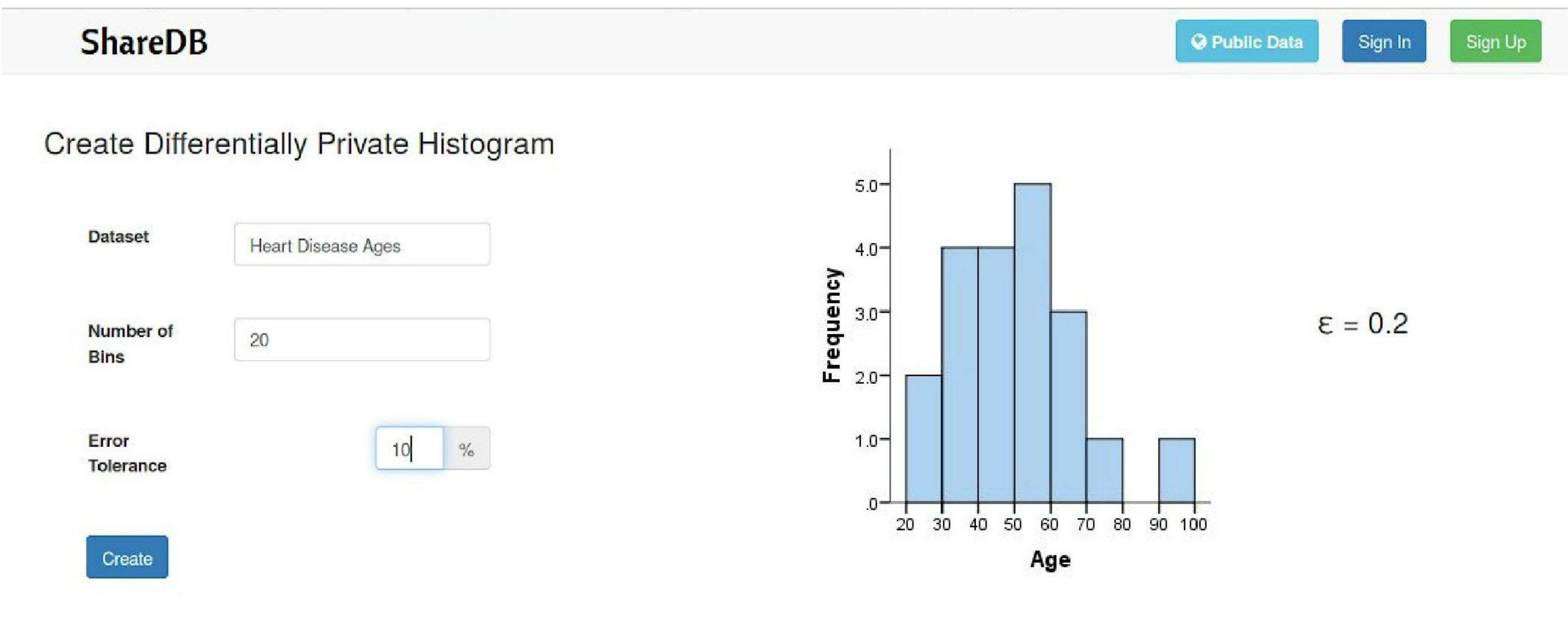
DIfferential Privacy in Licensing Model and Ecosystem for Data Sharing

Famien Koko[1], Jane Greenberg[2], Sam Grabus[2], Tim Kraska[1], Sam Madden[1]

[1]MIT, [2]Drexel University

Summary

The overall aim is to develop a data sharing system and an approach that addresses legal matters, policies, privacy concerns, and other challenges that too frequently hold up the process due. Specifically, we are working on a system, ShareDB (build on DataHub), that will expedite the process finalizing data sharing agreement. Privacy is often an important aspect of a data sharing agreement, and we have worked on a model to facilitate make and sharing privatized data.



Create Differentially Private Histogram

Dataset: Heart Disease Ages

Number of Bins: 20

Error Tolerance: 10 %

Figure 2: Creating a differentially private histogram

Differential Privacy

We have build a model which will allow users to easily privatize data queries. User simply select a dataset they have uploaded to ShareDB, input the queries they want to perform on the data (for example, a counting query to generate a histogram) and give an error percentage they are comfortable with. Our system will then select the differentially private algorithm which guarantees the most privacy, and return a privatized query using that algorithm. See Figure 2.

