Safe Data Technologies Validation Server Prototype

Sybil Mendonca, Josh Miller, Jean Clayton Seraphin, Deena Tamaroff, Silke Taylor & Erika Tyagi

URBAN INSTITUTE · ELEVATE · THE · DEBATE

Thank you to our collaborators at IRS SOI and NSF NCSES

IRS

Research, Analysis & Statistics

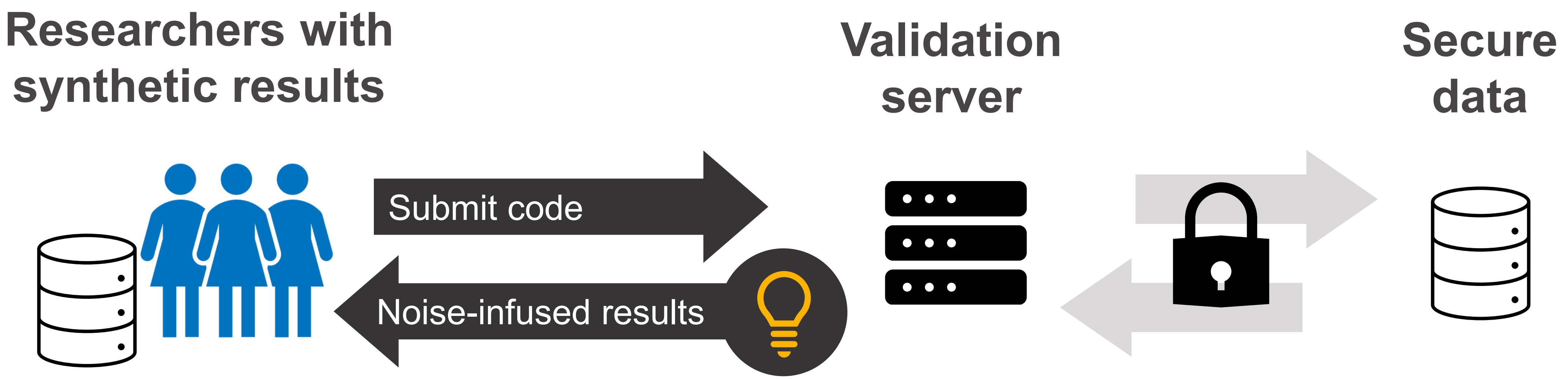
STATISTICS OF INCOME

NCSES

The Safe Data Technologies Project

Developing **innovative and practical tools** to safely expand access to confidential administrative data that advances evidence-based policymaking while protecting privacy.

- **Synthetic data** that represent the statistical properties of the data without revealing any individual taxpayer information.
- **A prototype validation server** that would allow researchers to perform statistical analyses on administrative data, using code that they develop using synthetic data, without revealing confidential information.



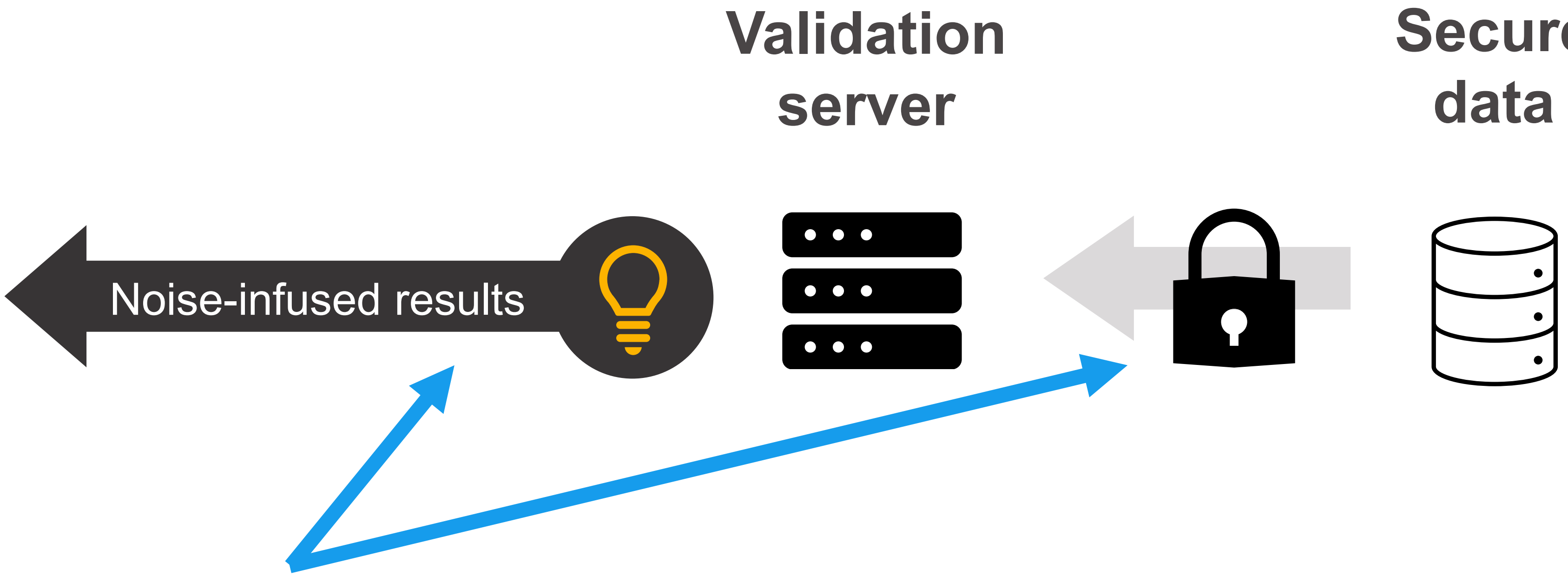
A validation server creates an intermediate layer between a researcher and the confidential data. With this intermediate layer, a researcher can **analyze confidential data without seeing them**.

Prototype Development History

- 2020-2021: Built the first automated validation server prototype.[1]
- 2022-2024: Built the **next generation prototype**.
  - Developed based on extensive user feedback on the initial prototype.
  - Finished developing a functional prototype in early 2024.
  - Plan to focus on dissemination and user testing for the rest of 2024 to help prioritize future improvements.

