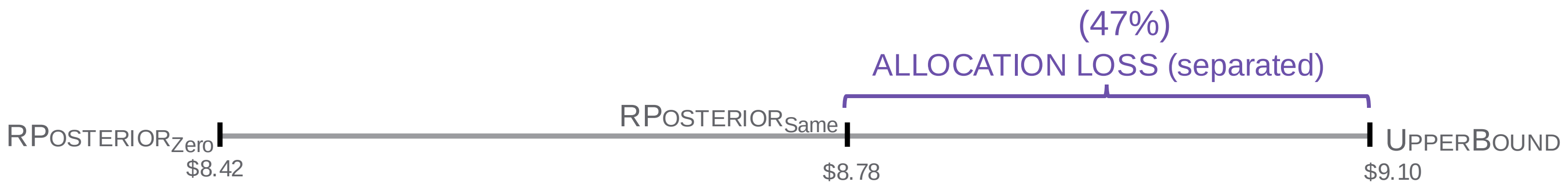
37% due to reporting loss **4% due to allocation loss**

Exploratory user study

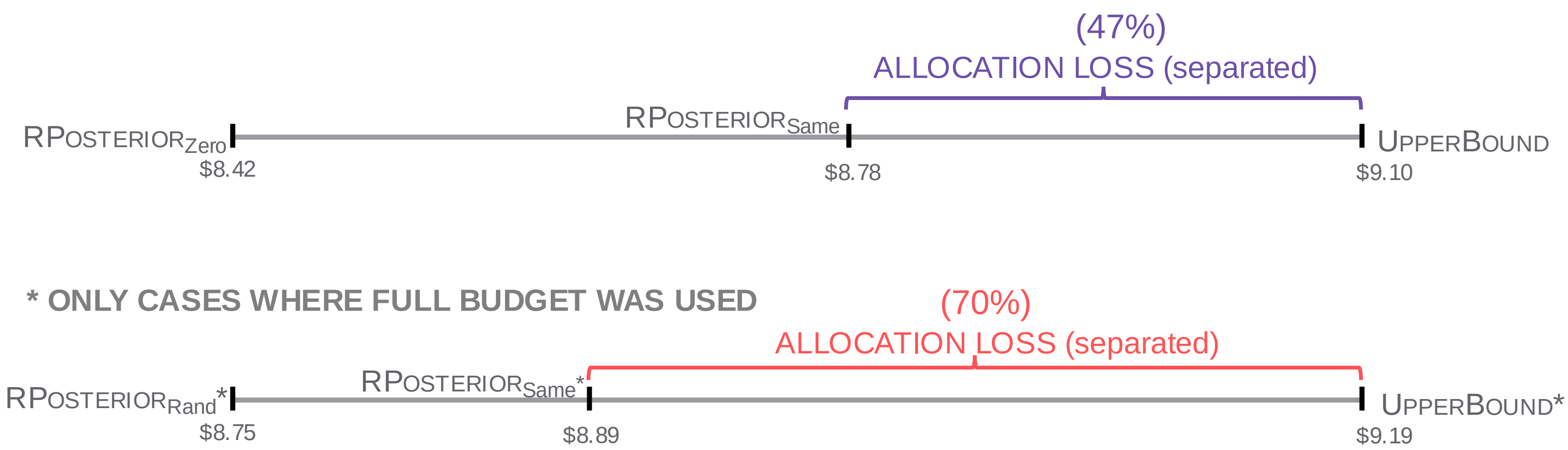
Exploratory user study



Exploratory user study



Exploratory user study



Interfaces as *technology probes*

- Interfaces can also help us learn potential challenges or opportunities related to bringing DP into practice
- *How?* By using interfaces as *technology probes*

Interfaces as *technology probes*

- Interfaces can also help us learn potential challenges or opportunities related to bringing DP into practice
- *How?* By using interfaces as *technology probes*

> "A probe is an instrument that is deployed to find out about the unknown - to hopefully return with useful or interesting data…Technology probes are a particular type of probe that combine the social science goal of collecting information about the use and the users of the technology in a real world setting, the engineering goal of field-testing the technology, and the design goal of inspiring users and designers to think of new kinds of technology to support their needs and desires."
> [Hutchinson, Mackay, Westerlund, Bederson, …, Sundblad 2003]

RESEARCH-ARTICLE

Don’t Look at the Data! How Differential Privacy Reconfigures the Practices of Data Science

Authors: Jayshree Sarathy, Sophia Song, Audrey Haque, Tania Schlatter, Salil Vadhan | Authors Info & Claims

