

Safe Data Technologies Validation Server Prototype

Thank you to our collaborators at IRS SOI and NSF NCSES



Research, Analysis & Statistics

STATISTICS OF INCOME

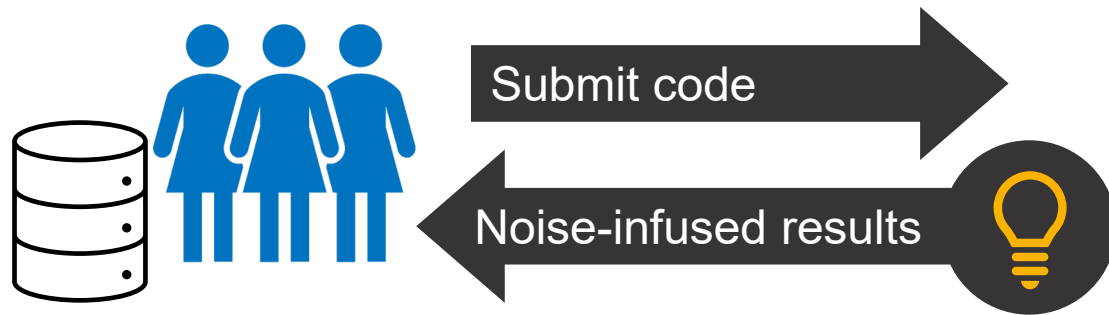


The Safe Data Technologies Project

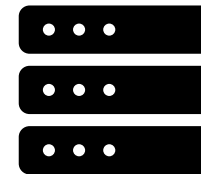
Developing **innovative and practical tools** to safely expand access to confidential administrative data that advances evidence-based policymaking while protecting privacy.

- **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.

Researchers with synthetic results



Validation server



Secure data



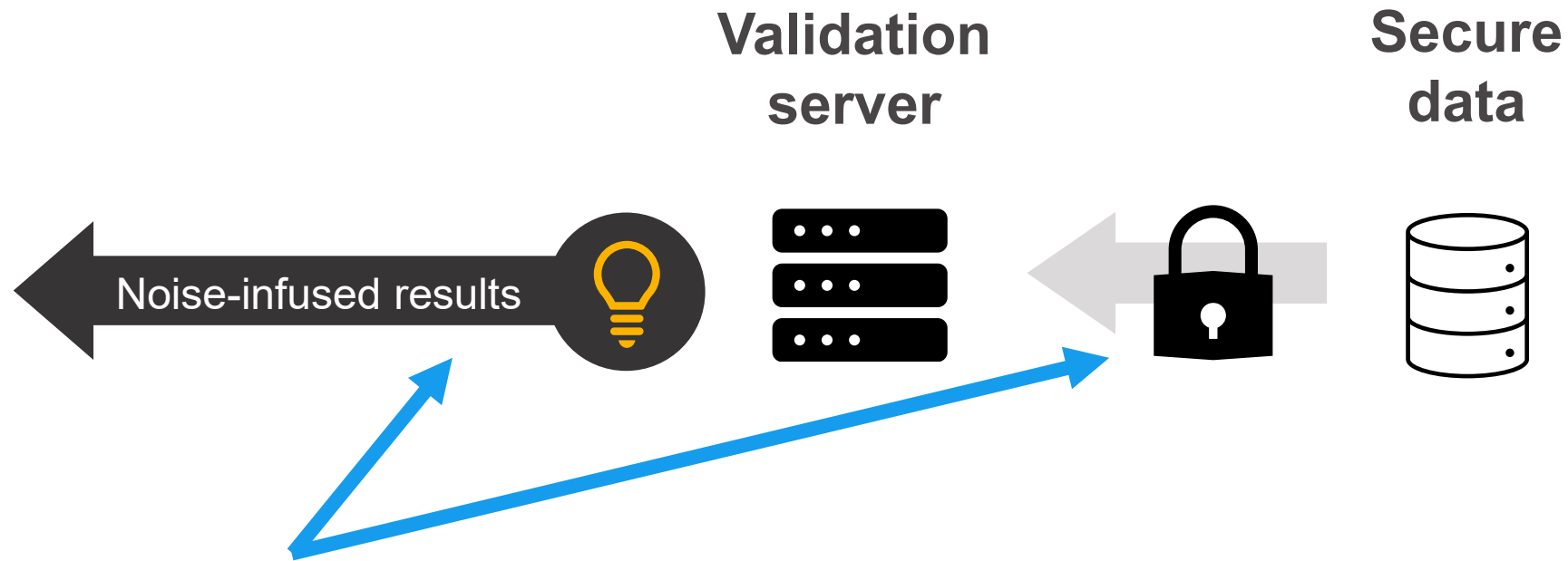
A validation server creates an intermediate layer between a researcher and the confidential data. With this intermediate layer, a researcher can **analyze confidential data without seeing them.**

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.