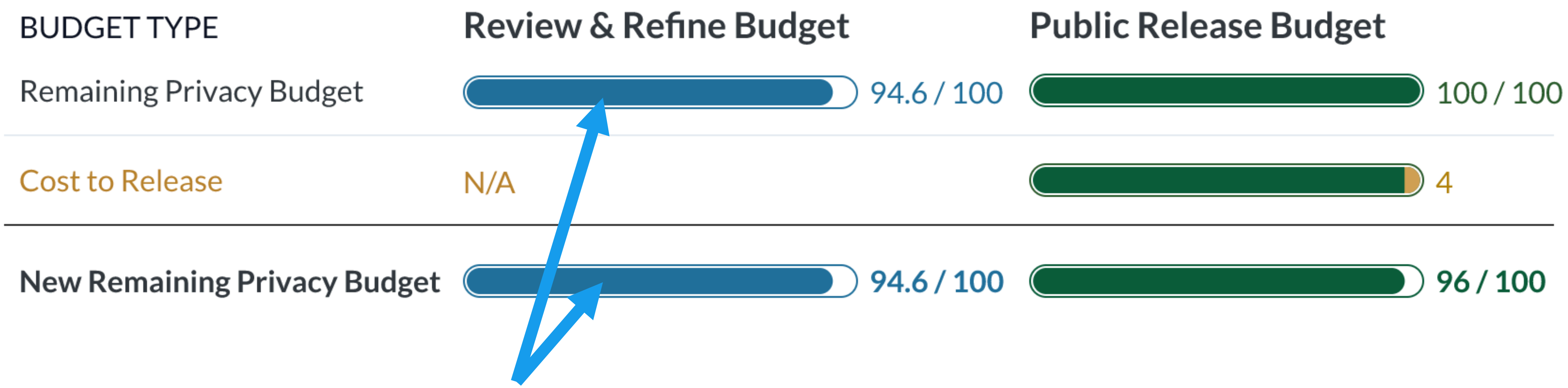
[1] See the technical white paper for a detailed overview of this prototype.

Key features of the validation server



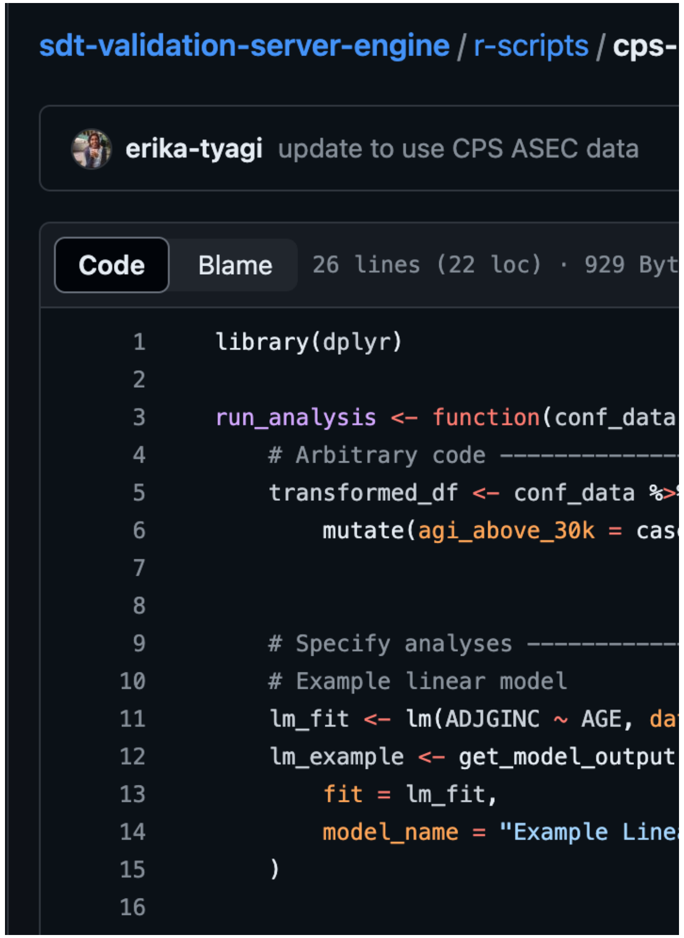
- **Automatically adds noise** to results to reduce staff burden.[1]

[1] This automated system differs from tools such as the U.S. Census Bureau's Synthetic Longitudinal Business Database (SynLBD), where agency staff manually review and validate results.



- Uses a **privacy budget mechanism** to automate the release process.
  - Researchers "spend" from a limited privacy budget to get more accurate results or produce more statistics.
  - A "Review & Refine" budget allows for iteration within a secure environment.
  - A "Public Release" budget controls results that can be published.

Key features of the validation server (cont.)



sdt-validation-server-engine / r-scripts / cps-

erika-tyagi update to use CPS ASEC data

Code Blame 26 lines (22 loc) · 929 Byt

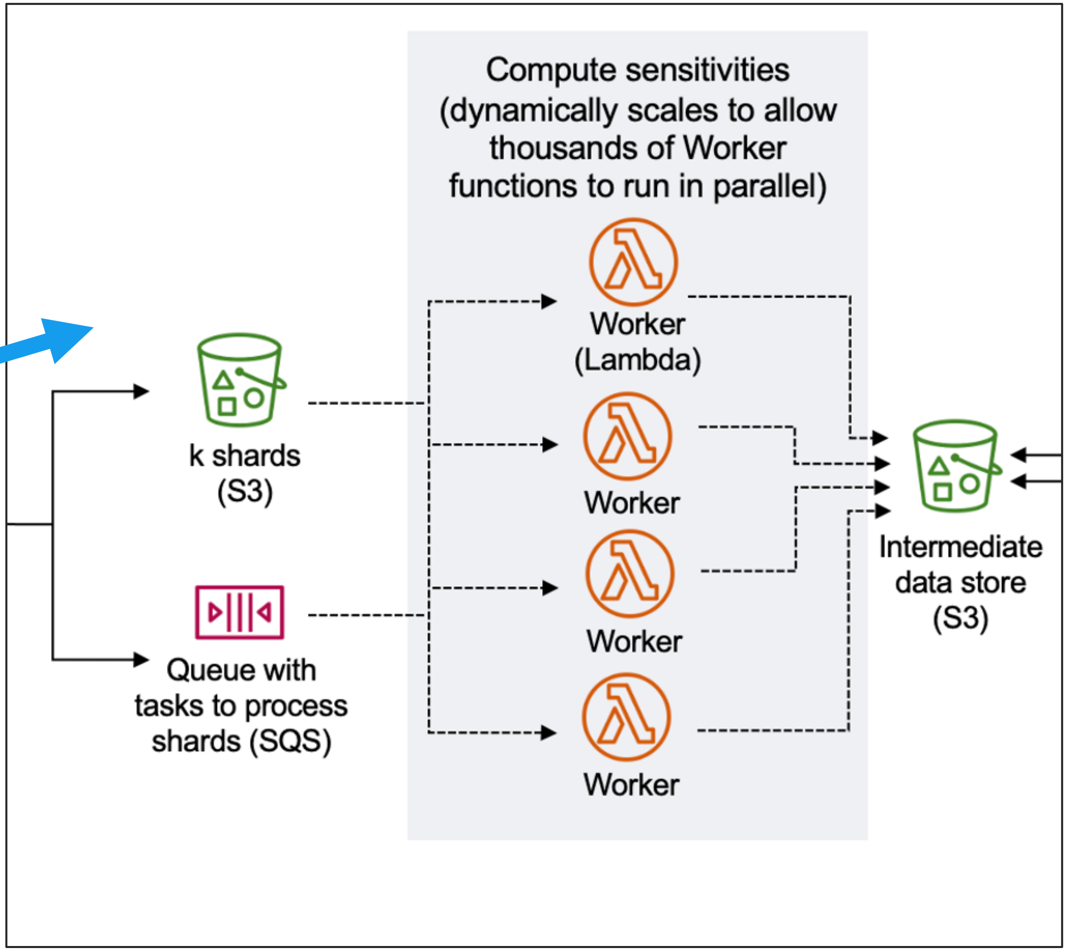
```r
library(dplyr)

run_analysis <- function(conf_data
    # Arbitrary code --------------
    transformed_df <- conf_data %>%
        mutate(agi_above_30k = cas


    # Specify analyses ------------
    # Example linear model
    lm_fit <- lm(ADJGINC ~ AGE, da
    lm_example <- get_model_output
        fit = lm_fit,
        model_name = "Example Line
    )