CHI '23: Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems • Article No.: 164, Pages 1 - 19
https://doi.org/10.1145/3544548.3580791

- Semi-structured interviews with practitioners using a DP analysis interface as a technology probe
- Exploratory insights about challenges & opportunities around DP in practice

Method & research questions

- Semi-structured interviews with 19 practitioners without DP expertise but experienced analyzing sensitive data. Observed them using DPCreator (descendant of PSI – similar goals, features, etc.)
- Research questions:
  - What barriers do data practitioners who are non-experts in DP face when using DP to share or analyze sensitive datasets?
  - What do data practitioners who are non-experts in DP perceive to be the potential utility of DP for expanding access of sensitive data to the public, facilitating exploratory data analysis, and enabling replication of scientific studies?
  - What changes need to be made in the data science workflow in order to address the barriers and realize the benefits (from RQ1 and RQ2) of DP?

DP Creator
from OpenDP

My Data   My Profile   Logout

0 **Validate Dataset** — 1 Confirm Variables — 2 Set Epsilon Value — 3 Create Statistics — 4 Generate DP

Continue

Last saved: 2021-09-21T12:07:55.981095Z

Remaining: 2d 14h 13min

Used dataset: Replication Data for: Eye-typing experiment

Validate Dataset

Firstly, we need to confirm the dataset's characteristics to determine if it's adequate for the differential privacy release process.

You must complete this questionnaire before starting the process.

▸ Does your dataset depend on private information of subjects?

- (●) Yes.
- ( ) No.
- ( ) I'm unsure.

▸ Which of the following best describes your dataset?

- ( ) Public information.
- ( ) Information that, if disclosed, **would not cause material harm**, but which the university has chosen to **keep confidential**.
- (●) Information that **could cause risk of material harm** to individuals or the university if disclosed.
- ( ) Information that **would likely cause serious harm** to individuals or the university if disclosed.

DP Creator
from OpenDP

My Data My Profile Logout

The DPcreator takes the first 20 variables of the dataset. The default type has been inferred from the dataset. Incorrect type labeling can result in privacy violation.

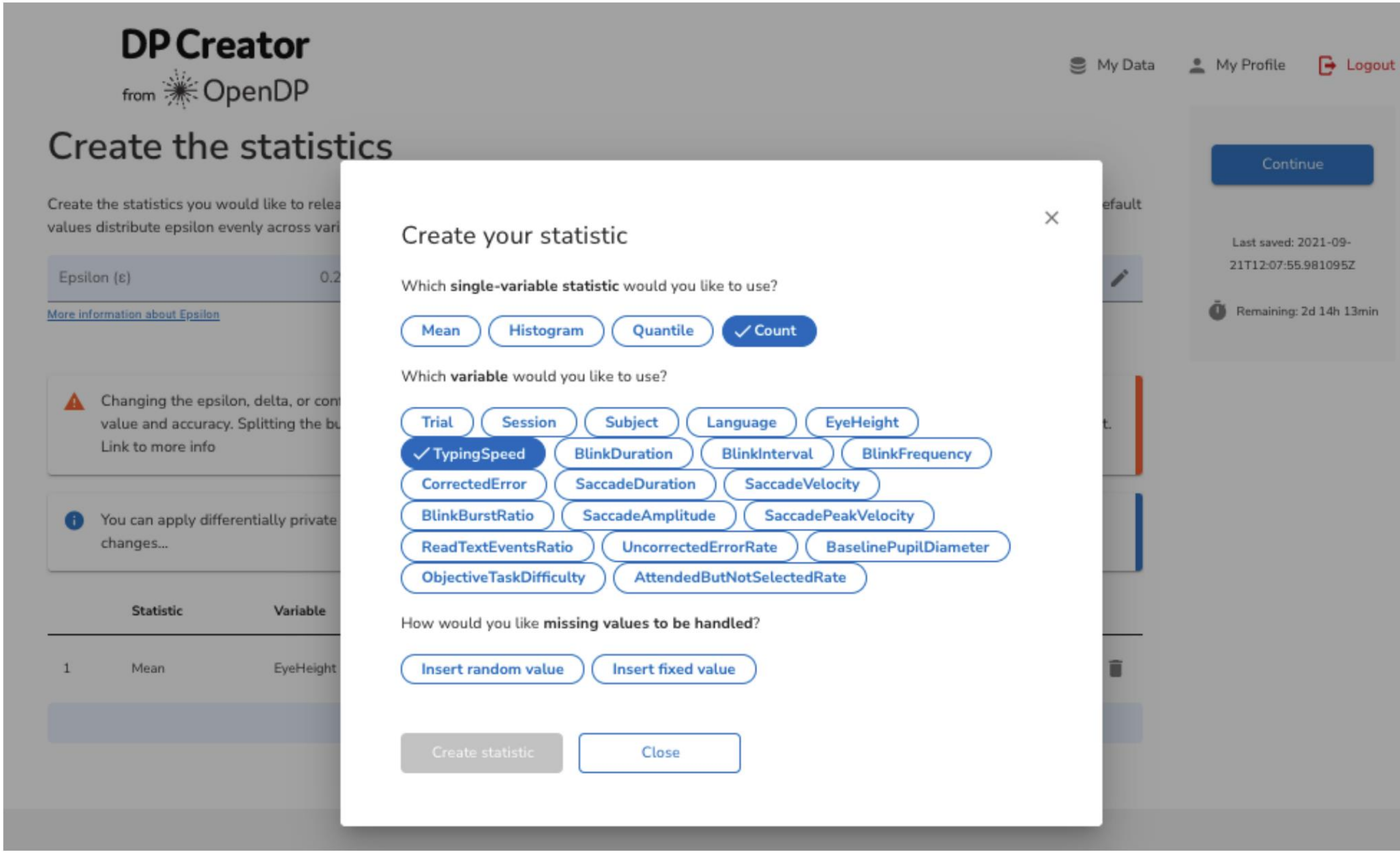
Any changes will be applied for the purpose of creating the differential privacy release only, and will **not affect** the original data file. dataset. Incorrect type labeling can result in privacy violation.

