2024 FCSM Research & Policy Conference

Building an Automated Validation Server Prototype

Safe Data Technologies: Safely Expanding Access to Administrative Tax Data

URBAN INSTITUTE · ELEVATE · THE · DEBATE

Erika Tyagi, Silke Taylor, Graham MacDonald, Deena Tamaroff, Josh Miller, Aaron R. Williams & Claire McKay Bowen

Tiered Access for Administrative Tax Data

Our goal is to enable more researchers to safely access confidential tax data.

I. **Basic Access:** Summary tables on the Statistics of Income Division website.

II. **Public Use File (PUF):** Researchers will have access to the synthetic PUF.

III. **Validation Server Access:** Researchers are trusted to access the validation server, where they submit statistical analyses that they have tested and debugged on the synthetic PUF. Researchers at this tier will have to undergo an application process.

IV. **Full Access:** Researchers who obtain clearance and therefore have access to the unaltered, confidential data, but will be still be limited on what information can be released.

Access Mechanism

User Interface Layer

Data User

Data User

Data User

Analysis 1

Analysis 2

Analysis 3

Analysis 4

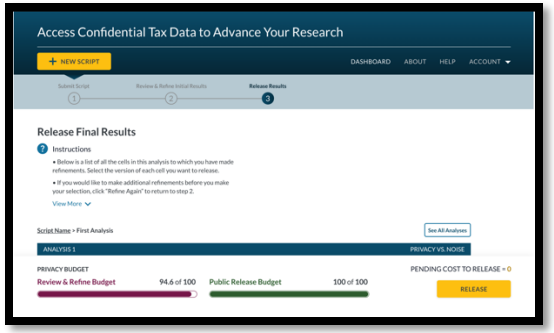
Analysis 5

Analysis 6

Analyses

Results

Privacy Layer



Apply privacy-preserving methodologies

Analysis 1.1

Analysis 1.2

n o i s e

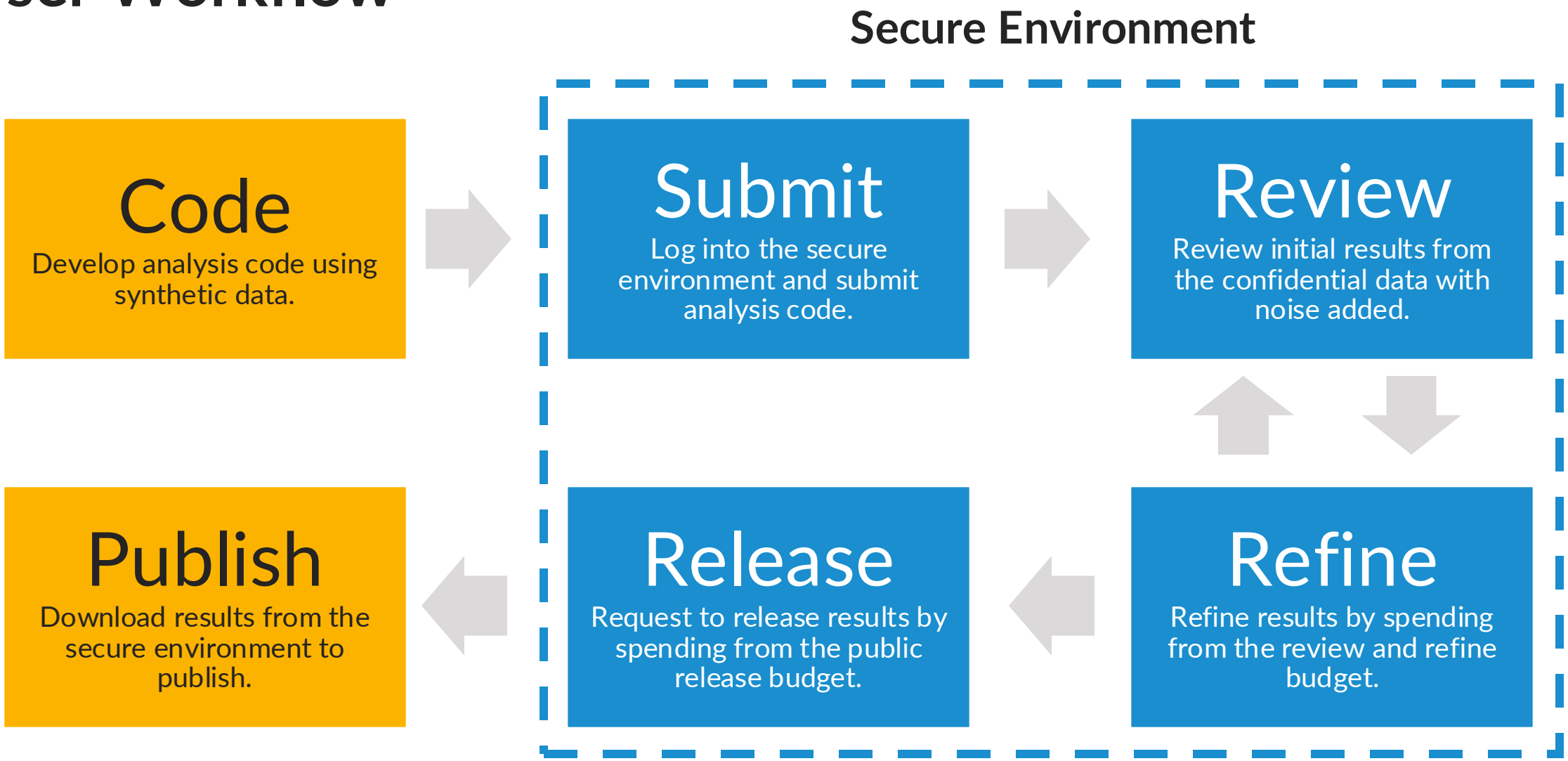
Analysis 1

Analyses

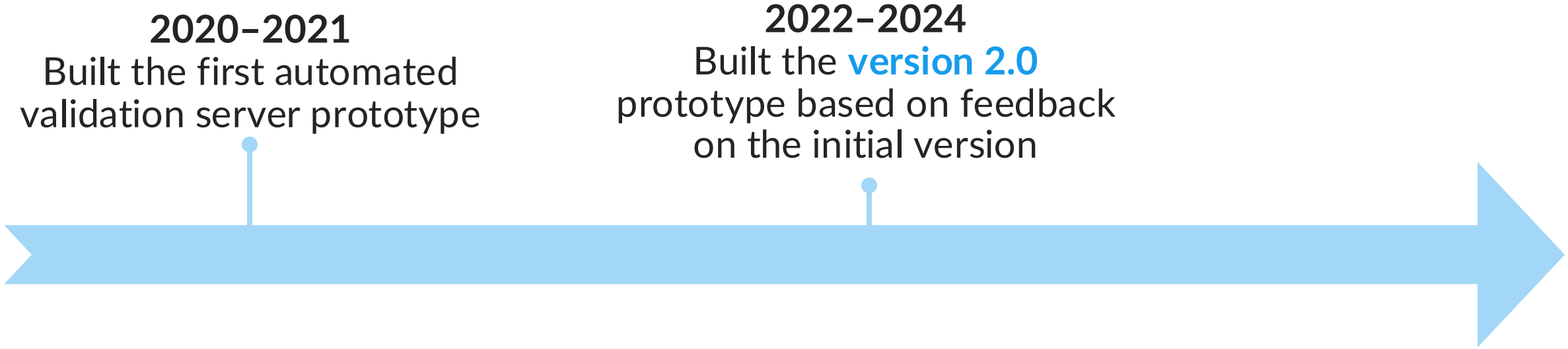
Statistics

Confidential Data

User Workflow



Prototype Development History



User testing from the initial prototype revealed that researchers wanted to:

- Submit a broader range of analyses (particularly regressions)
- Develop code using a more familiar programming language (such as R or Stata)
- Have more informed and granular control over their privacy budgets

Follows Familiar Researcher Workflows

- **Accepts R code**: Supports analyses developed using the R programming language and flexible preprocessing code.
- **Supports tabular and regression analyses**: Implements a local sensitivity approach to support a wide range of tabular and regression analyses.

