PRIVACY TOOLS FOR SHARING RESEARCH DATA

Salil Vadhan (lead PI), Harvard University
http://privacytools.seas.harvard.edu/

DataTags Tools

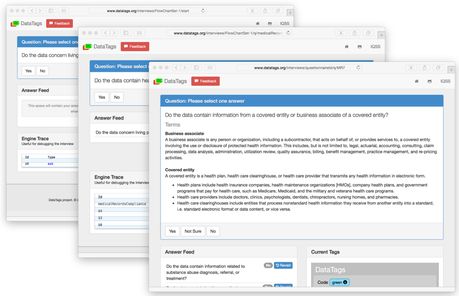
Tools that help generate
a policy for your sensitive data
that defines how to transfer, store,
access, and use those data.

DataTags Levels

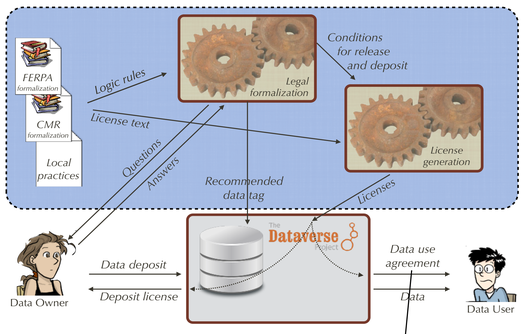
Tag Type	Description	Security Features	Access Credentials
Blue	Public	Clear storage, Clear transmit	Open
Green	Controlled public	Clear storage, Clear transmit	Email- or OAuth Verified Registration
Yellow	Accountable	Clear storage, Encrypted transmit	Password, Registered, Approval, Click-through DUA
Orange	More accountable	Encrypted storage, Encrypted transmit	Password, Registered, Approval, Signed DUA
Red	Fully accountable	Encrypted storage, Encrypted transmit	Two-factor authentication, Approval, Signed DUA
Crimson	Maximally restricted	Multi-encrypted storage, Encrypted transmit	Two-factor authentication, Approval, Signed DUA

DataTags and their respective policies
Sweeney L, Crosas M, Bar-Sinai M. Sharing Sensitive Data with Confidence: The Datatags System. Technology Science. 2015.

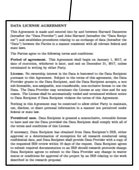
Automated Interviews



Robot Lawyers



Personnel associated with Robot Lawyers:
- Micah Altman
  Stephen Chong
- Alexandra Wood
- Obasi Shaw
- Aaron Bembenek
- Kevin Wang



Other Accomplishments

- Many theoretical results illuminating the limits of differential privacy (lower bounds, algorithms, hardness results, attacks).
- Theoretical and empirical work bridging differential privacy & statistical inference (confidence intervals, hypothesis testing, Bayesian posterior sampling).
- Framework for modern privacy analysis: catalogue privacy controls, identify information uses, threats, and vulnerabilities, and design data programs that align these over data lifecycle.

Motivation

Computational Social Science

The potential: massive new sources of data and ease of sharing will revolutionize social science.

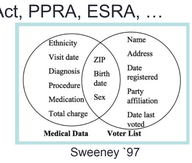
The problem: protecting the privacy of data subjects

e.g. NYT 5/21/12 "Troves of Personal Data, Forbidden to Researchers"

Challenges for Sharing Sensitive Data

Complexity of Law
- Thousands of privacy laws in the US alone, at federal, state, and local levels, usually context-specific: HIPAA, FERPA, CIPSEA, Privacy Act, PPRA, ESRA, …

Difficulty of Deidentification
- Stripping "PII" usually provides weak protections and/or poor utility



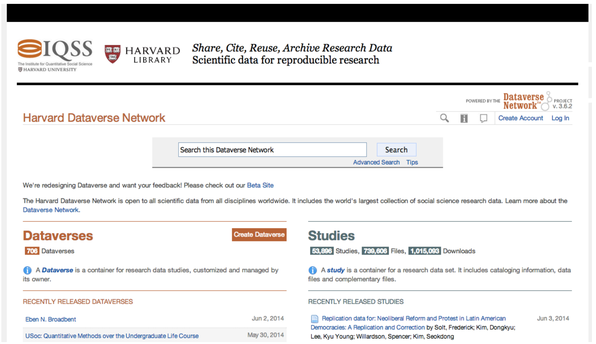
Sweeney `97

Inefficient Process for Obtaining Restricted Data
- Can involve months of negotiation between institutions, original researchers

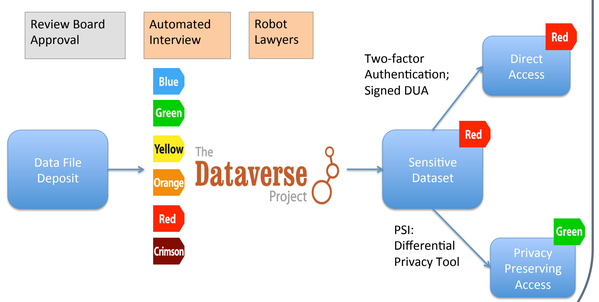
Vision

An array of computational, legal, and policy tools to make privacy-protective data-sharing easier for researchers without expertise in privacy law/CS/stats.