```

- Allows users to develop analyses using the R programming language and include pre-processing code to **mimic normal researcher workflows**.
- Supports a **wide range of tabular and regression analyses**.
- Implements a generalized version of the Maximum Observed Sensitivity (MOS) privacy algorithm that uses a **local sensitivity approach**.[1]

[1] The MOS algorithm was proposed by [Chetty and Friedman (2019)](#) in their work with the U.S. Census Bureau on the Opportunity Atlas.

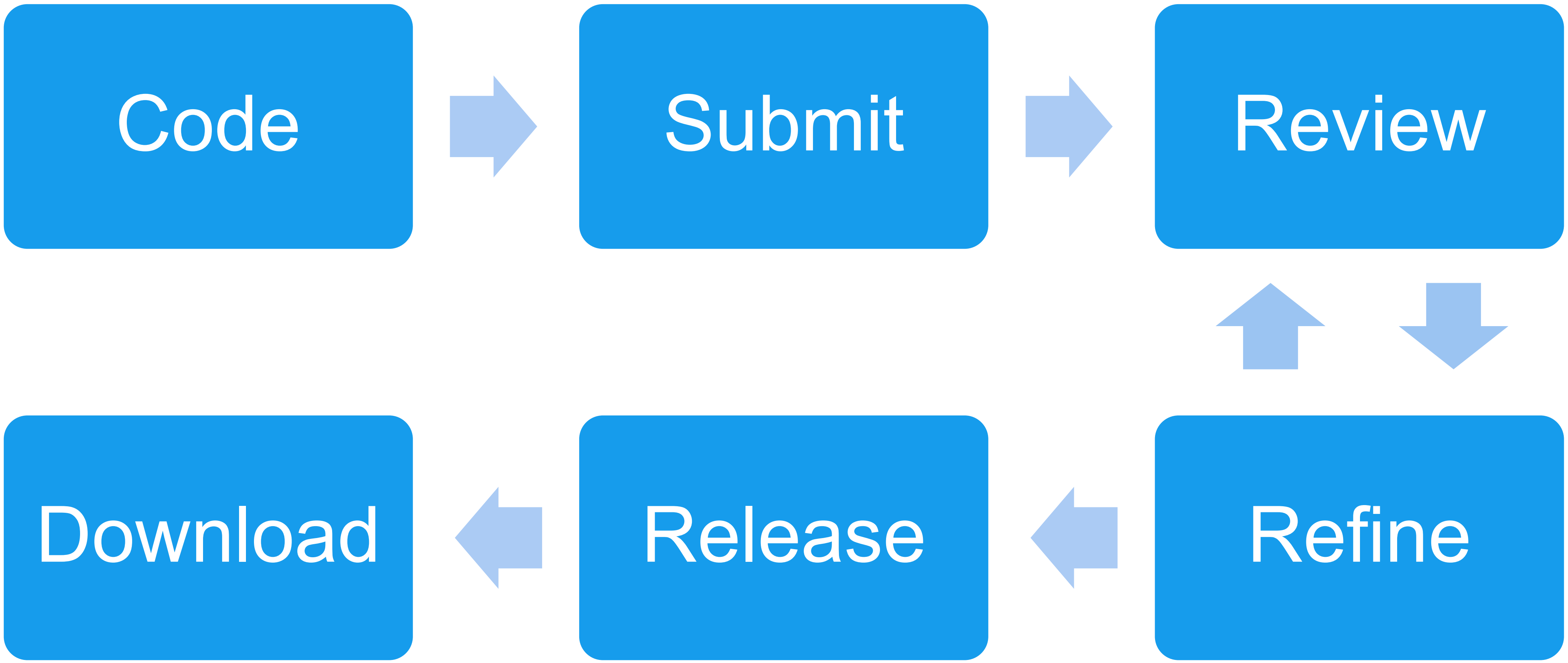
Key features of the validation server (cont.)

- Uses **public data** standing in for confidential data to facilitate user testing.
- Built using **scalable and secure** services in the AWS cloud that comply with the highest FedRAMP standards.



See the "Building the Prototype Backend" slide for full graphic

Validation server workflow



Demo

Future Challenges to Address

- Improve how errors in user-submitted analyses are reported. Errors can reveal sensitive information, but they are necessary to allow researchers to effectively debug code.
- Ensure the correct amount of noise is added for a given privacy budget for more complex statistical calculations.
- Speed up more complex, time-intensive analyses on big administrative datasets without compromising privacy.
- Balance algorithm improvements with the need for a simple interface that lets researchers interpret and interact with the privacy budget.

Upcoming Plans

- Disseminate to increase awareness of the prototype.
- Identify additional challenges for an automated validation server (including both infrastructure challenges and statistical data privacy challenges).
- Gather additional feedback to identify priorities to help inform a future National Secure Data Service.

