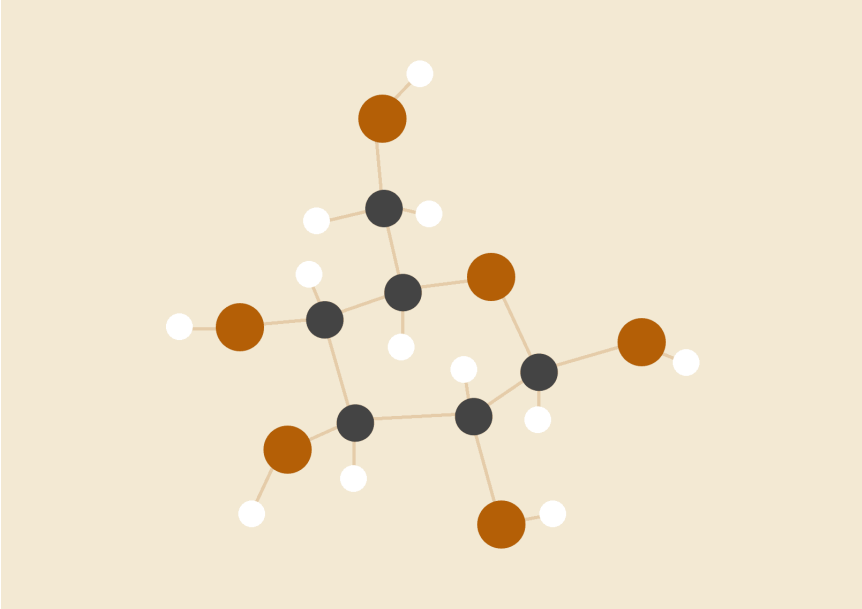
PSI Ψ: (PRIVATE DATA SHARING INTERFACE)

BUDGET TOOL

Privacy Tools for Sharing Research Data Project



Research Intern

Fanny Chow[1] | Nabib Ahmed[2]

Mentored by

Jack Murtagh[3] | James Honaker[4]

Summer 2016

Harvard School of Engineering & Applied Sciences (SEAS)

OVERVIEW

[1] University of California, Davis - fchow@ucdavis.edu

[2] Harvard University - nahmed8098@gmail.com

[3] Harvard University - murtagh.jack@gmail.com

[4] Harvard Institute for Quantitative Social Sciences (IQSS) - james@hona.kr

This project involves the design and implementation of a "privacy budget" user interface for the differentially private algorithms developed by the Privacy Tools Research Group. The Budget Tool will allow researchers to allocate a privacy budget to sensitive datasets by selecting statistics and variables for differentially private releases. The UI (user interface) should be intuitive, user-friendly, well-integrated with the existing components of the system, and easily extensible as the tool continues to expand.

INTRODUCTION

Data provides evidence for the basis of scientific knowledge. Open data improves both the transparency and reproducibility of research. However releasing research data must be made with considerations for protecting the privacy of individuals. Traditional approaches to privacy protection, such as de-anonymization, have shown to be unsuccessful. There have been many examples that these "re-identification" attacks are often quite easy to carry out in by using publicly available datasets as sources of auxiliary information (Vadhan 2016). Differential privacy aims to ameliorate some of the privacy risks of sharing sensitive data.

WHAT IS DIFFERENTIAL PRIVACY?

Differential privacy is a mathematical framework for enabling statistical analysis of sensitive datasets while ensuring that individual-level information cannot be leaked. A differentially private data release algorithm allows researchers to ask practically any question about a database of sensitive information and provides answers with added noise, so that they reveal virtually nothing about any individual's data — not even whether the individual was in the database in the first place (Klarreich 2012).

The project website (http://privacytools.seas.harvard.edu/courses-educational-materials) contains resources for learning more about differential privacy.

PSI (PRIVATE DATA SHARING INTERFACE)

The Privacy Tools for Sharing Research Group is developing a prototype system that will allow researchers to:

1. **upload** private data to a secured Dataverse archive
2. **budget -** *decide what statistics they would like to release about that data*
3. **release** privacy preserving versions of those statistics to the repository
4. that can be **explored** through a curator interface without releasing the raw data, including
5. interactive **queries**.

