Automatic, Fine-Grained Algorithmic Choice for Differential Privacy

Jacob Imola (Advisor: Jean Yang)

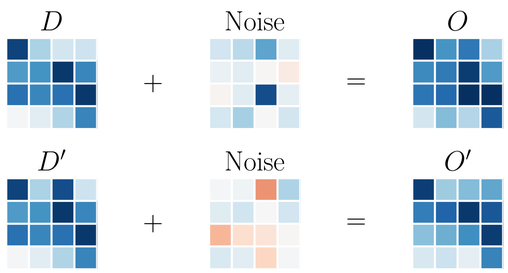
Carnegie Mellon University

Introduction

- Differential Privacy is useful but complicates code with noise.
- For all D and D' differing in 1 row: P is ϵ-DP if $\Pr(P(D) = O) < e^{\epsilon} \Pr(P(D') = O)$ for all O.

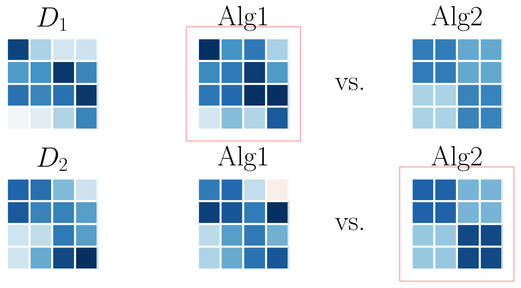
Example

- Algorithm One



$\Pr(P(D) = O) \approx 10^{-8}$ $\Pr(P(D') = O) \approx 2 \times 10^{-9}$

- Algorithm Two
  - Sum into 4 2x2 buckets, add noise, divide by 4
- Which is better?



Vision

Task: Remove burden of DP algorithm analysis: `ChoiceMaker`.

1. **Correctness** Differential privacy is never violated.
2. **Generalizability** Works on arbitrary code.
3. **Performance** Makes choice "close enough" to optimal.

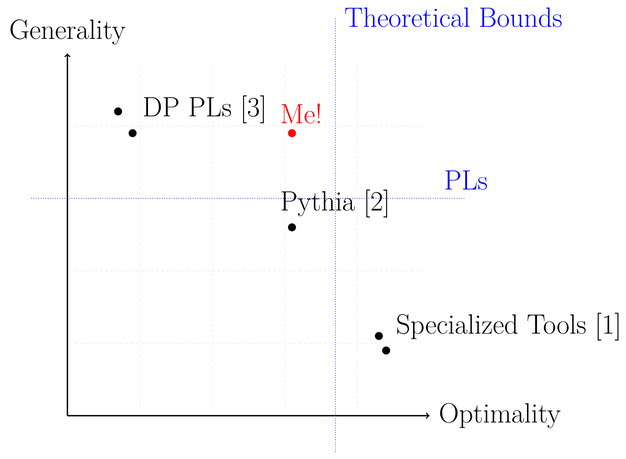
Solution: A programming language!

```
answerHistQueries = MkChoiceMaker among {Alg1, Alg2}
answers = answerHistQueries(data, queries)
```

Challenges

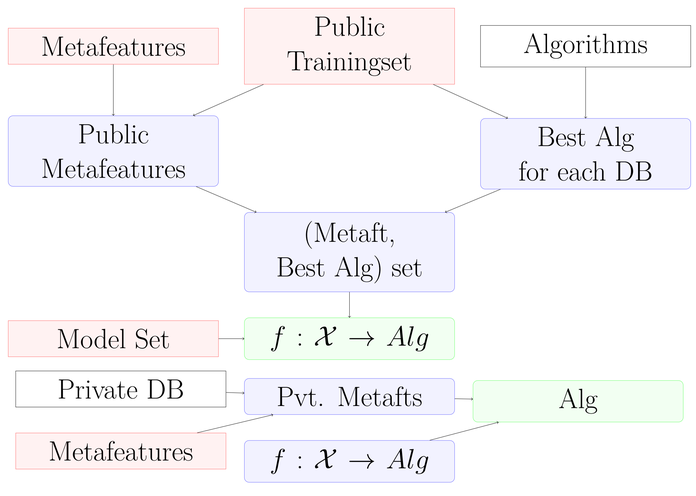
- Generality $\implies$ Can only run Alg1, Alg2.
- Meta-machine learning: function $f : DB \to Alg$.
- Intractable—data science cannot be automated well.

Existing Work



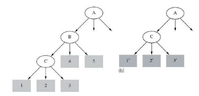
Solution Overview

- Modeled off data science: make problem tractable by specifying (meta-)features $\mathcal{X}$ of DB. Learn $f : \mathcal{X} \to Alg$.



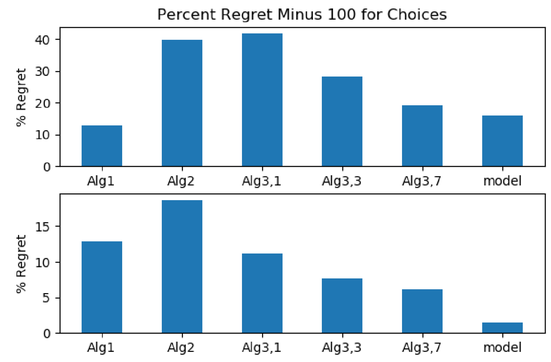
Experimental Setup

- **Algorithms** Stopping Criteria for Private Decision Trees. Not previously done.
- **Metafeatures** DB size, epsilon, domain size.
- **Model** Linear Classifiers, binary loss function.
- **Training Set**
  1. 300 real DB snapshots, 100 real DB snapshots.
  2. 300 synth. DB snapshots, 100 synth. DB snapshots.



Results

- Regret: my performance vs. best performance, averaged.



- Does as well as Pythia [2] with same expressiveness as PINQ [3].
- Only as good as how well the programmer frames the ML problem.

Bibliography

[1] Kamalika Chaudhuri and Staal A. Vinterbo.
A stability-based validation procedure for differentially private machine learning.
In *NIPS*, 2013.

[2] Ios Kotsogiannis, Ashwin Machanavajjhala, Michael Hay, and Gerome Miklau.
Pythia: Data dependent differentially private algorithm selection.
In *Proceedings of the 2017 ACM International Conference on Management of Data*, SIGMOD '17, pages 1323–1337, New York, NY, USA, 2017. ACM.

[3] Frank McSherry.
Privacy integrated queries: An extensible platform for privacy-preserving data analysis.
Commun. ACM, 53(9):89–97, September 2010.