PRIVACY TOOLS FOR SHARING RESEARCH DATA

Summer 2015 Orientation

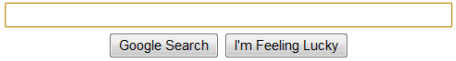
Salil Vadhan

Supported by an NSF Secure & Trustworthy Cyberspace (SaTC) "frontier" grant, a grant from the Sloan Foundation, and a gift from Google.

Computational Social Science

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

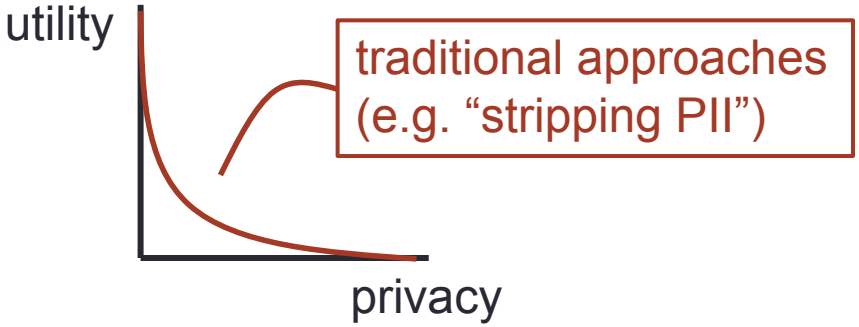
Google

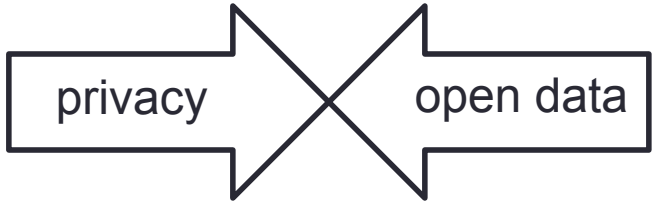


THE HUFFINGTON POST

amazonmechanical turk
Artificial Artificial Intelligence

The problem: protecting the privacy of data subjects



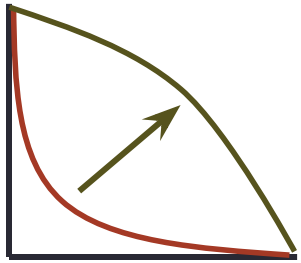


e.g. NYT 5/21/12 “Troves of Personal Data, Forbidden to Researchers”

Our Goal

Achieve:

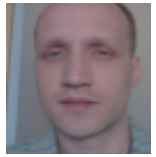
&



Via:

Nissim

Altman
(MIT)



