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.

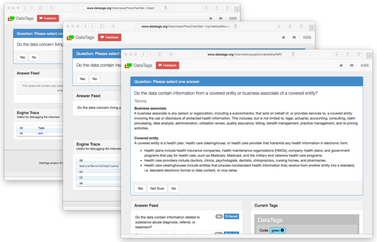
DataTags Levels

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



www.datatags.org/interviews/FlowChartSet-1/start

DataTags Feedback IQSS

Question: Please select one answer

Do the data concern living

Yes No

Answer Feed

This space will contain your answers

Engine Trace

Useful for debugging the interview

Id Type

$0 ask

www.datatags.org/interviews/FlowChartSet-1/q/medicalRecord

Do the data contain hea

Do the data concern living p

Id

medicalRecordsCompliance

$4

$3

$0

www.datatags.org/interviews/questionnaireId/q/MR7

Question: Please select one answer

Do the data contain information from a covered entity or business associate of a covered entity?

Terms

Business associate

A business associate is any person or organization, including a subcontractor, that acts on behalf of, or provides services to, a covered entity involving the use or disclosure of protected health information. This includes, but is not limited to, legal, actuarial, accounting, consulting, claim processing, data analysis, administration, utilization review, quality assurance, billing, benefit management, practice management, and re-pricing activities.

Covered entity

A covered entity is a health plan, health care clearinghouse, or health care provider that transmits any health information in electronic form.

- Health plans include health insurance companies, health maintenance organizations [HMOs], company health plans, and government programs that pay for health care, such as Medicare, Medicaid, and the military and veterans health care programs.
- Health care providers include doctors, clinics, psychologists, dentists, chiropractors, nursing homes, and pharmacies.
- Health care clearinghouses include entities that process nonstandard health information they receive from another entity into a standard, i.e. standard electronic format or data content, or vice versa.

Yes Not Sure No

Answer Feed

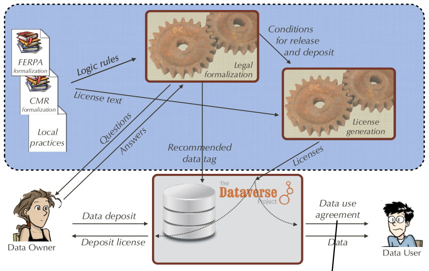
Do the data contain information related to substance abuse diagnosis, referral, or treatment? No Revisit

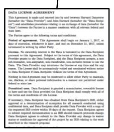
Current Tags

DataTags

Code green

Robot Lawyers





DATA LICENSE AGREEMENT

This Agreement is made and entered into by and between Harvard Dataverse (hereafter the "Data Provider") and John Harvard (hereafter the "Data Recipient") and establishes procedures relating to an exchange of data (hereafter the "Data") between the Parties in a manner consistent with all relevant federal and state laws.

The Parties agree to the following terms and conditions:

Period of agreement. This Agreement shall begin on January 1, 2017, or date of execution, whichever is later, and end on December 31, 2017, unless terminated in writing by either Party.

License. No ownership interest in the Data is bestowed to the Data Recipient pursuant to this Agreement. Subject to the terms of this agreement, the Data Provider grants to the Data Recipient, and the Data Recipient accepts, a non sub-licensable, non-assignable, non-transferable, non-exclusive license to use the Data. The Data Provider may terminate the License at any time and for any reason. The License shall be automatically voided and terminated without notice to Data Recipient if Data Recipient violates the terms of this Agreement.

Nothing in this Agreement may be construed to allow either Party to maintain, use, disclose, or share personal information in a manner not permitted under federal or state law.

Permitted uses. Data Recipient is granted a nonexclusive, revocable license to have and use the Data provided the Data Recipient shall comply with all of the terms and conditions of this License.

If necessary, Data Recipient has obtained from Data Recipient's IRB, either approval or a determination of exemption for all research conducted using confidential data, and Data Recipient shall provide Data Provider with a copy of the requested IRB review within 10 days of the request. Data Recipient agrees to submit required documentation to an IRB should research protocols change. Data Recipient agrees to submit to the Data Provider any change in waiver status or conditions for approval of the project by an IRB relating to the work described in the research proposal.

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

Other Accomplishments

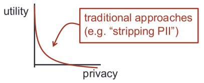
- Many theoretical results illuminating the limits of differential privacy (lower bounds, algorithms, hardness results, attacks).
- Theoretical and empirical work bridging differential privacy & statistical inference (confidence intervals, hypothesis testing, Bayesian posterior sampling).
- Framework 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



privacy
open data

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

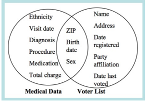
Challenges for Sharing Sensitive Data

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

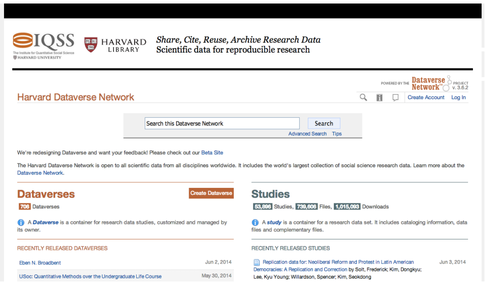
Inefficient Process for Obtaining Restricted Data