Learn More & Contact Us

- Learn about the Safe Data Technologies project: www.urban.org/projects/safe-data-technologies
- Reach out to our team: safedatatech@urban.org

Questions and Answers

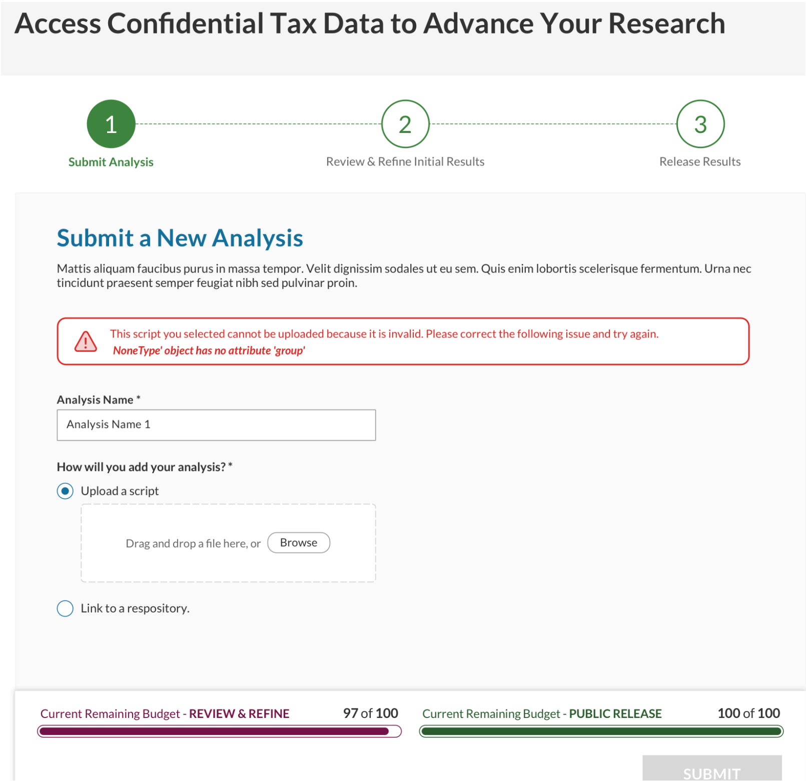
Appendix A:

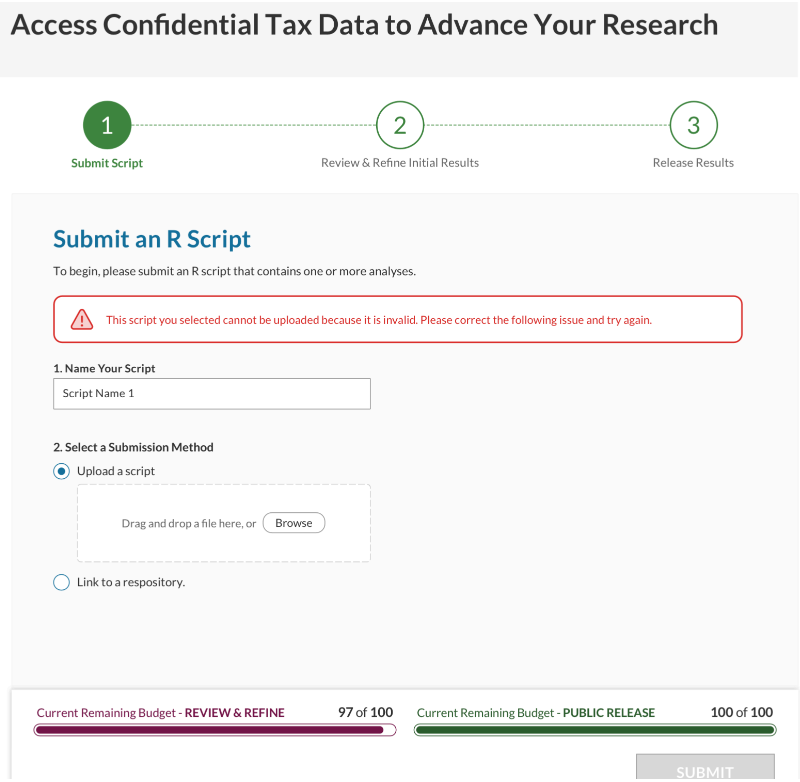
Development Process

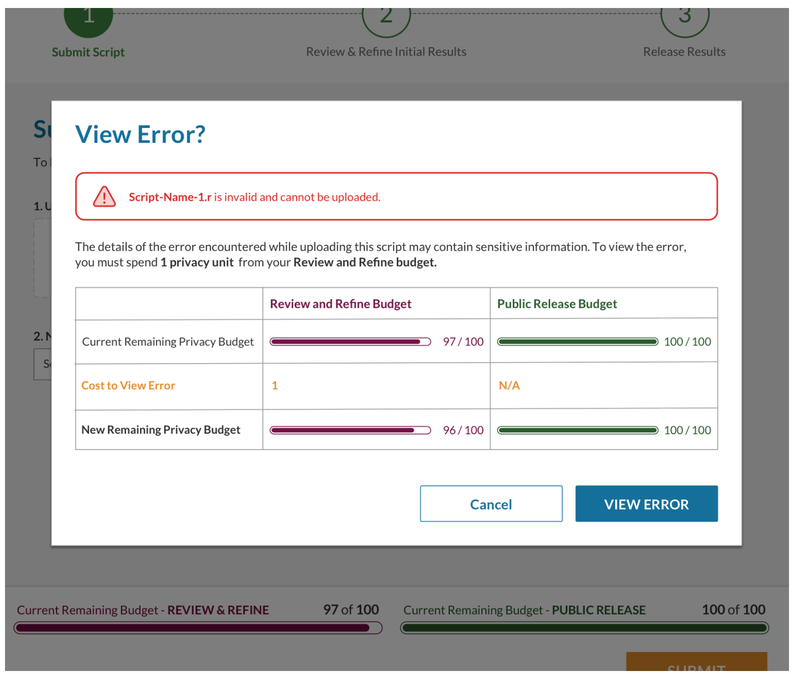
Building the Prototype Frontend

Using an iterative wireframing process informed by feedback sessions with users and stakeholders, we identified and devised solutions for several key challenges in the interface.

Challenge 1: How much information to reveal in error messages when uploading scripts?



