Privacy & Public Policy Conference | September 13, 2024

Validation Server: Version 2.0 Prototype

Safe Data Technologies Project: Safely Expanding Access to Administrative Data

URBAN INSTITUTE · ELEVATE · THE · DEBATE

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

Overview

1. Safe Data Technologies Project
2. Validation Server
   - Background
   - Version 2.0 Prototype
3. Challenges & Future Work

Safe Data Technologies Project Landing Page

Safe Data Technologies Project

Project Funding & Collaborators

ALFRED P. SLOAN
FOUNDATION

Project Goal

> *This body of work aims to safely expand access to confidential data that advances evidence-based policy making by creating new ways for researchers to use administrative data while protecting privacy.*

Project Framework

> *Our work is at the intersection of data privacy and public policy. We are implementing practical privacy-preserving technologies and tools (e.g., synthetic data generation) and exploring the feasibility of state-of-the-art methodologies (e.g., formal privacy) to provide better data access.*

Tiered Access for Administrative Tax Data

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

I. **Basic Access:** Researchers will have access to the synthetic public use file (PUF).

II. **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.

III. **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.

Tiered Access for Administrative Tax Data

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

I. **Basic Access:** Researchers will have access to the synthetic public use file (PUF).

II. **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.

III. **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.

Tiered Access for Administrative Tax Data: Validation Server

User Interface Layer

Data User

Data User

Data User

Analysis 1

Analysis 2

Analysis 3

Analysis 4

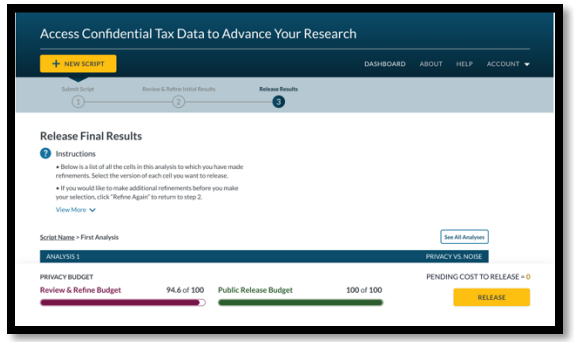
Analysis 5

Analysis 6

Analyses

Results

Privacy Layer



Apply privacy-preserving methodologies

Analysis 1.1

Analysis 1.2

noise

Analyses

Statistics

Data Layer

Analysis 1

Credit: Claire McKay Bowen

Validation Server Prototype

Validation Server: Prototype Development History

2020–2021
Built the first automated validation server prototype

2022–2024
Built the **version 2.0** prototype based on extensive feedback on the initial version

Feedback on initial version: Researchers wanted to submit a broader range of analyses (particularly regressions), use a more familiar programming language (such as R), and have more granular control over their privacy budgets.