Meta-Algorithm Model

Several differentially private algorithms have been developed, but some work better than others, depending on what the data looks like and what queries are going to be asked. We developed a model which takes a set of differentially private algorithms, a workload of queries and an error rate, and will return the algorithm which will achieve the desired error rate, and has the maximum amount of epsilon privacy. The model was build using a random forest regression algorithm, and trained over a set of representative data sets, workloads, and errors to predict the epsilon needed to acheive a particular error rate for that data and workload. We found that this approach acheived reasonable accuracy in predicted epsilons

Future Directions

The differential privacy meta-algorithm model supports generalized query workloads, and we plan to implement this in the ShareDB system

The model is also extendable, so that you can easily add new new differentially private algorithms to it.

We plan to continue adding functionality to ShareDB to provide users a rich toolkit and intuitive interface to share, privatize and license their data

Continue developing our ShareDB platform. Especially, we want to investigate additional anonymization, PII detection, and watermarking techniques

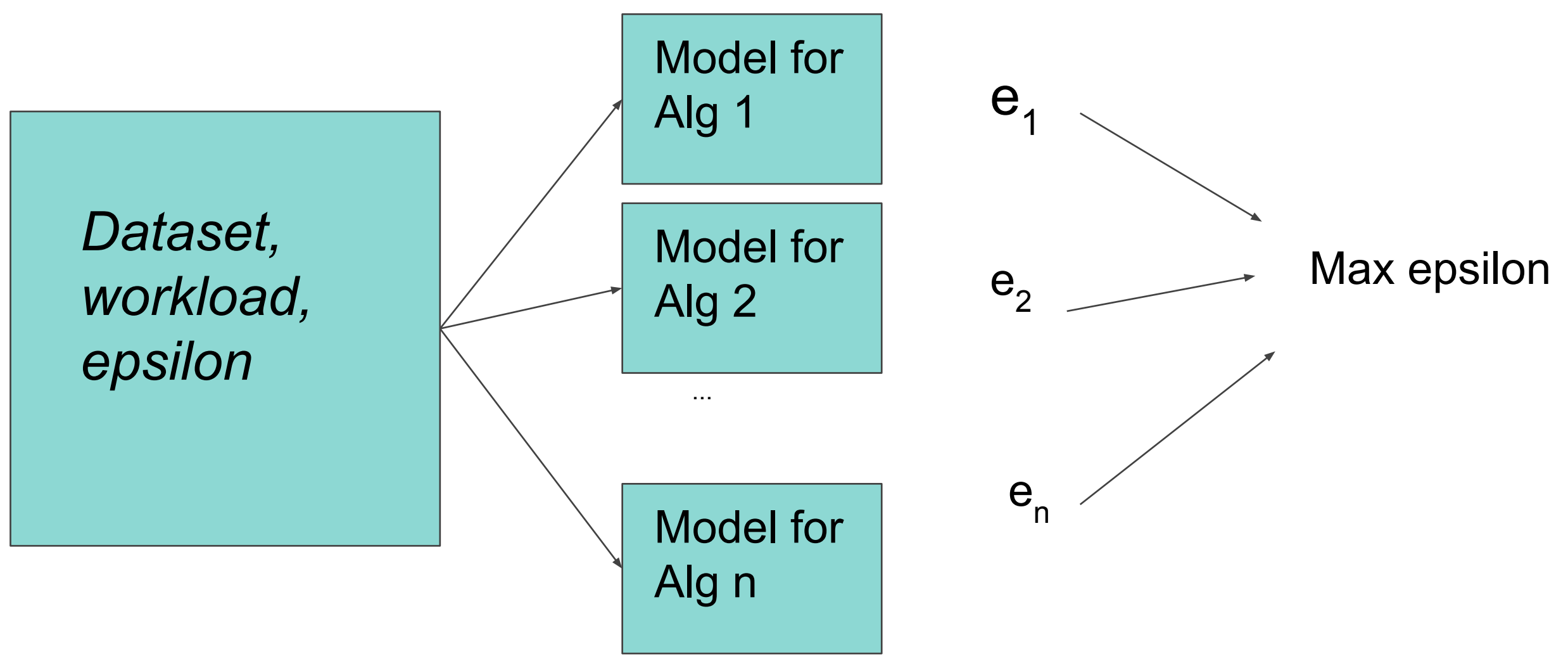


Figure 1: Creating the meta-algorithm model