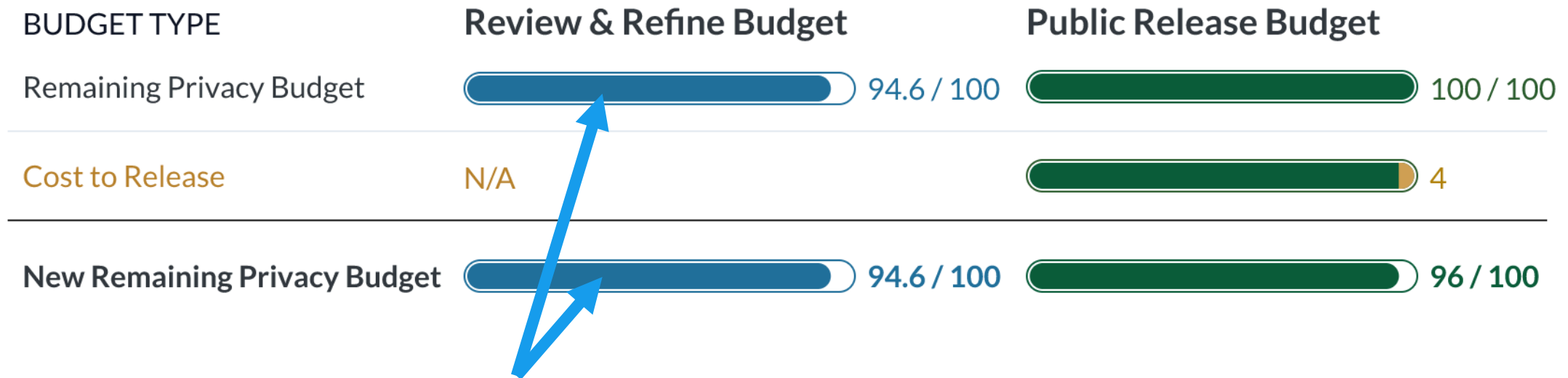
[1] See the [technical white paper](#) for a detailed overview of this prototype.

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.)