Key Feature: Follows Normal Researcher Workflows

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

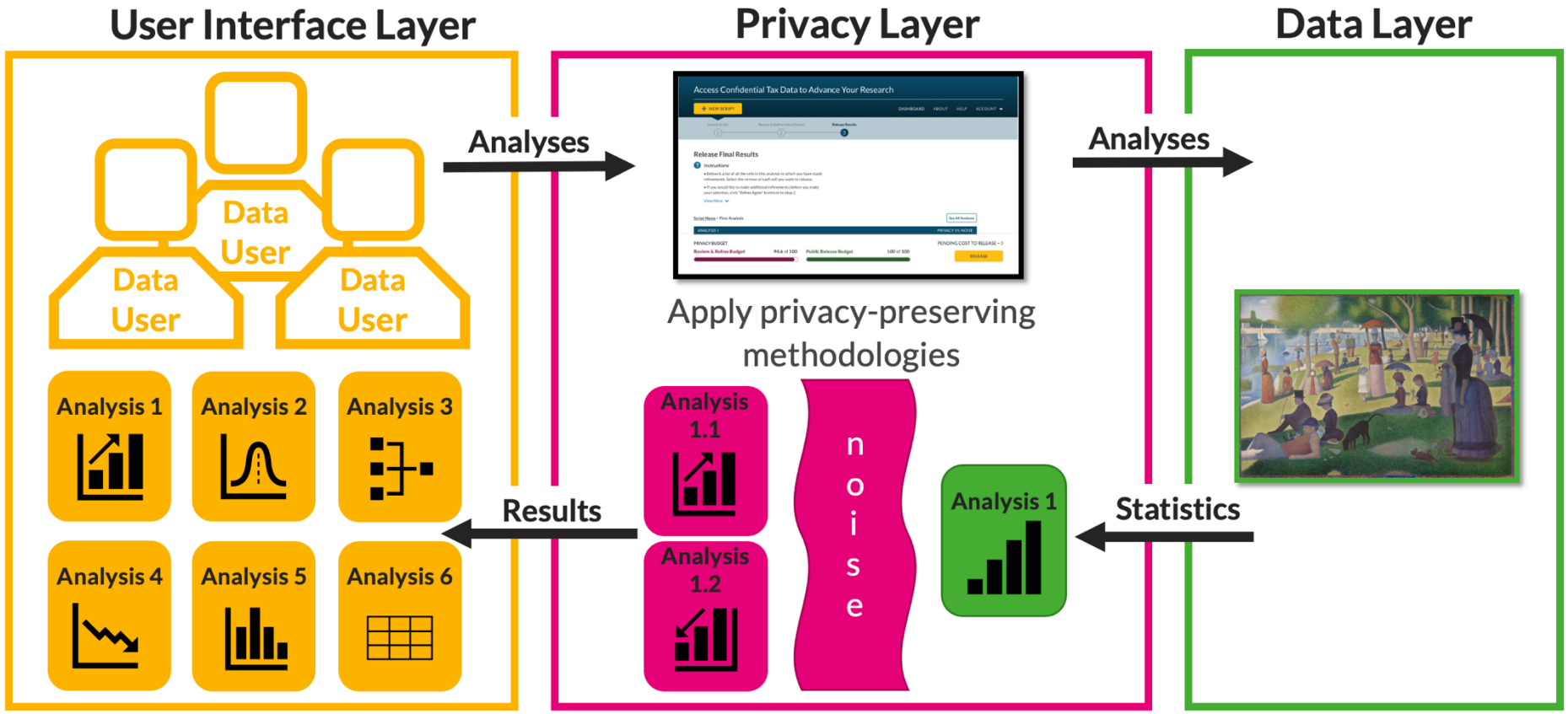
```r
# Arbitrary code ----------------------------------------------------
transformed_df <- conf_df %>%
    filter(AGE >= 18, AGE <= 65) %>%
    mutate(earned_income = INCWAGE + INCBUS + INCFARM)

# 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",
)
```

Key Feature: Automatically Adds Noise to Results

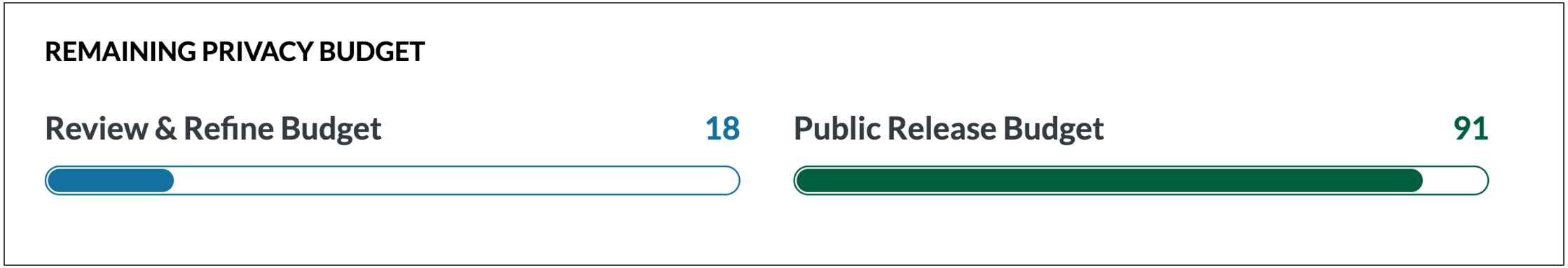
- **Reduces staff burden**: Doesn't require manual submission by agency staff with direct access to the confidential data.