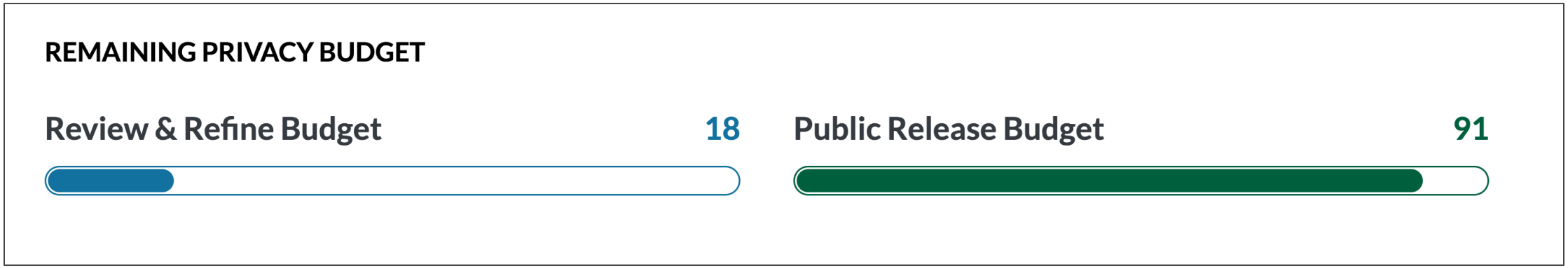
```r
# Arbitrary code ----------------------------------------------------
transformed_df <- conf_df %>%
    filter(AGE >= 18, AGE <= 65) %>%
    mutate(earned_income = INCWAGE + INCBUS + INCFARM)

# Analysis code -----------------------------------------------------
# Example regression
example_fit <- lm(earned_income ~ MARST + AGE, data = transformed_df)
example_model <- get_model_output(
    fit = example_fit,
    model_name = "Example Model"
)

# Example table
example_table <- get_table_output(
    data = transformed_df,
    stat = c("mean", "n"),
    var = "earned_income",
    by = "MARST",
    table_name = "Example Table",
)
```