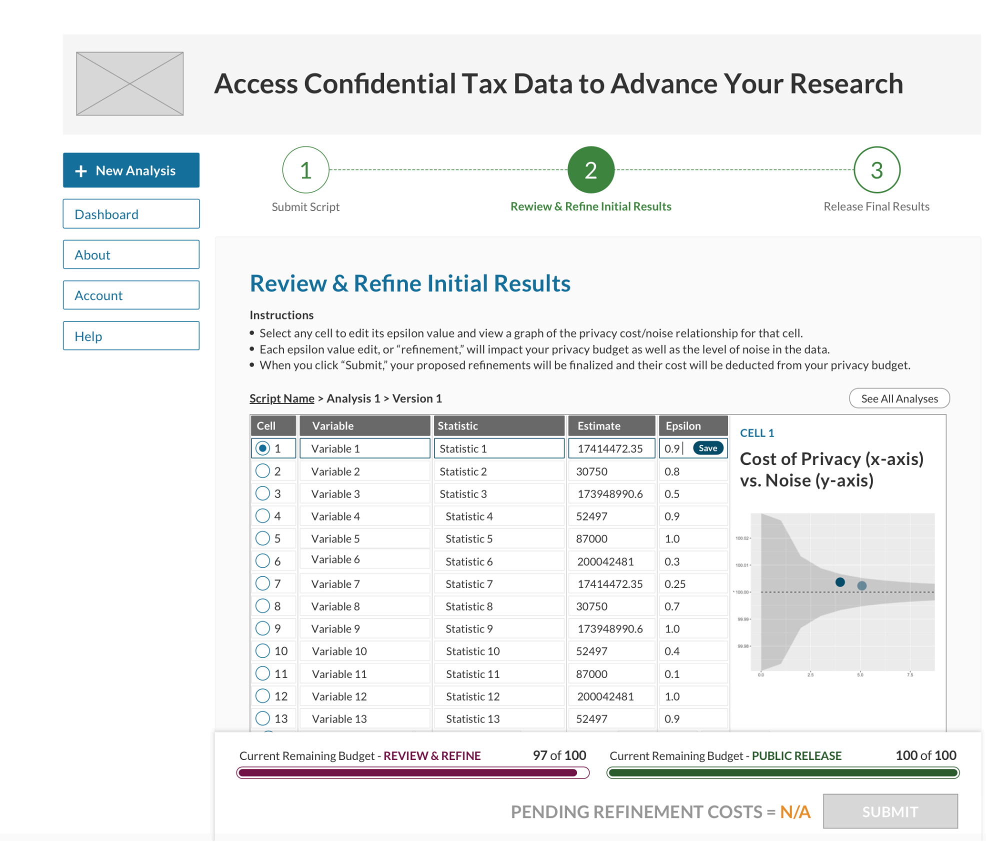


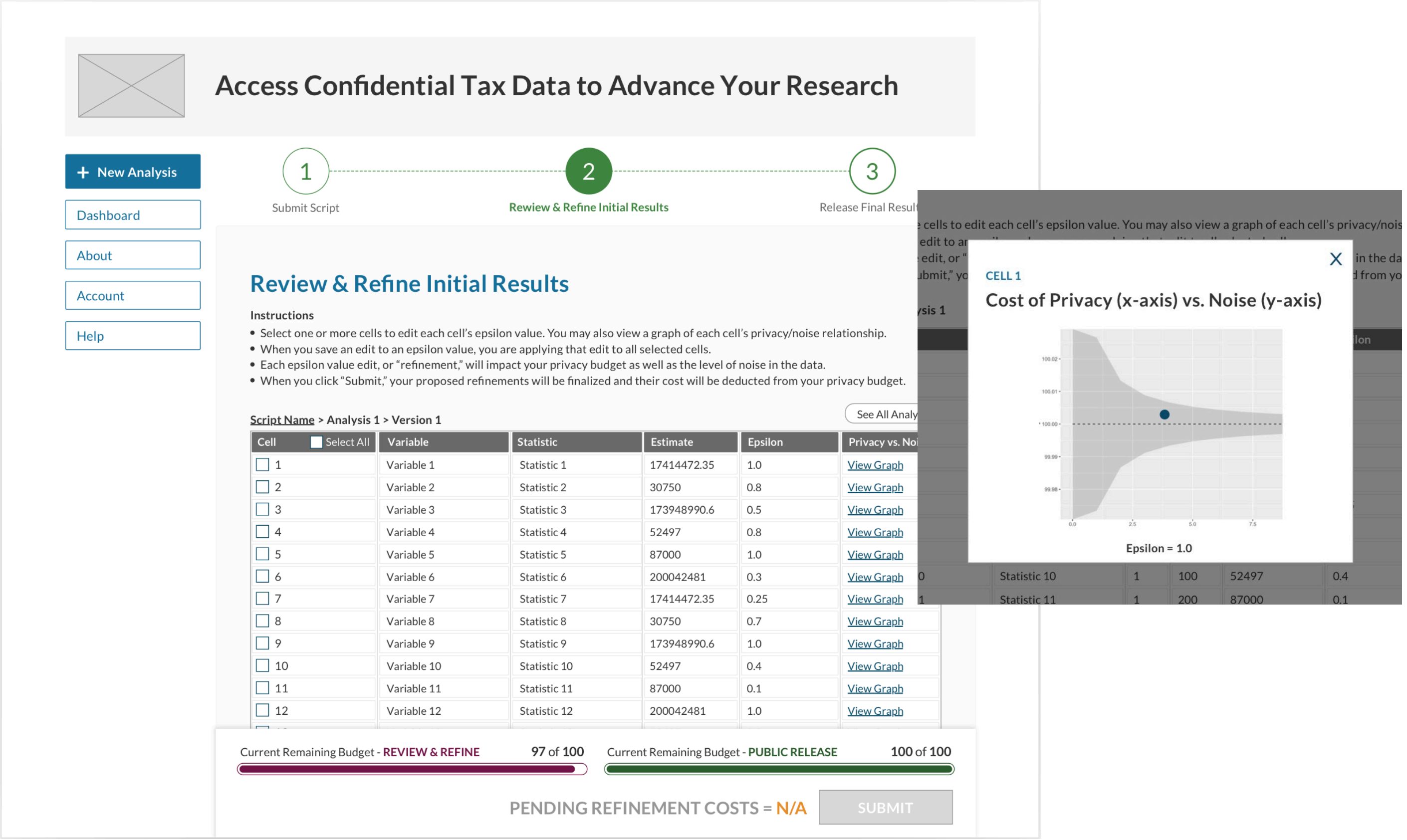


Building the Prototype Frontend (cont.)

Using an iterative wireframing process informed by feedback sessions with users and stakeholders, we identified and devised solutions for several key challenges in the interface.

Challenge 2: How to best visually communicate the privacy-noise tradeoff?