Continue

Last saved: 9/20/2021, 10:22:49 PM

Remaining: 2d 14h 16min

	Variable name	Variable label	Type	Additional variable information	
1	Trial	Trial	Categorical	Add categories	
2	Session	Session	Boolean		
3	Subject	Subject	Categorical	Add categories	
4	Language	Language	Boolean		
5	EyeHeight	EyeHeight	Numerical	-8	5
6	TypingSpeed	TypingSpeed	Numerical	Add min	Add max



DP Creator
from OpenDP
My Data
My Profile
Logout
Create the statistics
Create the statistics you would like to relea
values distribute epsilon evenly across vari
efault
Epsilon (ε)
0.2
More information about Epsilon
Changing the epsilon, delta, or con
value and accuracy. Splitting the bu
Link to more info
You can apply differentially private
changes...
Statistic
Variable
1
Mean
EyeHeight
Continue
Last saved: 2021-09-21T12:07:55.981095Z
Remaining: 2d 14h 13min
Create your statistic
Which single-variable statistic would you like to use?
Mean
Histogram
Quantile
Count
Which variable would you like to use?
Trial
Session
Subject
Language
EyeHeight
TypingSpeed
BlinkDuration
BlinkInterval
BlinkFrequency
CorrectedError
SaccadeDuration
SaccadeVelocity
BlinkBurstRatio
SaccadeAmplitude
SaccadePeakVelocity
ReadTextEventsRatio
UncorrectedErrorRate
BaselinePupilDiameter
ObjectiveTaskDifficulty
AttendedButNotSelectedRate
How would you like missing values to be handled?
Insert random value
Insert fixed value
Create statistic
Close