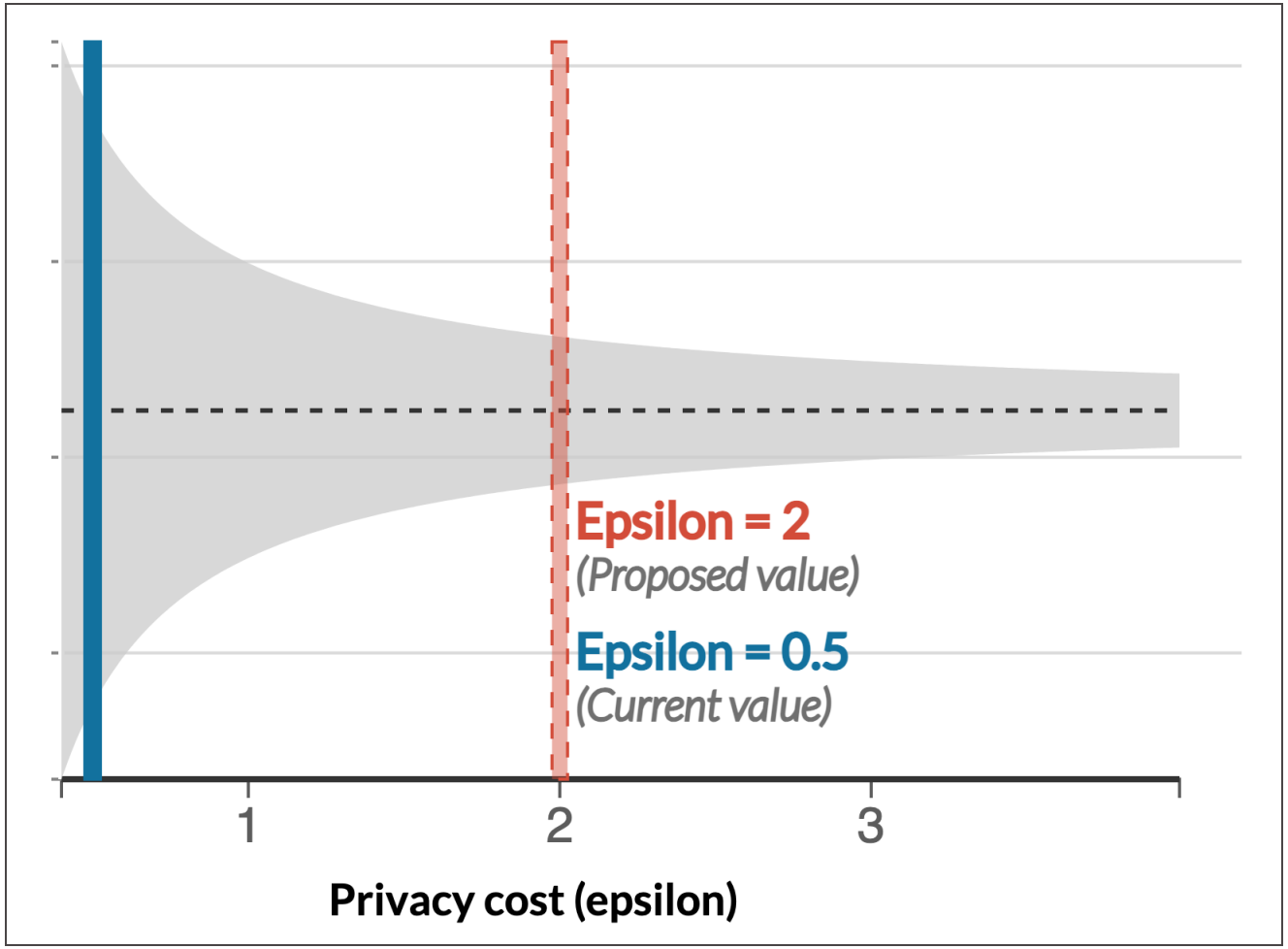
Implements Privacy Budgets to Manage Disclosure

- **Uses privacy budgets:** Researchers can *spend* from their limited privacy budgets to get more accurate results or produce more statistics.
- A *review and refine budget* allows for iteration within a secure environment.
  A *public release budget* controls results that can be published.