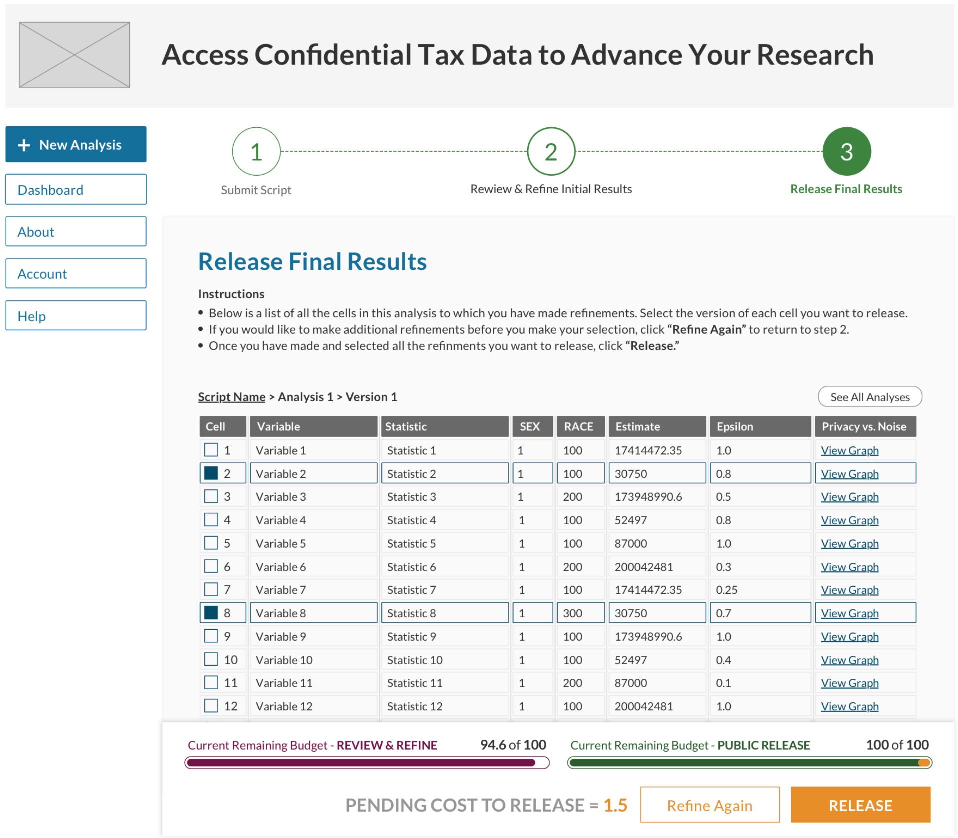


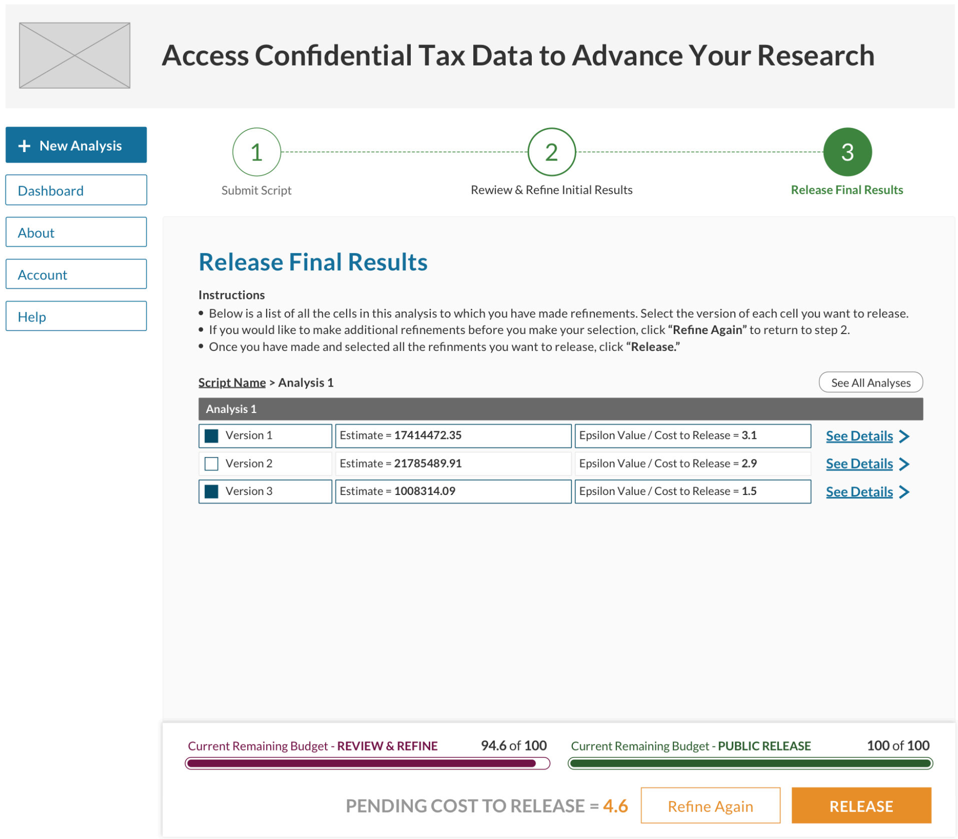


Building the Prototype Frontend (cont.)

Using an iterative wireframing process informed by feedback sessions with users and stakeholders, we identified and devised solutions for several key challenges in the interface.

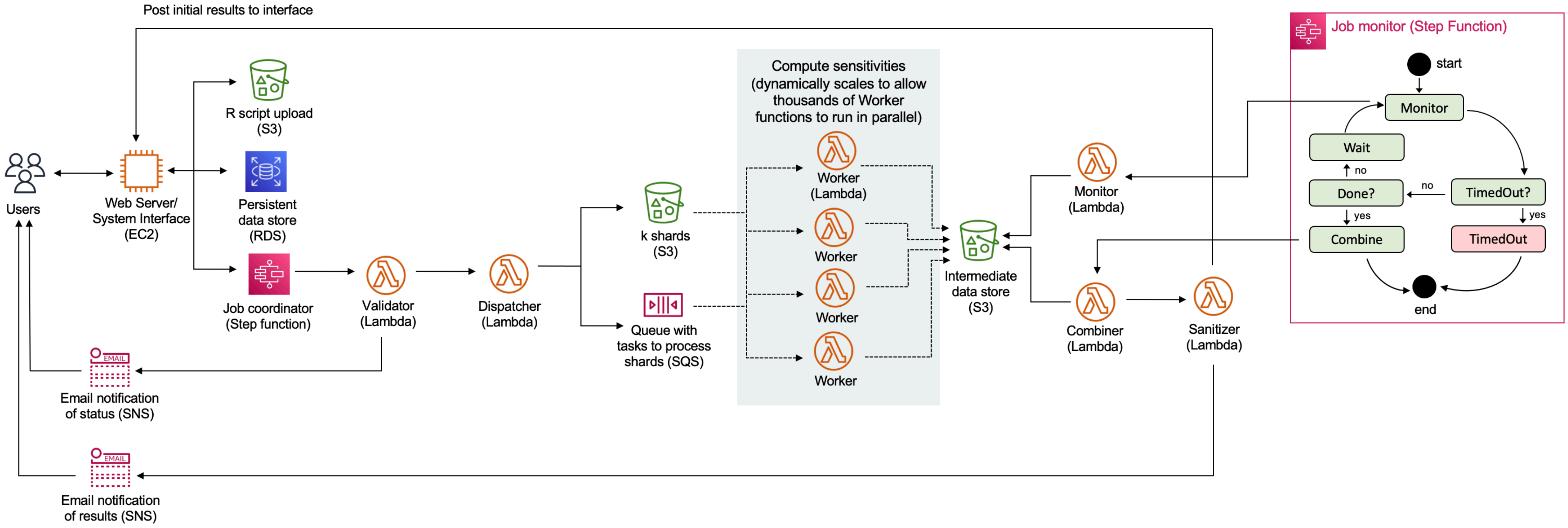
Challenge 3: Should users release results at the cell level or analysis level?