Target: Data Repositories



Approach: Integrated Privacy Tools



Bridging Law & CS Definitions of Privacy

Argue that Differential Privacy Satisfies FERPA and other privacy laws via two arguments:

1. The FERPA privacy standard is relevant for analyses computed with DP
   A legal argument supported by a technical argument

2. Differential privacy satisfies the FERPA privacy standard
   A technical argument supported by a legal argument
   FERPA allows dissemination of de-identified information → sufficient to show that DP analyses result in outcome that is not identifiable
   Extract a mathematical definition of privacy from FERPA and provide a mathematical proof that DP satisfies this definition

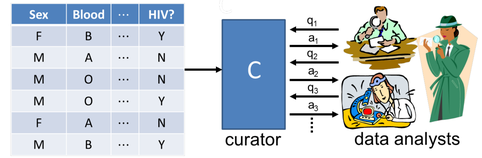
K. Nissim, A. Bembenek, A. Wood, M. Bun, M .Gaboardi, U. Gasser, D. O'Brien, T Steinke, and S. Vadhan. 2016. "Bridging the Gap between Computer Science and Legal Approaches to Privacy." In Privacy Law Scholars Conference (PLSC), 2016.

Broader Impacts

- Infrastructure for research in social science and other human subjects research fields
- Training in multidisciplinary research: ≈100 students, postdocs, interns from law, computer science, social science, statistics
- Policy impact: White House Big Data Privacy Study, National Privacy Research Strategy, NIST 800-188 Deidentifying Government Datasets, Federal Trade Commission
- Numerous workshops and symposia organized, including public symposium "Privacy in a Networked World" w/700+ registrants.
- New journal "Technology Science" utilizing DataTags
- Open-access pedagogical materials on data privacy for many audiences

Differential Privacy Tool:
PSI – A Private data-Sharing Interface

Marco Gaboardi, James Honaker, Gary King, Kobbi Nissim, Jonathan Ullman, and Salil Vadhan. "PSI (Ψ): a Private data Sharing Interface." Poster at Theory and Practice of Differential Privacy (TPDP) and arXiv:1609.04340, 2016.

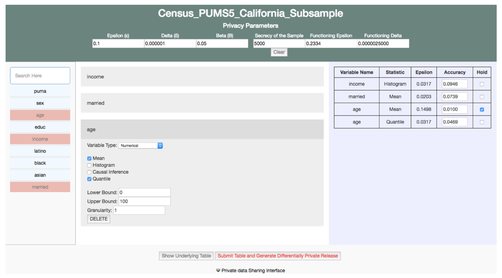


Privacy Definition: effect of each individual must be "hidden" [Dinur-Nissim '03+Dwork, Dwork-Nissim '04, Blum-Dwork-McSherry-Nissim '05, Dwork-McSherry-Nissim-Smith '06]

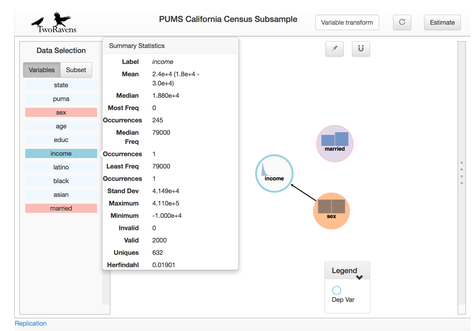
Goals of PSI

- **General-purpose:** applicable to most datasets in repository.
- **Automated:** no differential privacy expert optimizing algorithms for a particular dataset or application
- **Tiered access:** DP interface for wide access to rough statistical information, helping users decide whether to apply for access to raw data (cf. Census PUMS vs RDCs)

Privacy Budgeting Interface



Integration w/Statistical Tools for Social Science



Co-PIs & Senior Personnel

- Kobbi Nissim, co-PI, CRCS & Georgetown
- James Honaker, Sr. Researcher, CRCS
- Micah Altman, co-PI, MIT
- Steve Chong, co-PI, CRCS
- Merce Crosas, co-PI, IQSS
- Urs Gasser, co-PI, Berkman Klein Center
- Latanya Sweeney, co-PI, IQSS
- Edoardo Airoldi, co-PI, Harvard Stats Dept
- Gary King, co-PI, IQSS
- Marco Gaboardi, University of Buffalo
- David O'Brien, Sr. Researcher, Berkman Klein Center

CRCS Center for Research on Computation and Society
at Harvard School of Engineering and Applied Sciences

BERKMAN KLEIN CENTER
FOR INTERNET & SOCIETY AT HARVARD UNIVERSITY

The Institute for Quantitative Social Science at Harvard University

DATA PRIVACY LAB