Credit: Claire McKay Bowen

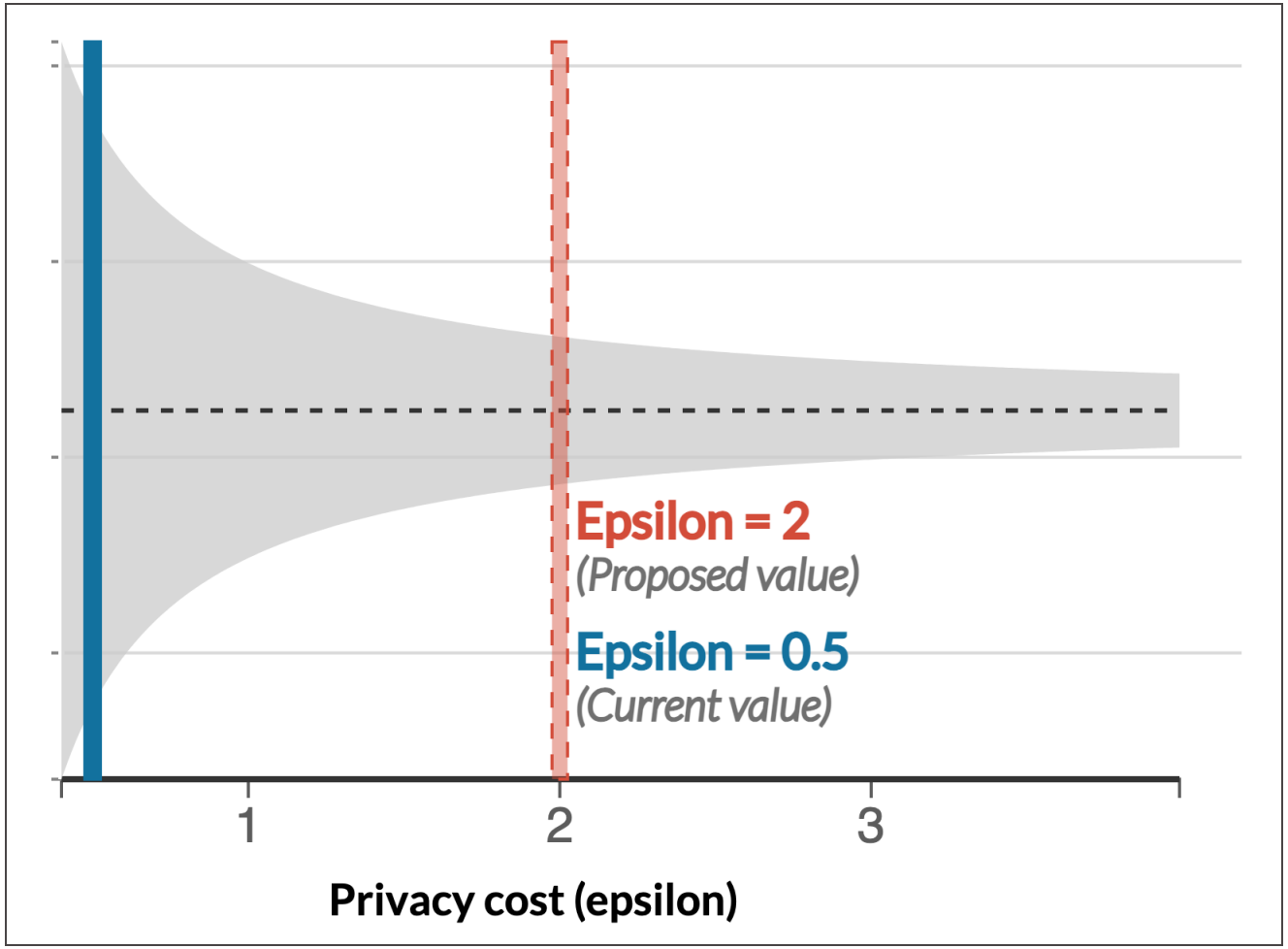
Key Feature: Uses Privacy Budgets to Manage Disclosure

- **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.



