Benchmarking Platform for Differentially Private Algorithms

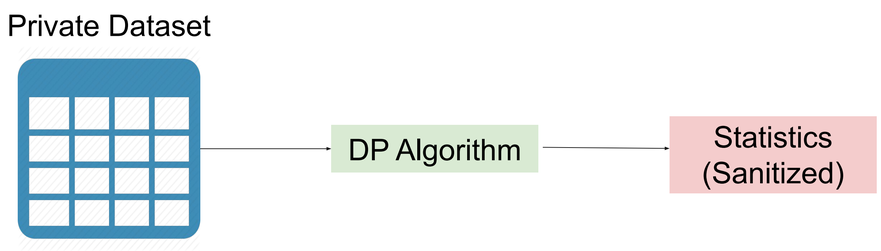
Allen Marquez[1,2], Zhiwei Steven Wu[1], Terrance Liu[1]
[1]Carnegie Mellon University, [2]California State University, Los Angeles

What is Differential Privacy?

In this setting, the goal is to **release queries,** which are **statistical questions** that can be asked of a dataset

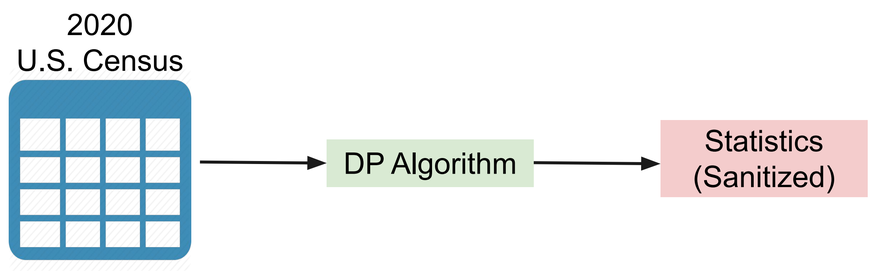
- *e.g. What fraction of people living in Pittsburgh are males between the ages of 20 and 30?*
- **Number of queries** is often larger than the number of people in the dataset itself and is **exploitable**
- Adversaries can **reconstruct datasets** from released statistics and datasets that **compromise people's privacy**

Differential privacy provides a mathematical platform for guaranteeing privacy that **disguises personal data when published,** allowing users to **calibrate** between **accuracy and privacy** when releasing sensitive information.



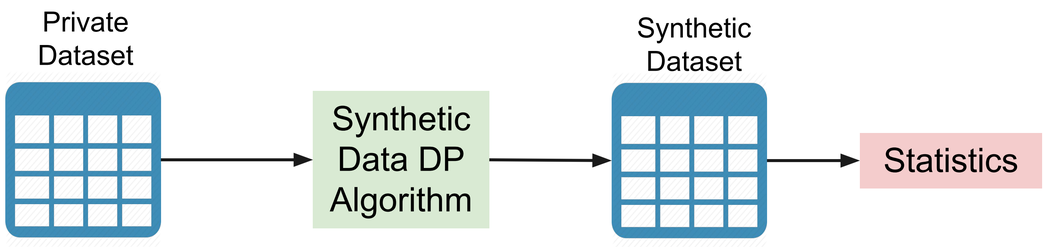
Differential Privacy in 2020

- U.S. Census Bureau releases a large amount of sensitive data and **simulated a reconstruction attack** on the 2010 Census and found **52 million people could be identified**
- U.S. Census Bureau is **modernizing privacy protections with differential privacy** to future proof against attacks



Synthetic Data Generation

- Our lab primarily focuses on synthetic data generation algorithms, which generate a **"fake" dataset** that can be used to answer queries
- There exists a **family of synthetic data generation algorithms** that satisfy differential privacy
  - Each has their own advantages (computationally efficient, better for higher dimensional data, ect.)



