Proposal

Group members:

- Marko Janic
- Leo Posva

What problem do we want to solve:

We are interested in comparing the performance of local differential privacy mechanisms such as unary encoding or randomized response and bloom filters (using for example Google's Rappor) on a dataset of our choice. We may adjust the number of mechanisms we are comparing in case we have too much or too little content for our project. The metrics we are interested in are mainly performance and accuracy.

What approach do we want to take (e.g.,what datasets,what algorithms/libraries):

We want to use the following dataset: CDC Diabetes Health Indicators

Description: The Diabetes Health Indicators Dataset contains healthcare statistics and lifestyle survey information about people in general along with their diagnosis of diabetes. The 35 features consist of some demographics, lab test results, and answers to survey questions for each patient. The target variable for classification is whether a patient has diabetes, is pre-diabetic, or healthy.

To test bloom filters we might use Google's Rappor from their github or the PyProbables Library.

Unary encoding or randomized response we can implement ourselves relatively easily. Or if that leads to problem we can look at the pure-LDP Library.

The project will mainly be in Python.