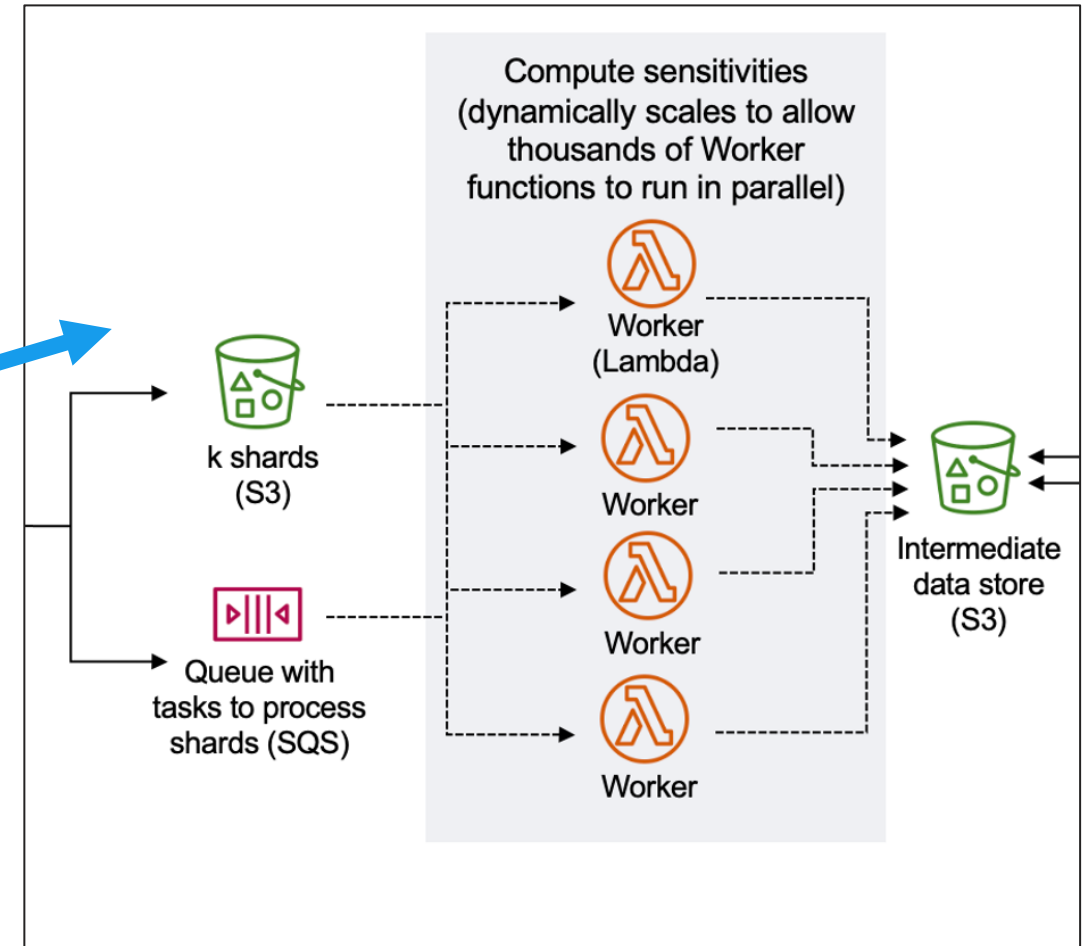
```
sdt-validation-server-engine / r-scripts / cps-  
erika-tyagi update to use CPS ASEC data  
Code Blame 26 lines (22 loc) · 929 Byt  
1 library(dplyr)  
2  
3 run_analysis <- function(conf_data  
4   # Arbitrary code -----  
5   transformed_df <- conf_data %>  
6     mutate(agi_above_30k = cas  
7  
8  
9   # Specify analyses -----  
10  # Example linear model  
11  lm_fit <- lm(ADJGINC ~ AGE, da  
12  lm_example <- get_model_output  
13    fit = lm_fit,  
14    model_name = "Example Line  
15  )  
16
```

- 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.

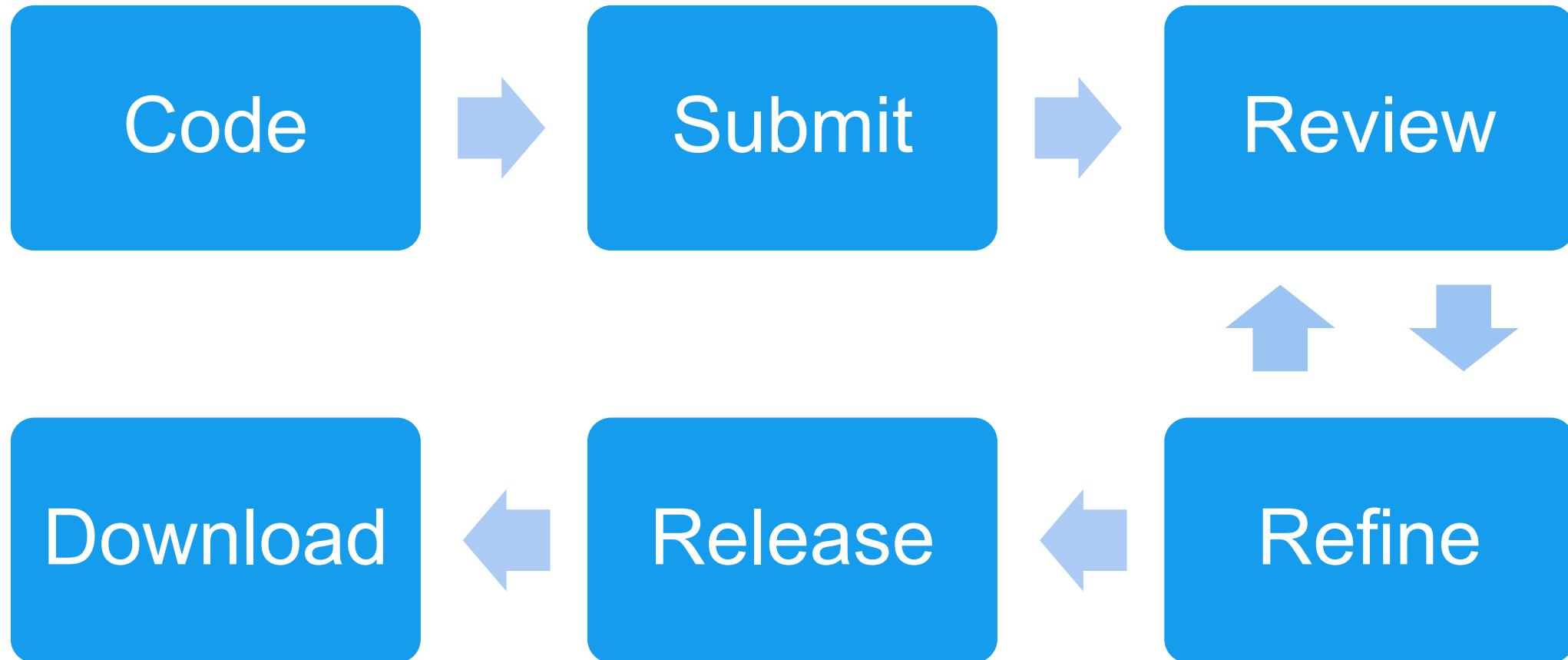
Key features of the validation server (cont.)