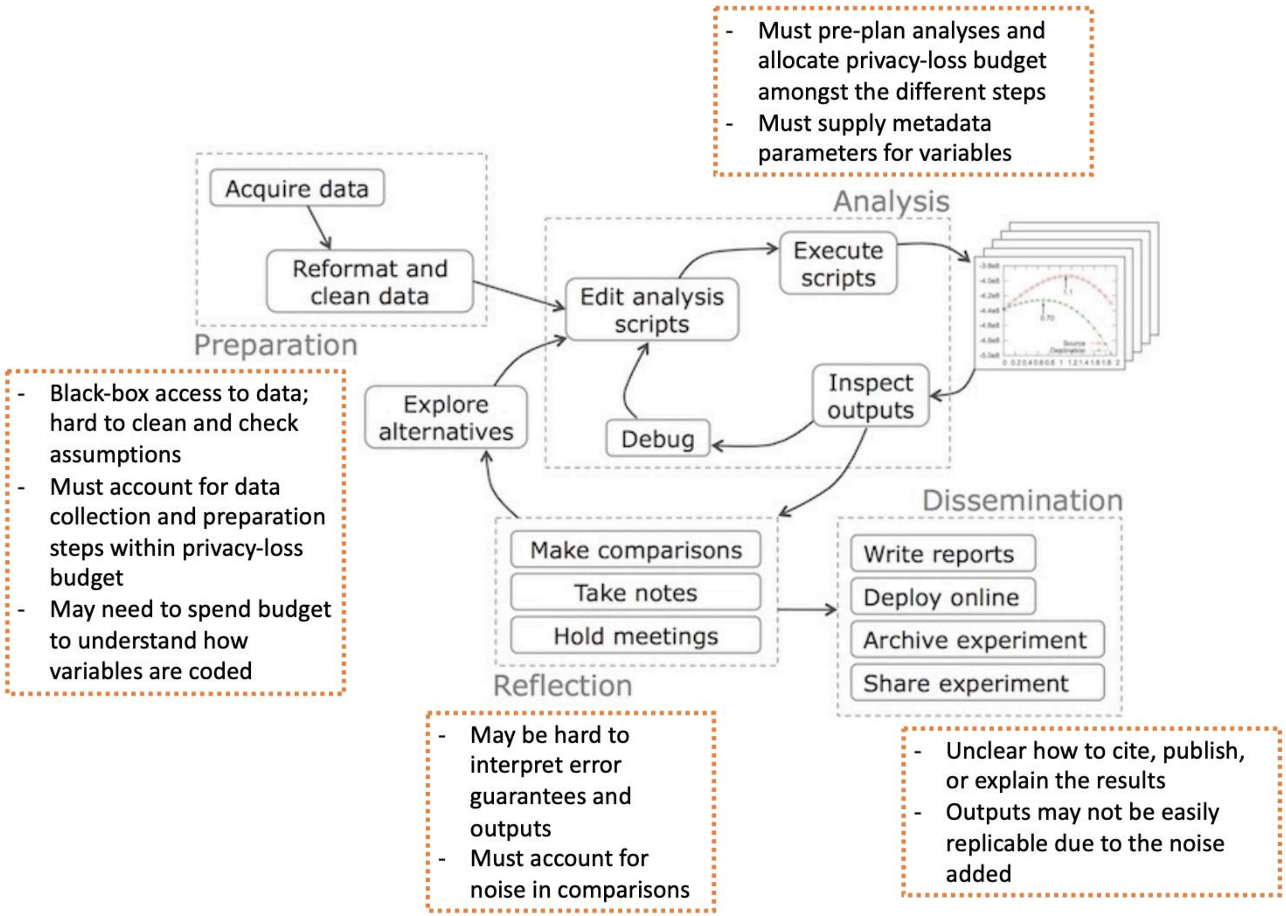
Task

Imagine that you are a researcher who has collected sensitive information about individuals in California, including attributes such as sex, race, marital status, and income level. Your dataset contains a simple random sample of 30,000 individuals from a population of 30 million individuals in California. Your goal is to use the DP Creator prototype to release privacy-preserving summary statistics that convey the main insights of your data. We'd like you to use the prototype to generate the following differentially private statistics:

1) Mean age of individuals in California; missing values should be replaced with the number 30.

2) OLS regression using x=age and y=sex as variables; missing values for age should be replaced with 30 and missing values for sex should be dropped.

Findings



Challenges posed by DP mapped onto Guo 2013’s data science workflow

Takeaways

- Designing DP interfaces requires being intentional about your audience, their goals, and their background
- Visualization enables presenting unfamiliar concepts in familiar formats
- Interactivity helps convey relationships between key DP concepts and embed DP into existing workflows – though gaps may still exist!

Takeaways continued

- There are multiple styles of user studies, which each enable us to answer different questions:
  - **Usability study**: Can potential users complete low-level interactions with the interface? I.e., can they "use" it?
  - **Controlled user study**: Can potential users complete higher-level conceptual tasks more accurately with the interface compared to with a baseline tool?
  - **Decision-theoretic user study**: How well do potential users complete higher-level tasks (decision problems) compared to a well-defined notion of best possible performance?
  - **Technology probe user study** (in context of DP): more open-ended exploratory questions, like what challenges might potential users <u>of DP</u> face in practice?

Suggestions for designing interfaces, running user studies

Geared toward course projects, but also broadly applicable to interface design.

- Carefully consider your audience
  - What are their goals?
  - What are their needs?
  - What is their background?
- Develop specific design goals (based on your prior knowledge, research, or formative research) for your interface
- When developing your interface, start with low-fidelity mockups, then work your way up to developing out a higher-fidelity prototype
- When running your user study (e.g., to evaluate whether your interface meets its design goals), carefully consider what tasks for participants will help you answer your research questions. Ensure that participants in your study represent intended users of the interface.
  - https://cuhs.harvard.edu/undergraduate-research-and-course-projects (IRB guidance for course projects with human subjects)