PSI BUDGET TOOL

The PSI Budget Tool will allow researchers with privacy sensitive datasets upload their data to a repository and select statistics to be released to the public under the guarantees of differential privacy. Data depositors are guided through the fundamentals of differential privacy and appropriate choices of privacy parameters, i.e. their "privacy budget". The tool then guides the depositors through distributing their privacy budget among the statistics they wish to release about the dataset. Once the depositor is finished, all of the statistics she selected are computed using differentially private algorithms and released for other researchers to see.

Eventually, the system will allow data depositors to optionally reserve some of the budget for future users to ask further questions about the data. This reserved budget will either be shared among all future users and once it is exhausted, nobody else may ask further interactive queries or each user will get a personal budget.

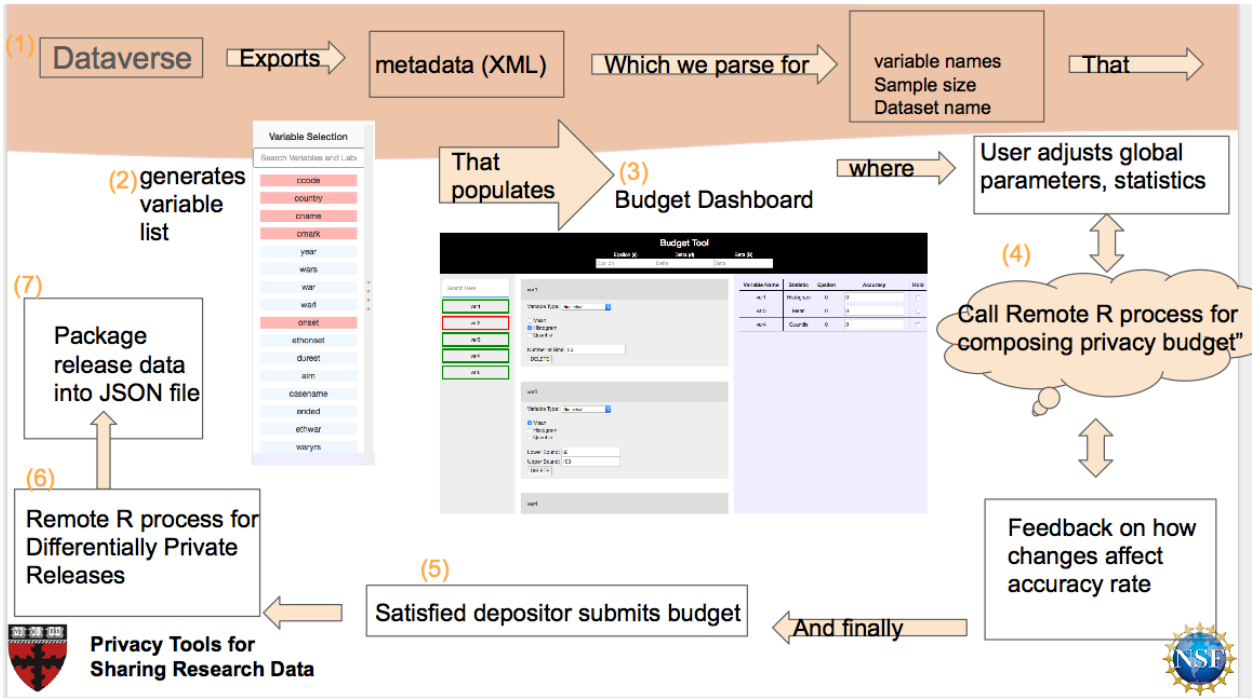
Once differentially private statistics are released and the interactive query interface is finished and integrated, other data analysts will be able to go to Dataverse, a data repository, and "explore" the uploaded dataset through an interface called TwoRavens. TwoRavens allows people to run statistical models on a dataset without ever touching the data - only using metadata and statistics about the data. If these statistics and metadata satisfy differential privacy, then so will any of the statistics run by TwoRavens (by the post-processing property of differential privacy). The interactive query interface will eventually be integrated with TwoRavens.