- 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

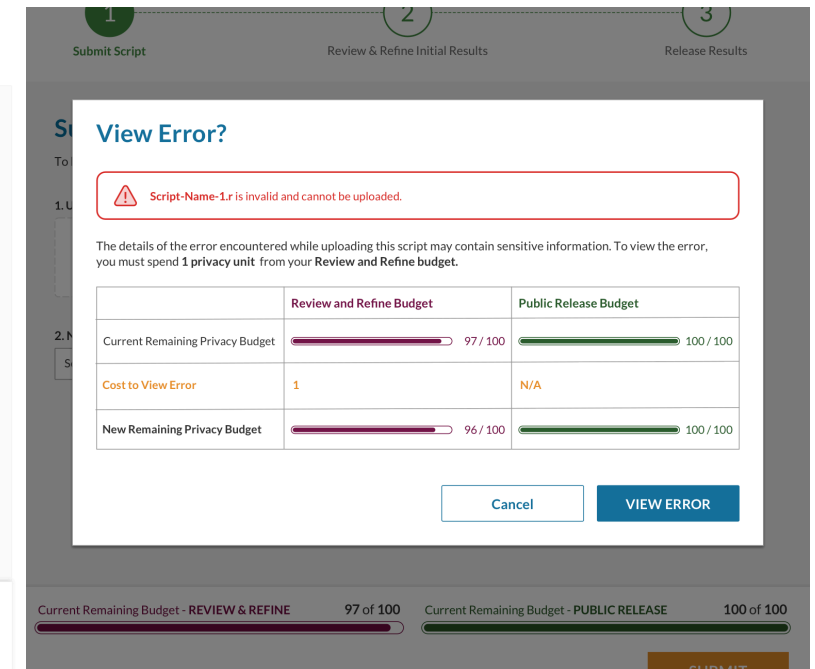
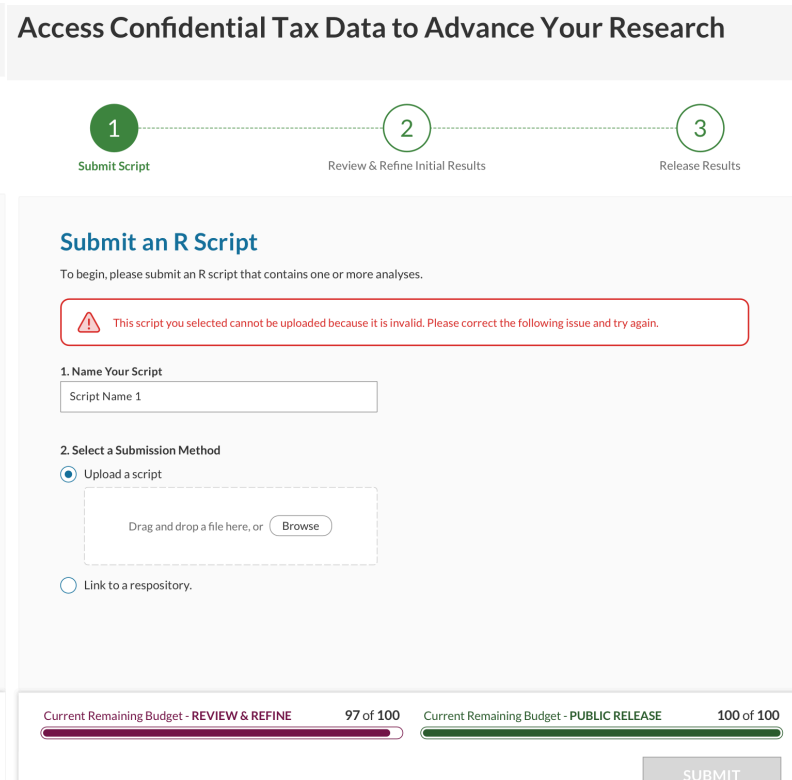
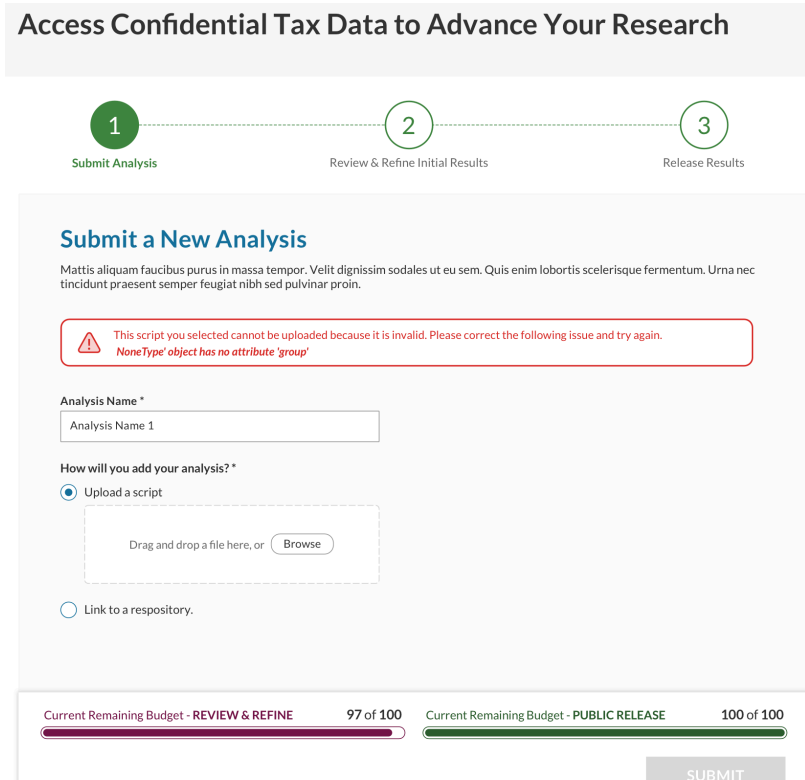
Questions and Answers