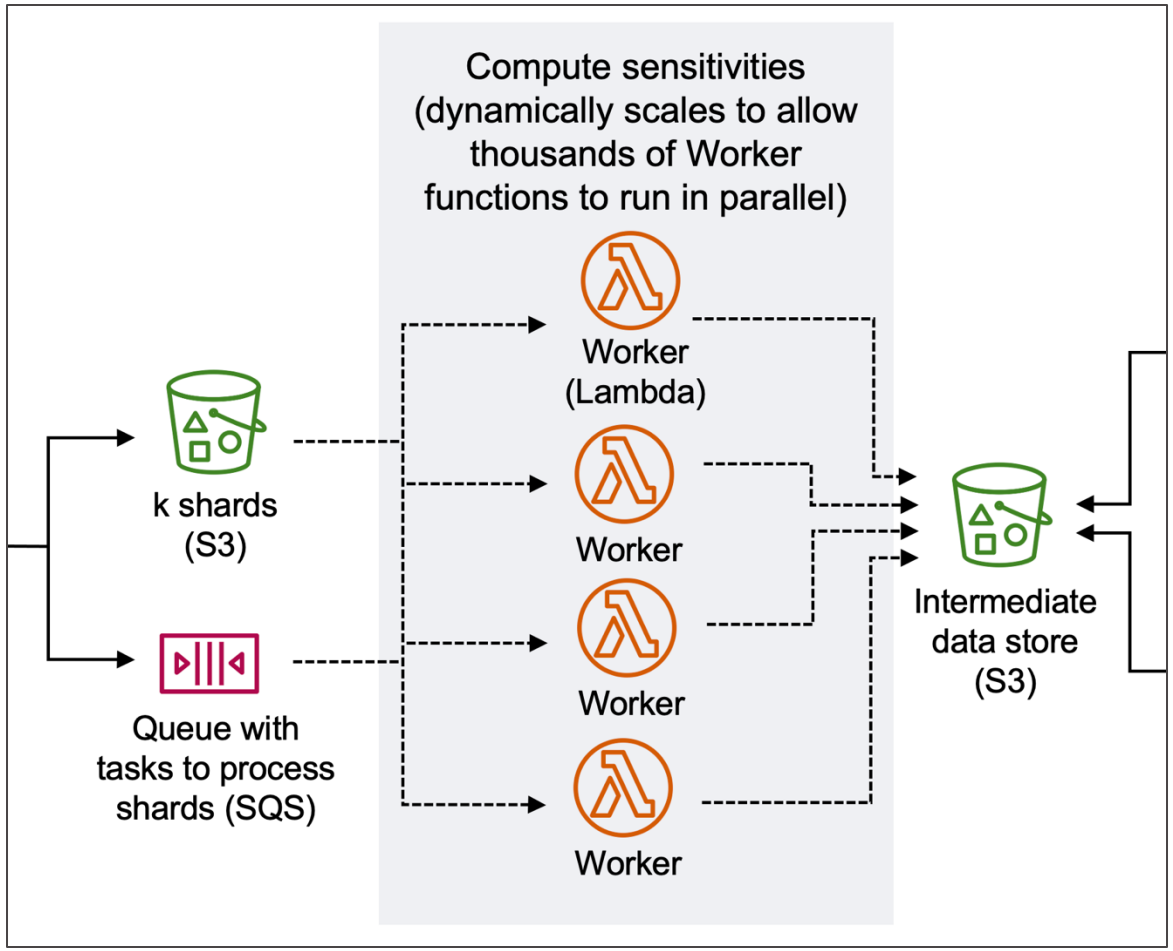
Key Feature: Displays Privacy & Usefulness Trade-off

