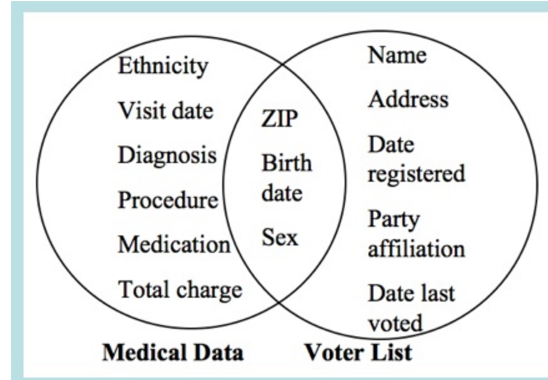
CS208: Applied Privacy for Data Science
Conclusions

James Honaker & Salil Vadhan

School of Engineering & Applied Sciences
Harvard University

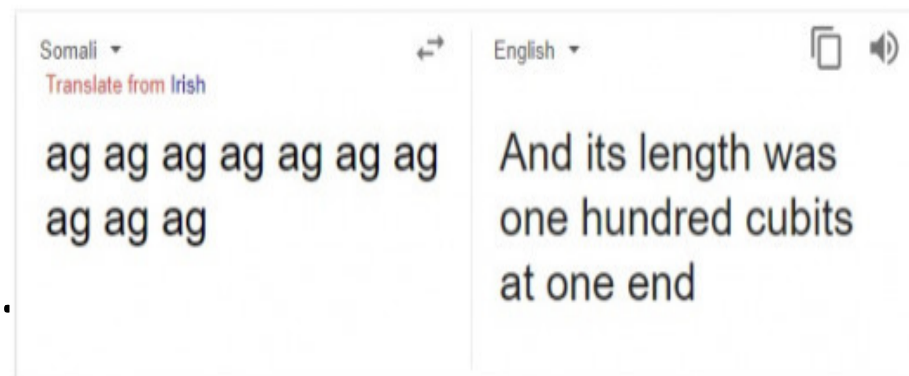
April 29, 2019

Privacy Risks



[Sweeney `97]

- Deidentified data can often be reidentifed.
- Naïve query systems are subject to differencing-style attacks.
- Releasing too many aggregate statistics allows for reconstruction or membership attacks (Census, Diffix).
- Machine learning models can memorize their data and allow for membership attacks (Shokri et al., Google translate).



Definition of Differential Privacy

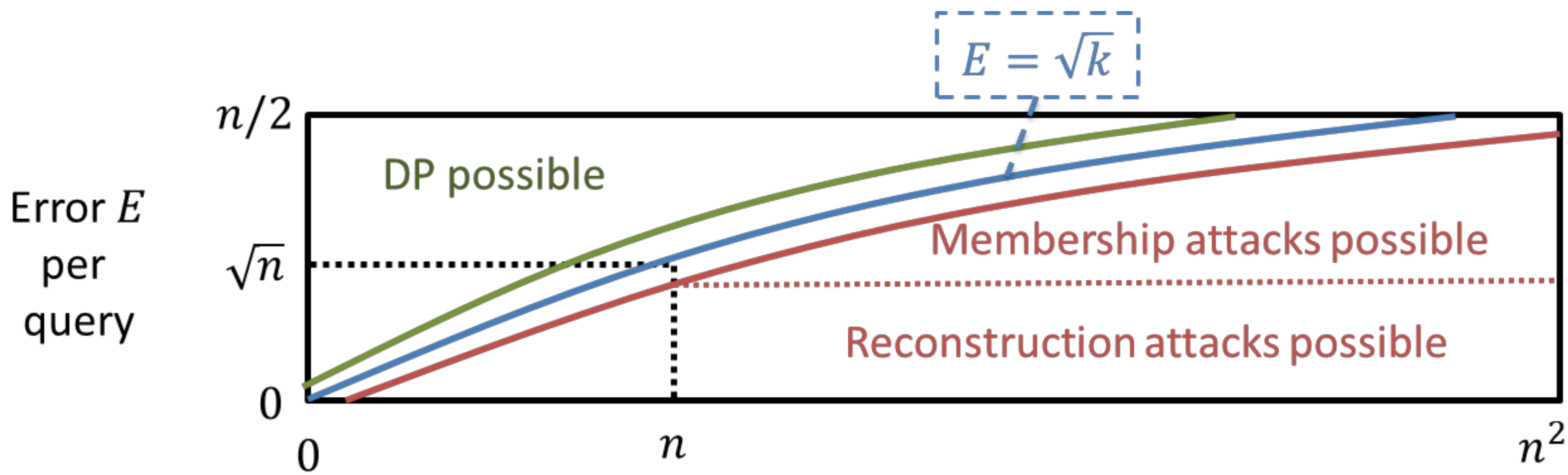
- Strong privacy definition.
- Compatible with many statistical analyses.
- Ensures that "individual-level information" does not leak.
- Applies regardless of adversary's auxiliary information.
- Adversary can be external "analysts" (centralized DP) or aggregator (local DP).