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?



Access Confidential Tax Data to Advance Your Research

+ New Analysis

Dashboard

About

Account

Help

- 1 Submit Script
- 2 Review & Refine Initial Results
- 3 Release Final Results

Release Final Results

Instructions


- Below is a list of all the cells in this analysis to which you have made refinements. Select the version of each cell you want to release.
- If you would like to make additional refinements before you make your selection, click "Refine Again" to return to step 2.
- Once you have made and selected all the refinements you want to release, click "Release."

Script Name > Analysis 1 > Version 1 [See All Analyses](#)

Cell	Variable	Statistic	SEX	RACE	Estimate	Epsilon	Privacy vs. Noise
<input type="checkbox"/> 1	Variable 1	Statistic 1	1	100	17414472.35	1.0	View Graph
<input checked="" type="checkbox"/> 2	Variable 2	Statistic 2	1	100	30750	0.8	View Graph
<input type="checkbox"/> 3	Variable 3	Statistic 3	1	200	173948990.6	0.5	View Graph
<input type="checkbox"/> 4	Variable 4	Statistic 4	1	100	52497	0.8	View Graph
<input type="checkbox"/> 5	Variable 5	Statistic 5	1	100	87000	1.0	View Graph
<input type="checkbox"/> 6	Variable 6	Statistic 6	1	200	200042481	0.3	View Graph
<input type="checkbox"/> 7	Variable 7	Statistic 7	1	100	17414472.35	0.25	View Graph
<input checked="" type="checkbox"/> 8	Variable 8	Statistic 8	1	300	30750	0.7	View Graph
<input type="checkbox"/> 9	Variable 9	Statistic 9	1	100	173948990.6	1.0	View Graph
<input type="checkbox"/> 10	Variable 10	Statistic 10	1	100	52497	0.4	View Graph
<input type="checkbox"/> 11	Variable 11	Statistic 11	1	200	87000	0.1	View Graph
<input type="checkbox"/> 12	Variable 12	Statistic 12	1	100	200042481	1.0	View Graph