- 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



IQSS The Institute for Quantitative Social Science HARVARD UNIVERSITY

HARVARD LIBRARY

Share, Cite, Reuse, Archive Research Data

Scientific data for reproducible research

POWERED BY THE Dataverse Network PROJECT v. 3.6.2

Harvard Dataverse Network

Create Account Log In

Search this Dataverse Network Search

Advanced Search Tips

We're redesigning Dataverse and want your feedback! Please check out our Beta Site

The Harvard Dataverse Network is open to all scientific data from all disciplines worldwide. It includes the world's largest collection of social science research data. Learn more about the Dataverse Network.

Dataverses Create Dataverse

706 Dataverses

A *Dataverse* is a container for research data studies, customized and managed by its owner.

RECENTLY RELEASED DATAVERSES

Eben N. Broadbent Jun 2, 2014

USoc: Quantitative Methods over the Undergraduate Life Course May 30, 2014

Studies

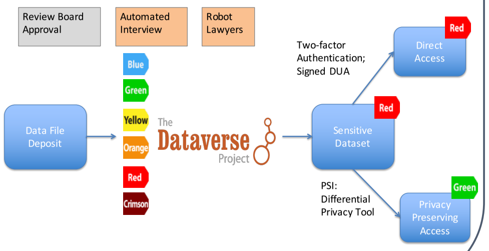
53,896 Studies, 739,606 Files, 1,015,093 Downloads

A *study* is a container for a research data set. It includes cataloging information, data files and complementary files.

RECENTLY RELEASED STUDIES

Replication data for: Neoliberal Reform and Protest in Latin American Democracies: A Replication and Correction by Solt, Frederick; Kim, Dongkyu; Lee, Kyu Young; Willardson, Spencer; Kim, Seokdong Jun 3, 2014

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

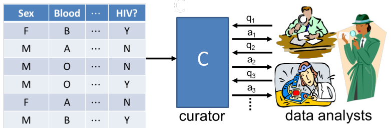
K. Nissim, A. Bembenek, A. Wood, M. Bun, M .Gaboardi, U. Gasser, D. O'Brien, T Steinke, and S. Vadhan. 2016. "Bridging the Gap between Computer Science and Legal Approaches to Privacy." In Privacy Law Scholars Conference (PLSC), 2016.

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

Ψ 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.



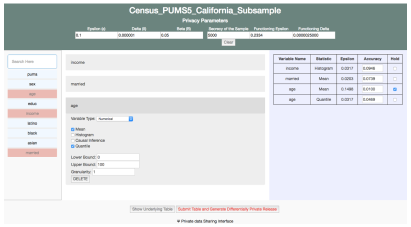
Sex	Blood	⋯	HIV?
F	B	⋯	Y
M	A	⋯	N
M	O	⋯	N
M	O	⋯	Y
F	A	⋯	N
M	B	⋯	Y

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



Census_PUMS5_California_Subsample

Privacy Parameters

Epsilon (ε)	Delta (δ)	Beta (β)	Secrecy of the Sample	Functioning Epsilon	Functioning Delta
0.1	0.000001	0.05	5000	0.2334	0.0000025000

Clear

Search Here

puma
sex
age
educ
income
latino
black
asian
married

income

married

age

Variable Type: Numerical

☑ Mean
☐ Histogram
☐ Causal Inference
☑ Quantile

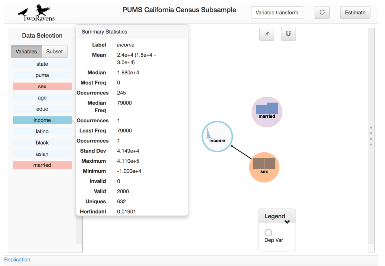
Lower Bound: 0
Upper Bound: 100
Granularity: 1
DELETE

Variable Name	Statistic	Epsilon	Accuracy	Hold
income	Histogram	0.0317	0.0946	☐
married	Mean	0.0203	0.0739	☐
age	Mean	0.1498	0.0100	☑
age	Quantile	0.0317	0.0469	☐

Show Underlying Table | Submit Table and Generate Differentially Private Release

Ψ Private data Sharing Interface

Integration w/Statistical Tools for Social Science



TwoRavens

PUMS California Census Subsample

Variable transform | Estimate

Data Selection

Variables | Subset

state
puma
sex
age
educ
income
latino
black
asian
married

Summary Statistics

Label	*income*
Mean	2.4e+4 (1.8e+4 - 3.0e+4)
Median	1.880e+4
Most Freq	0
Occurrences	245
Median Freq	79000
Occurrences	1
Least Freq	79000
Occurrences	1
Stand Dev	4.149e+4
Maximum	4.110e+5
Minimum	-1.000e+4
Invalid	0
Valid	2000
Uniques	632
Herfindahl	0.01901

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