Building the Prototype Backend

We built the backend infrastructure in the Amazon Web Services (AWS) cloud using services that are FedRAMP High compliant using a scalable, secure, and cost-efficient serverless architecture.

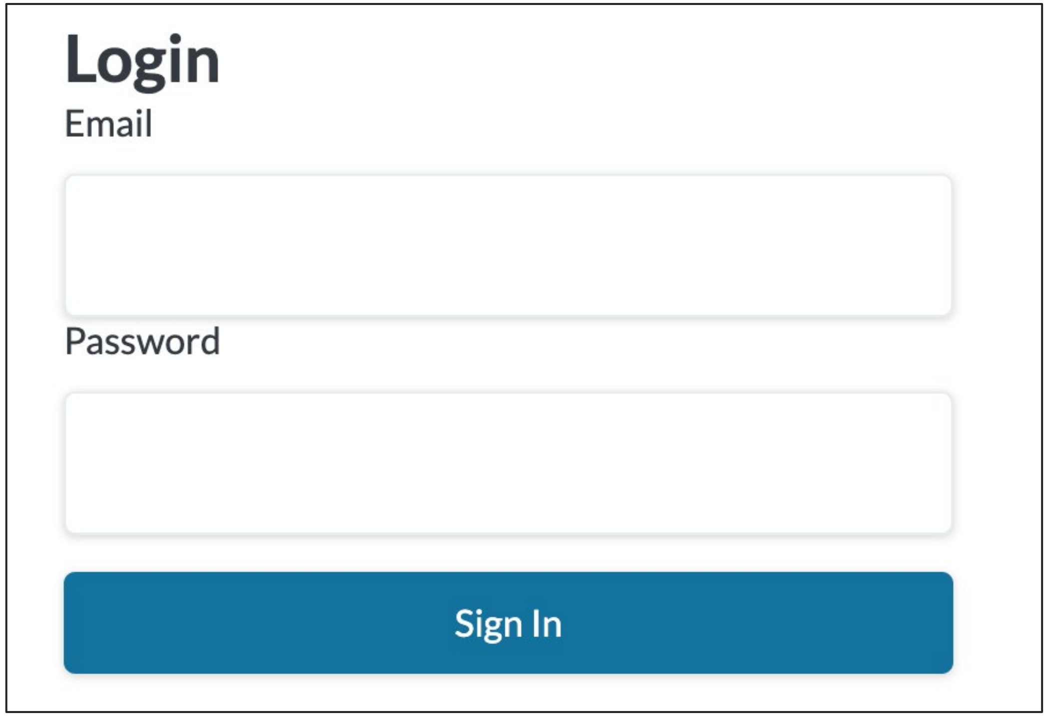


Appendix B:

How a user interacts with the prototype

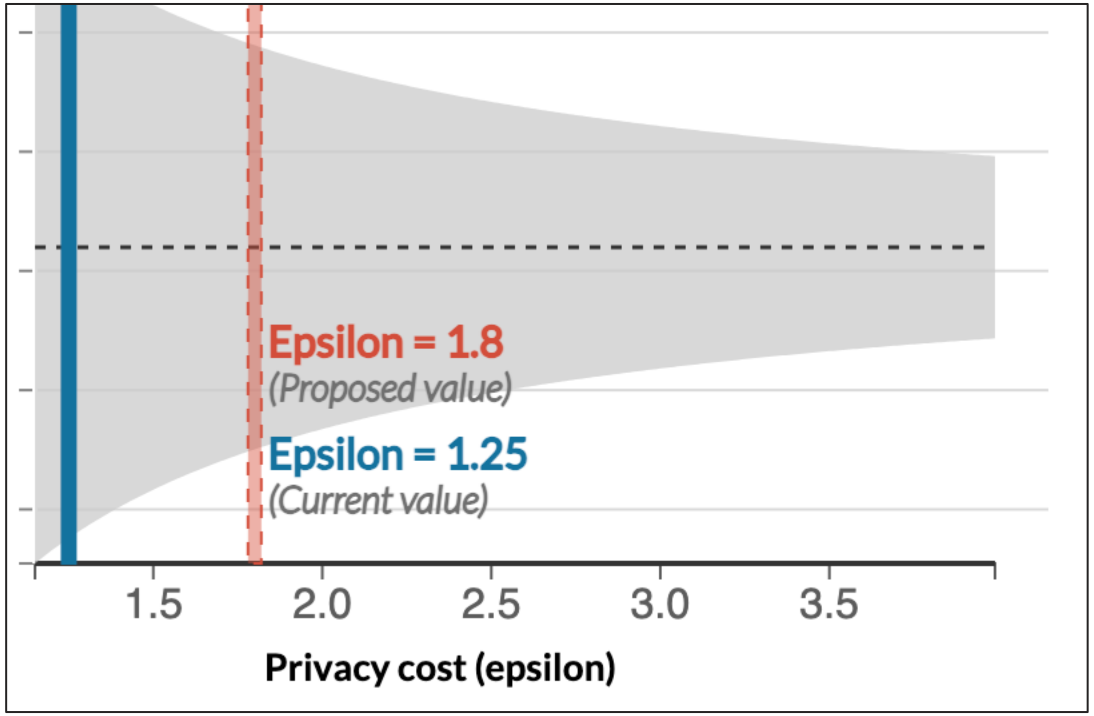
How a User Interacts with the Prototype

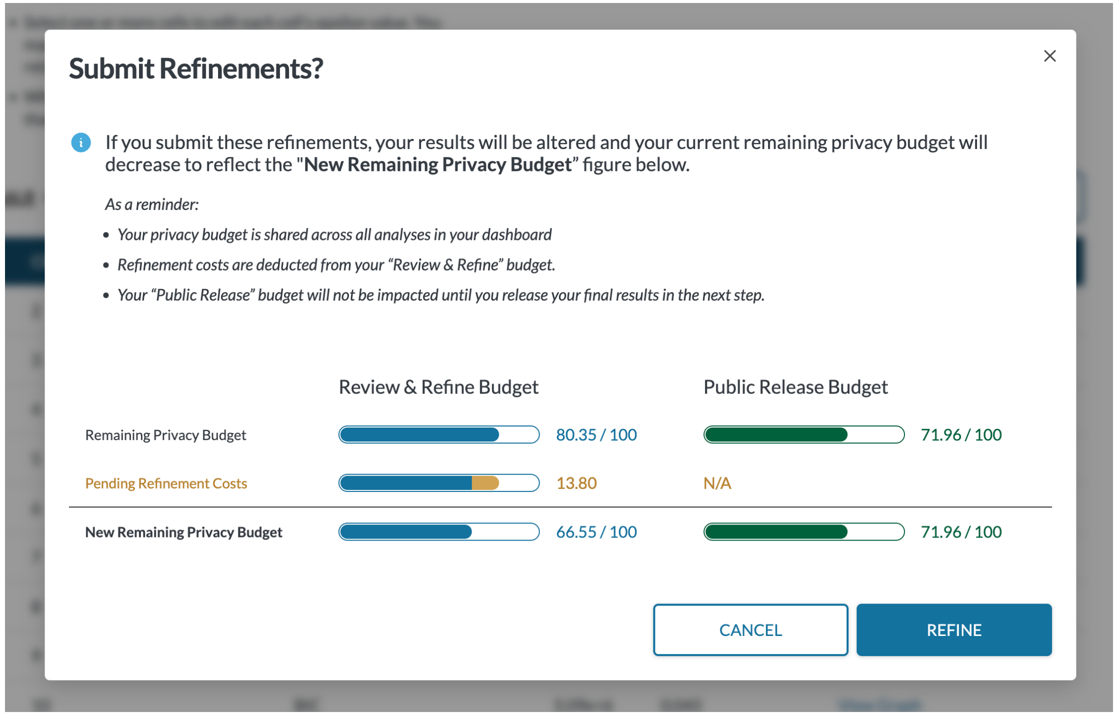
```r
    # Specify analyses ------------------------------------------------------------
    # Example linear model
    lm_fit <- lm(ADJGINC ~ AGE, data = transformed_df)
    lm_example <- get_model_output(
        fit = lm_fit,
        model_name = "Example Linear Model"
    )

    # Example binomial model
    glm_fit <- glm(agi_above_30k ~ AGE, family = binomial, data = transformed_df)
    glm_example <- get_model_output(
        fit = glm_fit,
        model_name = "Example Binomial Model"
    )

    # Submit analyses -------------------------------------------------------------
    submit_output(lm_example, glm_example)
}
```



1. Develop an R script using synthetic data.
2. Log into the validation server interface and submit the R script.

How a User Interacts with the Prototype (cont.)

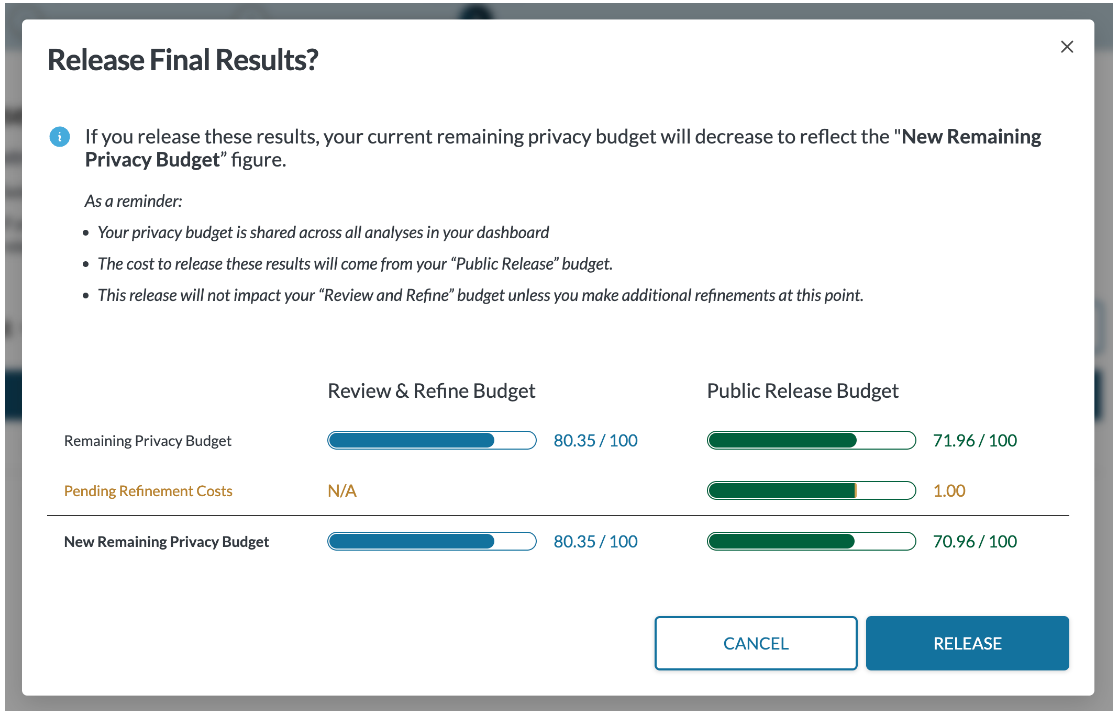




3. Review results with graphs that show the tradeoff between privacy and noise.

4. Refine results by spending from a "Review & Refine" budget.

How a User Interacts with the Prototype (cont.)



5a514c7bdbf5466995076a48652de1ad-run1-release

var	statistic	analysis_type	analysis_name	chi	epsilon	noise_90	value_
	r.squared	model	Example Model	2.366034846566010	0.045454545454545500	3.32604888752963E-05	0.000
	adj.r.squared	model	Example Model	2.386398874382060	0.045454545454545500	3.35467558005763E-05	0.000
	sigma	model	Example Model	43112656.13013590	0.045454545454545500	606.0553257201810	1
	statistic	model	Example Model	402.0018850646160	0.045454545454545500	0.00565113368699780	4
	p.value	model	Example Model	58.62287490583350	0.045454545454545500	0.000824089924741928	-0.00
	df	model	Example Model	0.0	0.045454545454545500	0.0	
	logLik	model	Example Model	17105.069909996900	0.045454545454545500	0.24045418784864200	-
	AIC	model	Example Model	34210.139819993800	0.045454545454545500	0.4809083756972850	
	BIC	model	Example Model	34213.148784154900	0.045454545454545500	0.4809506741525050	
	deviance	model	Example Model	7319115355249538.0	0.045454545454545500	102888321870.49400	2.02
	df.residual	model	Example Model	169.0	0.045454545454545500	0.002375714215740840	1
	nobs	model	Example Model	169.0	0.045454545454545500	0.002375714215740840	
(Intercept)	estimate	model	Example Model	19028052.26742870	0.045454545454545500	267.4864749679880	2
(Intercept)	std.error	model	Example Model	8838380.858354190	0.045454545454545500	124.2453671557620	
(Intercept)	statistic	model	Example Model	272.3166854783260	0.045454545454545500	0.003828086514049040	
(Intercept)	p.value	model	Example Model	125.63341798264300	0.045454545454545500	0.001766089331795720	-0.00
MARS	estimate	model	Example Model	13949136.060709000	0.045454545454545500	196.08970909307400	
MARS	std.error	model	Example Model	4430335.997351510	0.045454545454545500	62.279362185895600	
MARS	statistic	model	Example Model	159.549076163222	0.045454545454545500	0.0022428580375697300	

5. Request to release results by spending from a "Public Release" budget.
6. Download results from the interface that can be published.