Current Remaining Budget - REVIEW & REFINE 94.6 of 100 Current Remaining Budget - PUBLIC RELEASE 100 of 100

PENDING COST TO RELEASE = 1.5 [Refine Again](#) [RELEASE](#)



Access Confidential Tax Data to Advance Your Research

+ New Analysis

Dashboard

About

Account

Help

- 1 Submit Script
- 2 Review & Refine Initial Results
- 3 Release Final Results

Release Final Results

Instructions

- Below is a list of all the cells in this analysis to which you have made refinements. Select the version of each cell you want to release.
- If you would like to make additional refinements before you make your selection, click "Refine Again" to return to step 2.
- Once you have made and selected all the refinements you want to release, click "Release."

Script Name > Analysis 1 [See All Analyses](#)

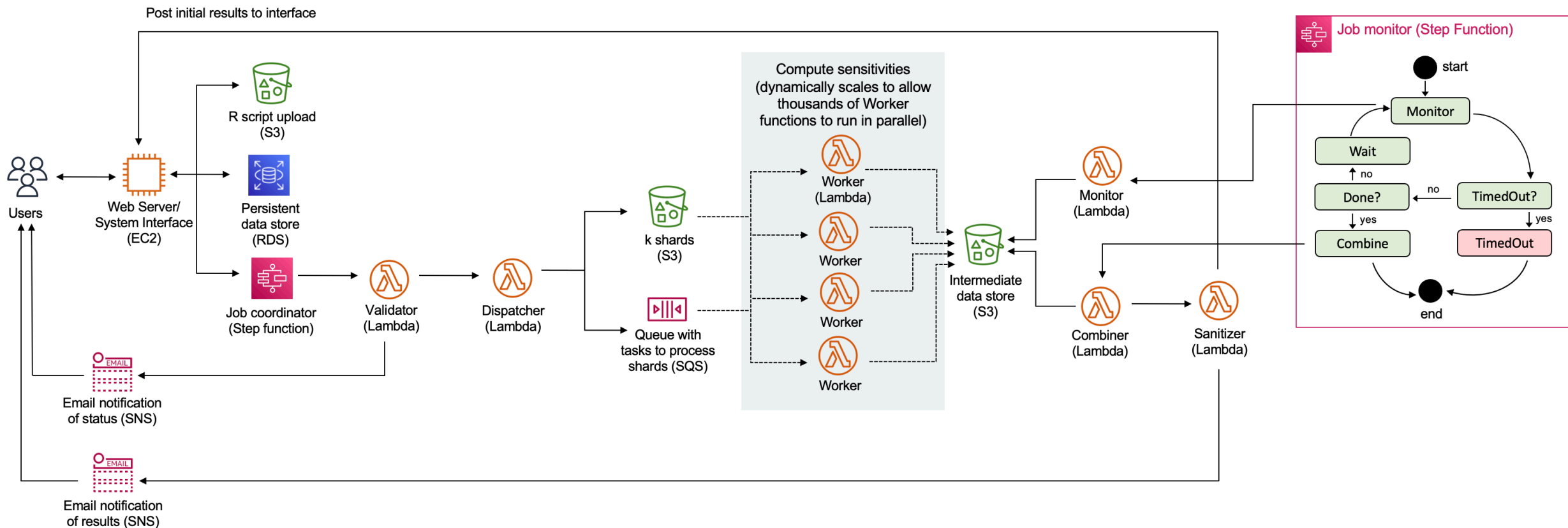
Analysis 1	Estimate	Epsilon Value / Cost to Release	See Details
<input checked="" type="checkbox"/> Version 1	Estimate = 17414472.35	Epsilon Value / Cost to Release = 3.1	See Details
<input type="checkbox"/> Version 2	Estimate = 21785489.91	Epsilon Value / Cost to Release = 2.9	See Details
<input checked="" type="checkbox"/> Version 3	Estimate = 1008314.09	Epsilon Value / Cost to Release = 1.5	See Details

Current Remaining Budget - REVIEW & REFINE 94.6 of 100 Current Remaining Budget - PUBLIC RELEASE 100 of 100