The Problem

There is not a standardized and comprehensive platform for testing and deploying differentially private algorithms

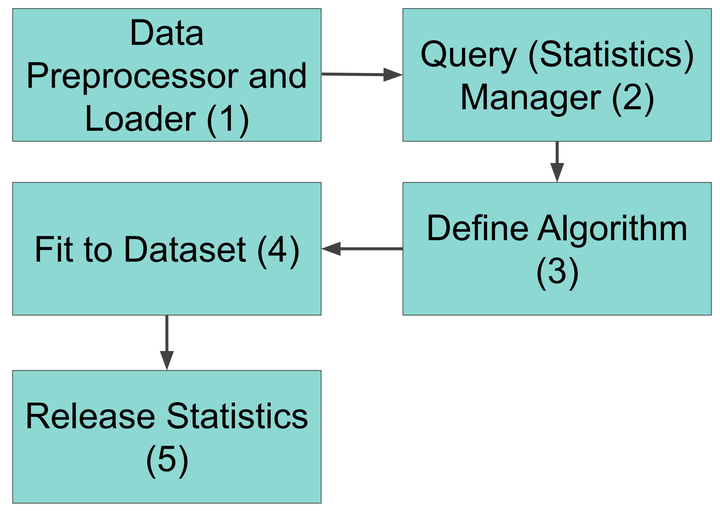
Our Proposed Solution

- Create a **modular platform** that that will make it **easier for research and development**
- Goals for Target Audience:
  - **Differential privacy Researchers**
    - Develop new algorithms
    - Test against standardized datasets
  - **Non experts in differential privacy** (social scientists, policy makers, etc.)
    - Evaluate on their own datasets

Our Approach to Designing this Platform

End Product Goal: User enters a few lines of code to run algorithms

```
(1) data = load_data("census")
(2) qm = QueryManager(data)
(3) algo = MWEM()
(4) algo.fit(data, qm)
(5) answers = algo.predict()
```



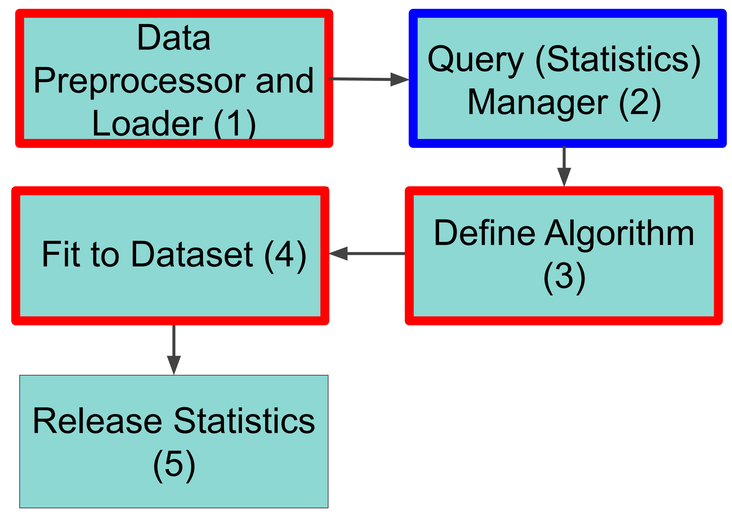
Design Philosophy

- Provide tools that are **user friendly**
- Allow platform to be modular
  - **Query Manager (2) is able to interchangeable** if user decides load a different set of queries
  - Define Algorithm (3) allows **users to load different algorithms** that are part of the platform, but also **allows users to define new algorithms** and implement it into our platform

Current Progress in Pipeline

```
(1) data = load_data("census")
(2) qm = QueryManager(data)
(3) algo = MWEM()
(4) algo.fit(data, qm)
(5) answers = algo.predict()
```

Done
In progress



Acknowledgements

I would like to thank my advisor Dr. Steven Wu for providing advice and guidance throughout this project. I'd also like to thank my mentor Terrance Liu for his invaluable help throughout this project. I would also like to thank the Institute of Software Research for their support throughout the duration of this program. Without all the support from this project wouldn't have been possible.