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

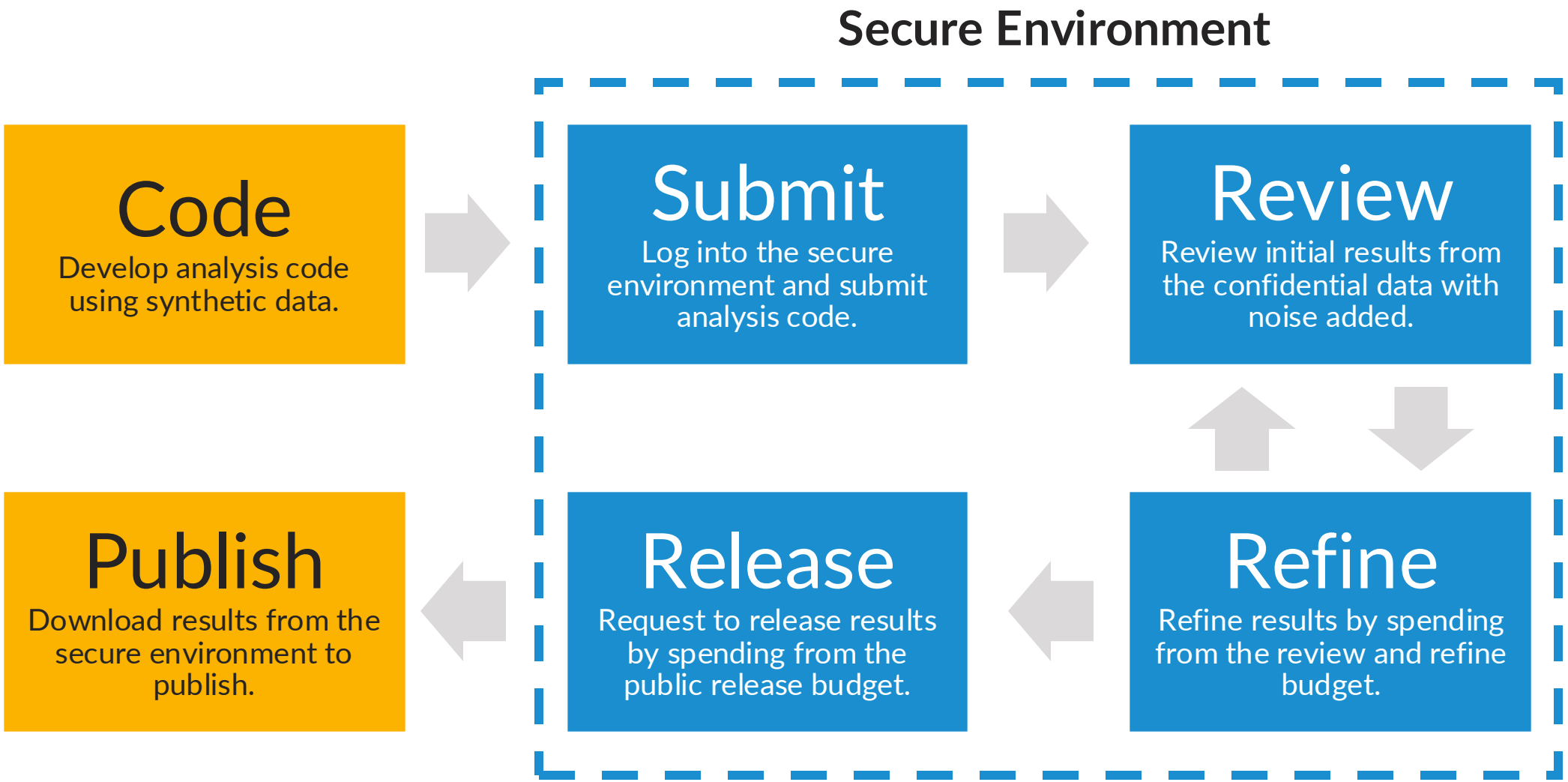


Key Feature: Uses Flexible, Scalable & Secure Technology

- **Meets the needs of different data stewards**: Built to accommodate different privacy algorithms, optional manual review steps, and other features.
- Implements a scalable, parallelized back-end architecture in the AWS cloud with services that comply with FedRAMP standards.



Researcher Perspective & Workflow



Challenges & Future Work

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
- 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