At the beginning of the Summer, we began with Budget Tool 1.0, a prototype user interface, and a command line tool for the interactive queries. We redesigned and implemented a new budgeting interface called Budget Tool 2.2.

HIGH LEVEL GOALS

UX (USER EXPERIENCE) GOALS	DESIGN GOALS	DATA ARCHITECTURE GOALS
User should be able to choose what statistics they want to release & the value of global parameters	Scale for univariate statistics (ie.Mode, Box and Whisker Plots, Frequency Polygons)	"stateless" (store as little information as possible about the history of the interaction with the user)
Should be simple and intuitive and only require a conceptual understanding of differential privacy	Scale for multivariate statistics added later (ie. multivariate regression, PCA, covariance matrix)	Integrating code bases that underlie these systems (Dataverse, the budgeting interface, TwoRavens, interactive queries)
Should minimize work of data depositor & check against erroneous data entry	Should be versatile enough to scale for additional types of datasets or privacy features	

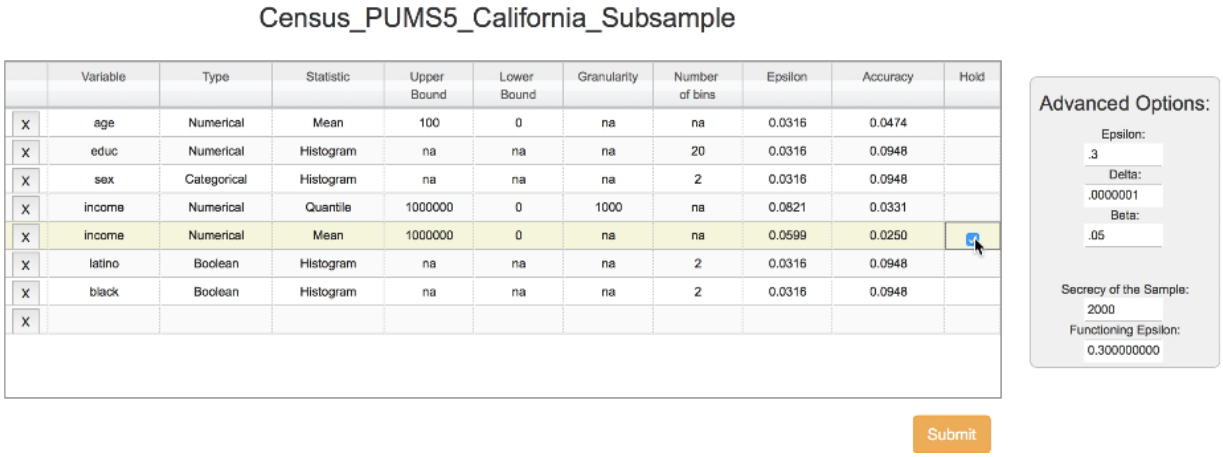
DATA ARCHITECTURE



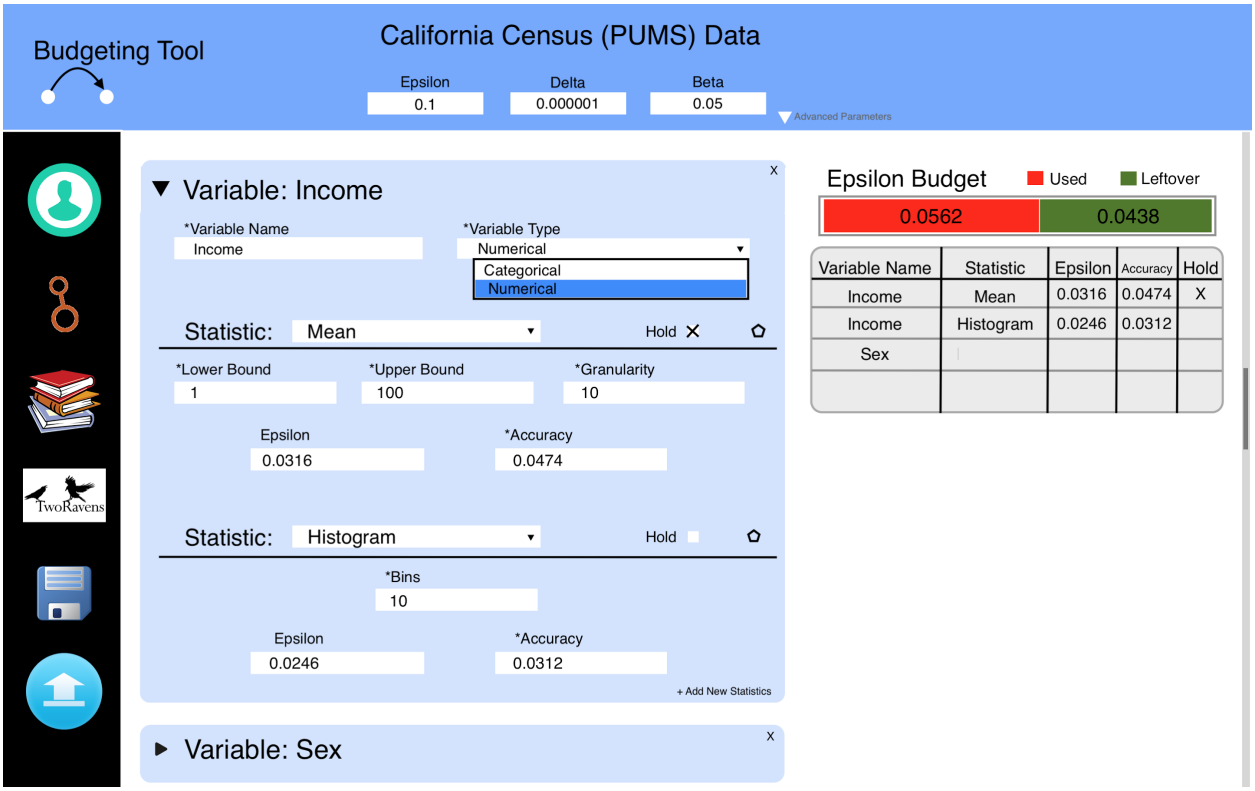
The Budget Tool is comprised of a data pipeline corresponding to the user workflow:

1. Dataverse
    a. Upload raw data set
    b. We parse metadata XML file for:
        i. Variable names
        ii. Sample size
        iii. Name of data set
2. Generate variables list in left sidebar of Budget Tool
    a. User selects or deselects variables of interest
3. Budget Dashboard
    a. User can adjust global parameters