But:

- Adversary may still infer sensitive attributes.
- Not applicable when utility requires individual-level data.
- "Privacy" has many other meanings beyond what is captured by DP (cf. Solove taxonomy).

Composition of DP

- DP and variants (pure, approximate, zCDP, moments accountant) satisfy composition thms.
- Allows for modular design of DP algorithms (w/post-processing).
- Leads to tradeoff of # queries vs. accuracy (vs. privacy)



- Tradeoff is worse in local model.

DP Algorithms & Tools We've Seen

- Means
- Medians/Quantiles/Ranges
- Histograms
- Regression
- Synthetic Data Generation
- Empirical Risk Minimization
- Deep Learning/SGD
- Subgraph Counts
- Degree Distribution

- $\boldsymbol{\varepsilon}$ktelo
- PSI
- RAPPOR
- Tensorflow privacy

There are many more!

Core Components

A small number of primitives form the building blocks of some of the most complicated models, including:

- Clipping/Clamping
- Laplace (and Gaussian) Mechanisms
- Exponential Mechanism
- Randomized Response
- Composition
- Binning, One-hot encoding

As well as some core recurring ideas:

- Post-processing
- Lipschitz transformations
- Subsampling

Experimental Investigation

Monte Carlo simulation methods are a valuable tool for investigating utility and other performance measures of algorithms. We have used this underlying template repeatedly:

1. Simulate data from distribution with known properties
(or bootstrap from large dataset as if a population).
2. Release DP estimate and compare to true estimand.
3. Repeat 1 & 2 to integrate over simulation error and summarize.
4. Repeat 3 over free parameters of interest.

Value of Rigorous Thinking in Privacy & Security

- Break cycle of attack-defense-attack-defense-…
- Separates goal from solutions.
  - Can evaluate privacy/security definition on its own.
  - Opens design space for solutions.
- Makes assumptions about adversary and implementation explicit, evaluable.
- Allows for study of tradeoffs (e.g. privacy vs. utility) and limits (impossibility, hardness).

Deployments of DP

Census, Opportunity Atlas, Google, Apple, Uber, Privitar, Leapyear, …

Challenges and Open Problems:

- Getting both sufficient utility and satisfactory privacy.
- Managing privacy budget over many queries and analysts.
- Compatibility with existing data science workflows.
- Practical methods for generating synthetic data.
- Enabling analysts to interpret noise, perform inference, measure uncertainty.
- Bridging with law & policy.
- Relational data (joins).
- Side channel attacks (e.g. randomness, timing).
- Vetted and general-purpose software tools.

To Pursue Further at Harvard

- Some final projects may lead to publishable papers.
- Join the Privacy Tools Project
  - Email Louisa Bloomstein lbloomstein@seas.harvard.edu and cc us to join mailing lists for regular meetings.
  - Apply for summer or term-time internship or developer position https://privacytools.seas.harvard.edu/participate/positions
  - Probably coming soon: “OpenDP project”
- Explore annotated bibliography.
- Come discuss with us in office hours.
- Take Cynthia Dwork’s CS227r next year for more theory of DP and related topics (e.g. preventing overfitting, algorithmic fairness).

To Pursue Further Elsewhere

- Apply for a job as a privacy engineer/data scientist/researcher.
  - Big & small tech companies
  - Privacy start-ups
  - Government agencies
  - Privacy non-profits and advocacy organizations
  - Industries grappling with data privacy (healthcare, finance, …)

- Apply to graduate programs at places doing DP
  (we're happy to provide advice).