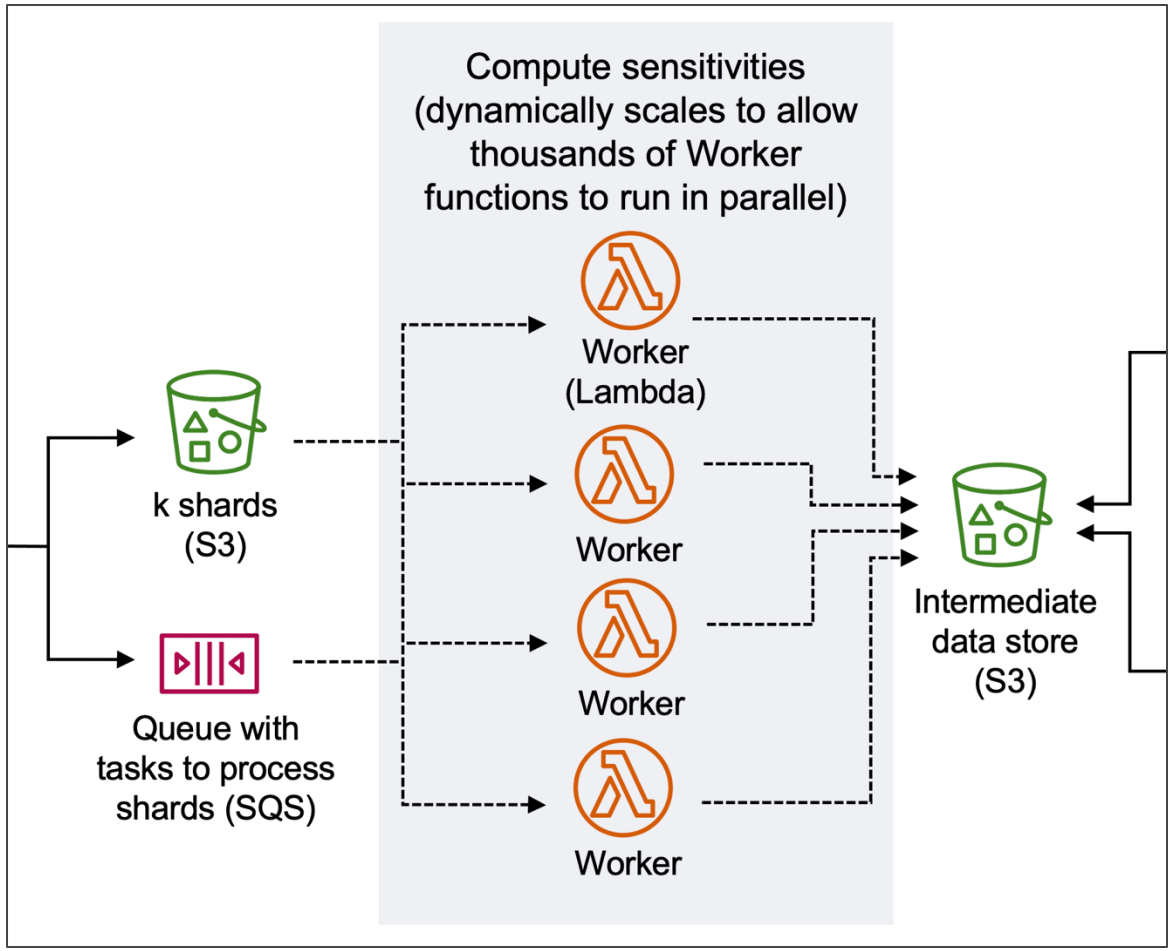
Displays Privacy & Usefulness Trade-offs

- **Helps users understand privacy concepts:** Displays an estimate of the 90th percentile of noise to help researchers identify appropriate epsilon values for their needs.



Uses Flexible, Scalable & Secure Technology

- **Designed to meet the needs of different data stewards**: Can accommodate different privacy algorithms, optional manual review steps, and other features.
- Implements a scalable back-end architecture in the AWS cloud with services that meet FISMA and FedRAMP compliance.



Future Challenges to Address

- Allow researchers to incorporate survey weights, join external datasets, and submit a wider range of input.
- Develop robust learning libraries and interfaces for researchers as well as other stakeholders such as data stewards.
- Appropriately display errors in user-submitted code.
- Ensure the correct amount of noise is added for complex analyses.
- Speed up time-intensive analyses on big datasets.

Upcoming Plans for Version 3.0

- Identify additional challenges for an automated validation server across the following categories:
  - Security & infrastructure
  - User experience
  - Data privacy
- Integrate the tool into a secure compute environment testbed to support the National Secure Data Service Demonstration project.
- Solicit feedback from various stakeholders to identify priorities and inform a future National Secure Data Service.

This prototype allows us to provide a testable solution to government agencies looking to improve and automate statistical disclosure control processes.

We hope that testing on a fully operational system, building trust with practitioners, continuously improving as the privacy field evolves, and disseminating our learnings will lead to increased access to valuable data and insights used to craft better public policy.

Contact Us & Learn More

safedatatech@urban.org

Validation Server Version 2.0 White Paper

Safe Data Technologies Project Landing Page