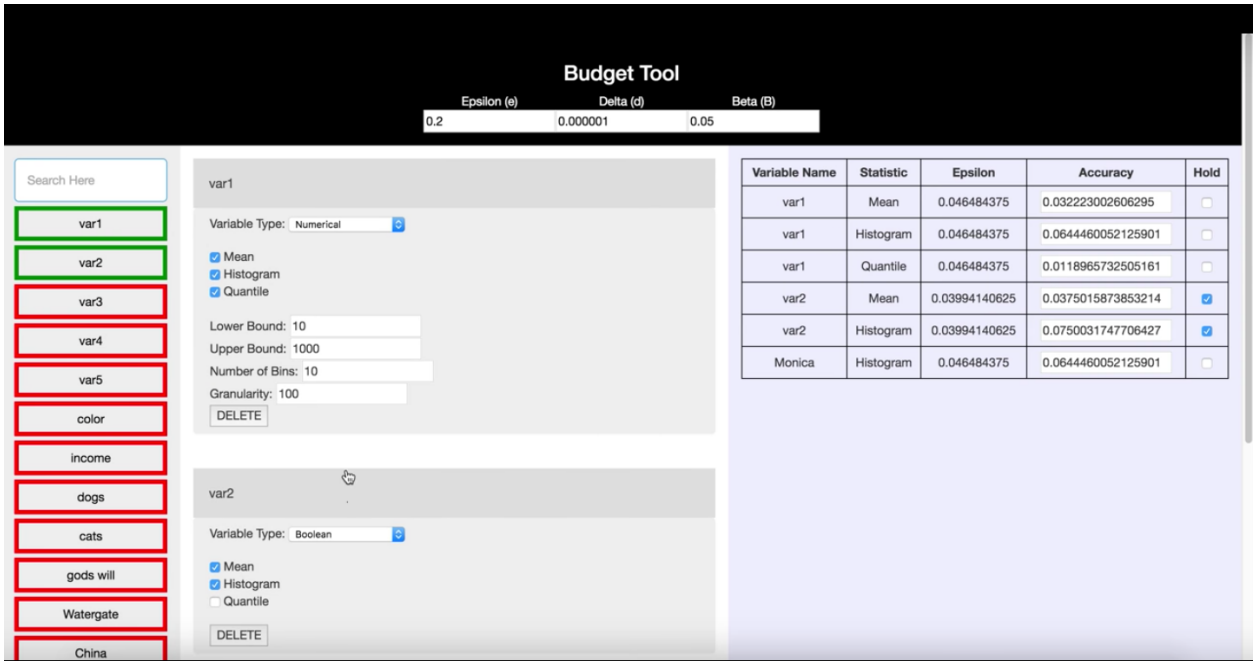
BUDGET TOOL 1.0



BUDGET TOOL 2.0 (STATIC WIREFRAME)

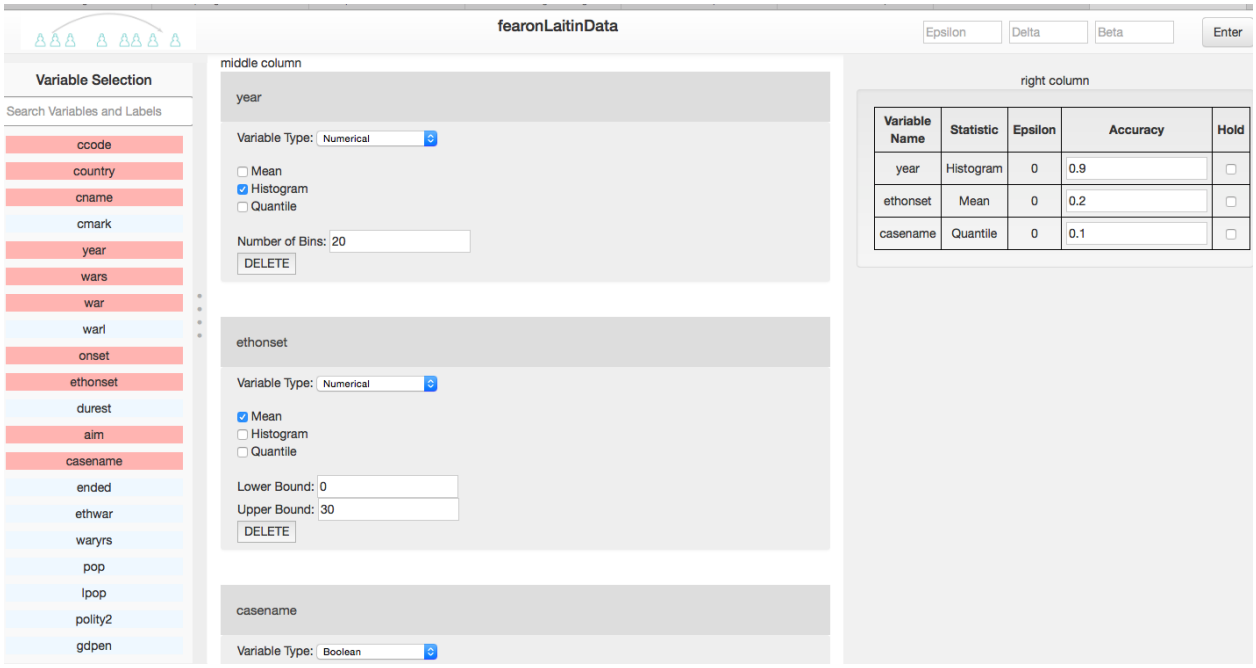


BUDGET TOOL 2.1 (IMPLEMENTATION)



