Core Questions

• How do we make differentially private releases of greatest use to statistical analysts?
Core Questions

• How do we make differentially private releases of greatest use to statistical analysts?
• What are the algorithms we need to implement?
Core Questions

- How do we make differentially private releases of greatest use to statistical analysts?
- What are the algorithms we need to implement?
- What are the algorithms we need to research?
Core Questions

• How do we make differentially private releases of greatest use to statistical analysts?
• What are the algorithms we need to implement?
• What are the algorithms we need to research?
• What are the properties of these algorithms we need to expose?
• What is the infrastructure for doing that best?
Core Questions

• Differential privacy is a definition of worst-case privacy loss from the release of a function on the data.
Core Questions

- Differential privacy is a definition of worst-case privacy loss from the release of a function on the data.
- Utility is an expectation of information gain from the release of the function.
Population and sample

![Histogram of educational attainment](image)
Population and sample
Population and sample

![Histogram of educ](image)

![Histogram of educ](image)
Population and sample
Population and sample
Estimator
Estimator
Estimator
Uncertainty
Uncertainty
Uncertainty
Posterior Incorporation

Sensitive Data

DP Release

correlations

Density
Statistical Utility Metrics

- Error - how far the estimate is from the underlying “true” value
- Precision/Recall - similar for classification
- Bias - the expected value of the error.
- Variance - RMSE can be decomposed into bias and variance
- Coverage - correctness of measures of uncertainty
- Consistency - converge to the population value as the sample grows
- Invariance - estimates should not depend on arbitrary scaling choices
- Power - how small an effect can be observed with a finite sample of data.
Key Points

Propositions for the OpenDP community:

1. To facilitate applied research, the OpenDP library should immediately prioritize incorporating a collection of algorithms that provide uncertainty estimates, analytical error promises, and data generating processes that are easy to incorporate into statistical models.

2. Analysts need different categories of algorithms, and a priority should be to round out methods in each of these.

3. There is not one single or unifying utility measure. Library components should support different systems of inference and notions of utility.

4. Contributions to the OpenDP library ought to come with some measure or evidence of utility.

5. The OpenDP platform should provide automated support for Monte Carlo experiments on public datasets or synthetic data including infrastructure for automated benchmarking.
Statistics Prioritization List

• Data Curation
  ▶ Methods for metadata creation

• Data Cleaning and Preprocessing
  ▶ Outlier detection, Missing data treatment and imputation, Matching methods

• Exploratory Data Analysis
  ▶ Visualization, Histograms, Univariate Statistics, Medians, Quantiles, Range Queries, PCA

• Modelling
  ▶ Generalized Optimization Methods, Sample-and-Aggregate approaches, Linear and Logistic regression, Likelihood methods, Hypothesis Testing, Synthetic Data Generation
Call for Lightning Talks

Community Meeting, May 13th - 15th, 2020

Breakout Session: Statistical Functionality

We would love to use the statistics break out session to coalesce a community of privacy researchers, statisticians, and other use case partners to form a joint understanding of what are the priorities, goals and research needs of the OpenDP library to best facilitate applied statistically oriented research with sensitive data.

We would welcome anyone who could see themselves in that broad community, to offer to give 4 minute, 1 slide lightning presentations on Friday, May 15, 11am-12:45pm EST, so we can collectively understand each other's needs and abilities. This could be a researcher offering an algorithm they have worked on, or a statistician describing their approach or work on bridging DP with statistical properties, or an analyst describing what they have done (or what they haven't done because methods don't exist) or future possible use cases with sensitive data. Our understanding is broad, undoubtedly we have not thought of everything, and don't feel limited in your proposals.

James and Sesa
James Honaker
Aleksandra (Sesa) Slavkovic