PENDING COST TO RELEASE = 4.6 [Refine Again](#) [RELEASE](#)

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

```
# 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)  
}
```

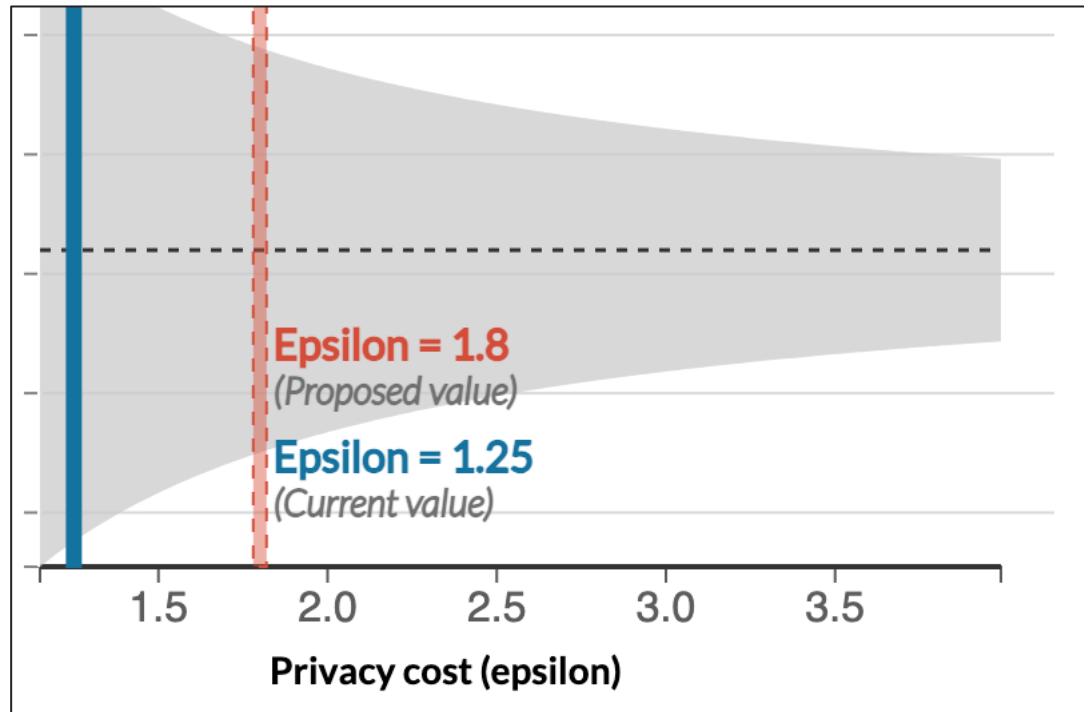
Login
Email

Password

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.

Submit Refinements?

1 If you submit these refinements, your results will be altered and your current remaining privacy budget will decrease to reflect the "New Remaining Privacy Budget" figure below.

As a reminder:

- Your privacy budget is shared across all analyses in your dashboard
- Refinement costs are deducted from your "Review & Refine" budget.
- Your "Public Release" budget will not be impacted until you release your final results in the next step.

	Review & Refine Budget	Public Release Budget
Remaining Privacy Budget	80.35 / 100	71.96 / 100
Pending Refinement Costs	13.80	N/A
New Remaining Privacy Budget	66.55 / 100	71.96 / 100

CANCEL REFINE

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

How a User Interacts with the Prototype (cont.)

Release Final Results?

i If you release these results, your current remaining privacy budget will decrease to reflect the "New Remaining Privacy Budget" figure.

As a reminder:

- Your privacy budget is shared across all analyses in your dashboard
- The cost to release these results will come from your "Public Release" budget.
- This release will not impact your "Review and Refine" budget unless you make additional refinements at this point.

Review & Refine Budget

Remaining Privacy Budget 80.35 / 100

Pending Refinement Costs N/A

New Remaining Privacy Budget 80.35 / 100

Public Release Budget

71.96 / 100

1.00

70.96 / 100

CANCEL
RELEASE

5. Request to release results by spending from a "Public Release" budget.

5a514c7bdf5466995076a48652de1ad-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	0.005651133686997780	4
	p.value	model	Example Model	58.62287490583350	0.045454545454545500	0.000824089924741928	-0.00
	df	model	Example Model		0.0		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	0.002375714215740840	1
	nobs	model	Example Model		169.0	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	

6. Download results from the interface that can be published.