Video Demo: https://youtu.be/4GwX5Br7OWs | **Github Repo:** https://github.com/Nashmed28/REU

BUDGET TOOL 2.2 (IMPLEMENTATION)



Video Demo: https://vimeo.com/180649061 | **Github Repo:** https://github.com/fbchow/budget_tool

RESULTS

Some main accomplishments:

1. Designed a working prototype (version 2.2) that is suitable enough to replace the old implementation (version 1.0).
2. Build modularly so that features, such as multivariate statistics, could easily added
    a. Avoid "hard-coding"
    b. JSON file with list of available statistics to populate and generate the web app
3. Begin connecting frontend to back-end R server processes
    a. design mock-up → working prototype

CONCERNS ABOUT DIFFERENTIAL PRIVACY: THEORY VS PRACTICE

1. Is the current implementation considered a "stateless" process (as little information as possible about the history of the interaction with the user is stored)?
2. Is the user interface a truthful interpretation of differential privacy?
3. Will allowing an analyst to save budget tool and work it on it later allow data dependent decisions?

FUTURE WORK

Here are some future steps in the development of the Budget Tool:

1. UX (User Experience) Testing
    a. Jack Landry (2016 Summer REU Intern) and Caper Gooden (2015 Summer REU Intern) have developed a series of educational documents and user experience questions to assess understanding and effectiveness of the budget tool for social scientists and researchers with minimal understanding of differential privacy
2. Formal Documentation
    a. Many aspects of the current implementation are auto-generated, including the statistics and types available. In order to meet our goal of scalability, the tool is designed such that the information regarding what statistics and types are available only need to be passed in once. Currently they are passed through as a JSON object. This JSON object must meet a certain format in order for the tool to be executed properly. This means that future researchers who wish to create additional r-servers for more statistics and types must add information to the JSON object in order for the site to properly reflect the additions. Thus formal documentation needs to be created where researchers can learn the correct method of adding this vital information.
3. Other Univariate Statistics
    a. Currently the tool only allows the release of three statistics when in fact there is a myriad of univariate statistics that can be used, so one goal is to develop r-libraries that can apply differentially private methods for these statistics such that researchers can release more on their

data.

4. Multivariate Statistics
    a. Researchers would like to release statistics that show the relationship between two or more variables. Currently the tool is design to release statistics about one variable at a time and so the above becomes very tricky to do with our current tools. Thus, one goal is to create a design that allow researchers to release both univariate and multivariate statistics with ease and simplicity.

CONCLUSION

Ultimately the Budget Tool will allow researchers to work with the guarantee of privacy while sacrificing minimal data utility. With this working tool, data depositors can begin experimenting with differential privacy releases without having to learn the specifics of differentially private algorithms. Developing the privacy budget tool is one step towards developing the entire suite of PSI tools that will make privacy-protective data-sharing easier for researchers.

REFERENCES

"PSI (Ψ): A Private data Sharing Interface." The Privacy Tools Project Differential Privacy Group. Working Paper.

Erica Klarreich in Scientific American, "Privacy by the Numbers: A New Approach to Safeguarding Data" *Scientific American* Dec. 2012: 62-69. Print.

Salil Vadhan, "The Complexity of Differential Privacy," The Privacy Tools Project Differential Privacy Group. Working Paper. 2016.

Honaker, James and Vito D'Orazio. "Statistical Modeling by Gesture: A graphical, browser-based statistical interface for data repositories." *Extended Proceedings of ACM Hypertext 2014*. PDF

Jack Murtagh and Salil P. Vadhan. "The complexity of computing the optimal composition of differential privacy." In *Theory of Cryptography - 13th International Conference, TCC 2016-A, Tel Aviv, Israel, January 10-13, 2016, Proceedings, Part I*, pages 157–175, 2016.

Kobbi Nissim, Thomas Steinke, Alexandra Wood, Mark Bun, Marco Gaboardi, David O'Brien, Salil Vadhan. "Differential Privacy: An Introduction for Social Scientists". Working Paper.