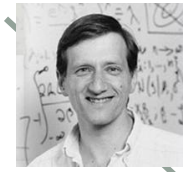
Honaker

CRCS Center for Research on Computation and Society

Chong

Vadhan

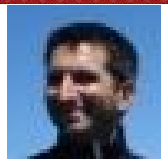
social science



King

IQSS
The Institute for Quantitative Social Science
HARVARD UNIVERSITY

Crosas



Gaboardi

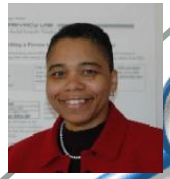
computer science

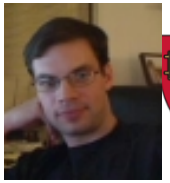
Airoldi

statistics

Gasser

Sweeney



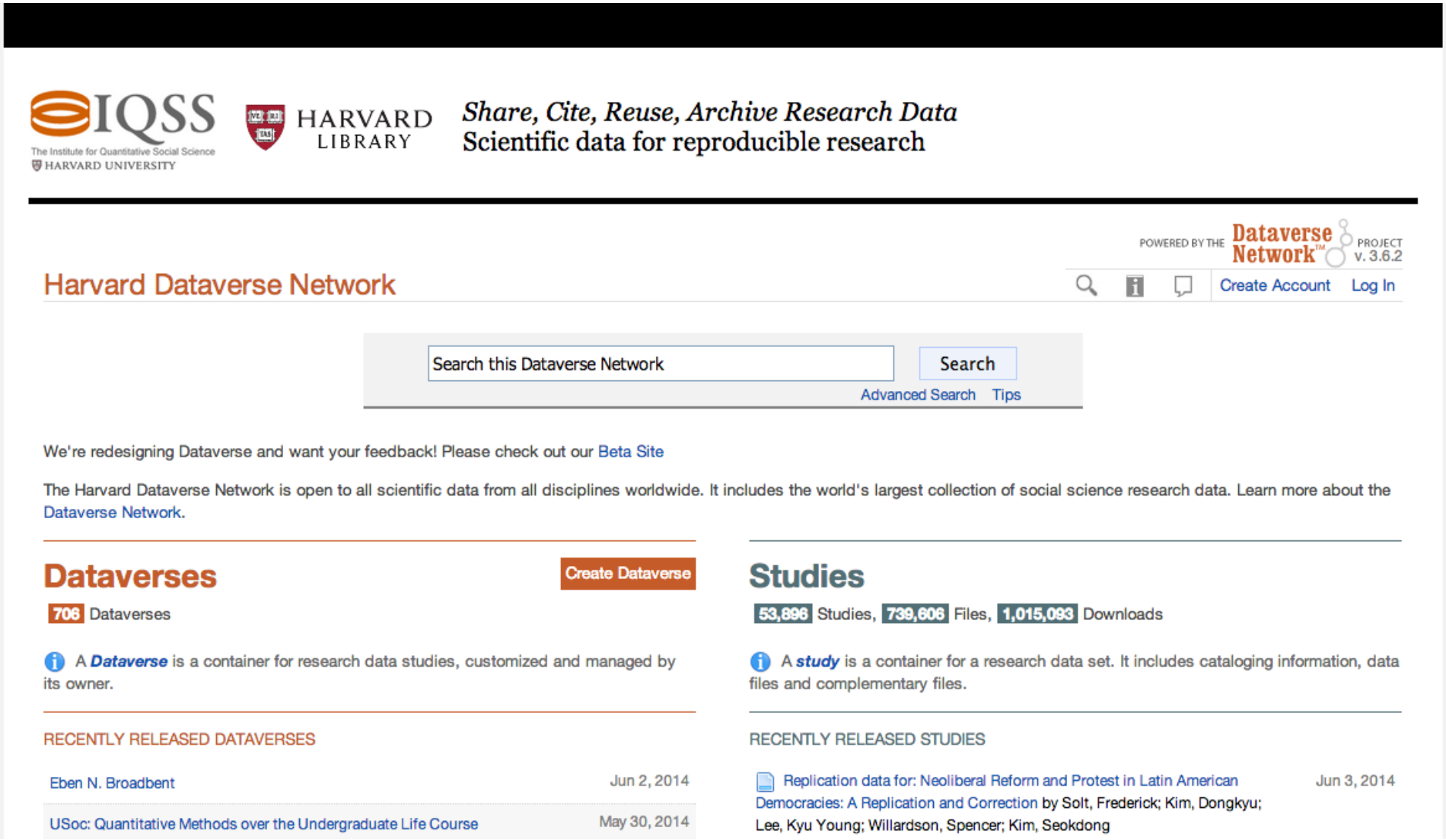


Berkman
The Berkman Center for Internet & Society at Harvard University

law & policy

O'Brien

Use Case: Data Repositories



IQSS Dataverse Network >

POWERED BY THE Dataverse Network™ PROJECT v. 3.2

Murray Research Archive Original Collection Dataverse

Create Account Log In

INTERGENERATIONAL STUDIES, 1932-1982

hdl:1902.1/00627UNF:3:jYQzhUZ5MxpaKGMvIojITA==

Version: 5– Released: Tue Jun 19 13:50:23 EDT 2012

Cataloging Information | **DATA & ANALYSIS** | Comments (0) | Versions

Use the check boxes next to the file name to download multiple files. Data files will be downloaded in their default format. You can also download all the files in a category by checking the box next to the category name. You will be prompted to save a single archive file. Study files that have restricted access will not be downloaded.

Access to some files is restricted, and those files are not available for downloading. Check the Terms of Use for more information.

Select all files Download All Selected Files

1. Documentation

00627IHD-InterGenerational-Co… Adobe PDF - 41 MB - …	Download	Description of coded data variables
00627IHD-InterG… Adobe PDF - 7 M…	Download	Blank measures for study
00627IHD-InterG… Adobe PDF - 173 K…	Download	Overview: abstract, research methodology, publications, and other info.

2. Berkeley Data

00627IHD-InterGenerational-BerkSpou-Data.por SPSS Portable - 29 KB - 0 downloads	Restricted	Data on Spouses in Berkeley Sample in SPSS Portable Format
00627IHD-InterGenerational-BerkSpou-Data.tab Tab Deli… TABUL…	Restricted	Data on Spouses of Berkeley Sample in Tab
00627IHD-InterGenerational-BerkSubj-Data.por SPSS Portable - 217 KB - 0 downloads	Restricted	Data on Subjects in Berkeley Sample in SPSS Portable Format

Many datasets are restricted due to privacy concerns

Goal: enable wider sharing while protecting privacy

Challenges for Sharing Sensitive Data

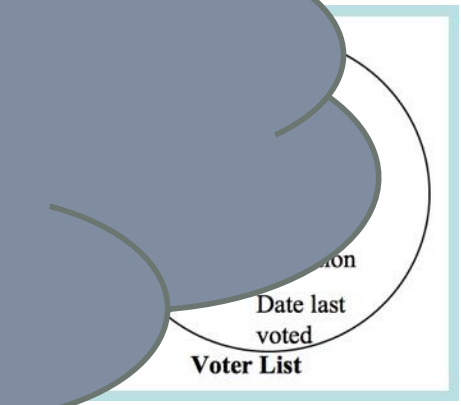
Complexity of Law

- Thousands of privacy laws in the US alone, at federal, state and local level, usually context-specific:
  HIPAA, [illegible] SRA, ….

D

- S[illegible]

Goal: make sharing easier for researcher without expertise in privacy law/cs/stats

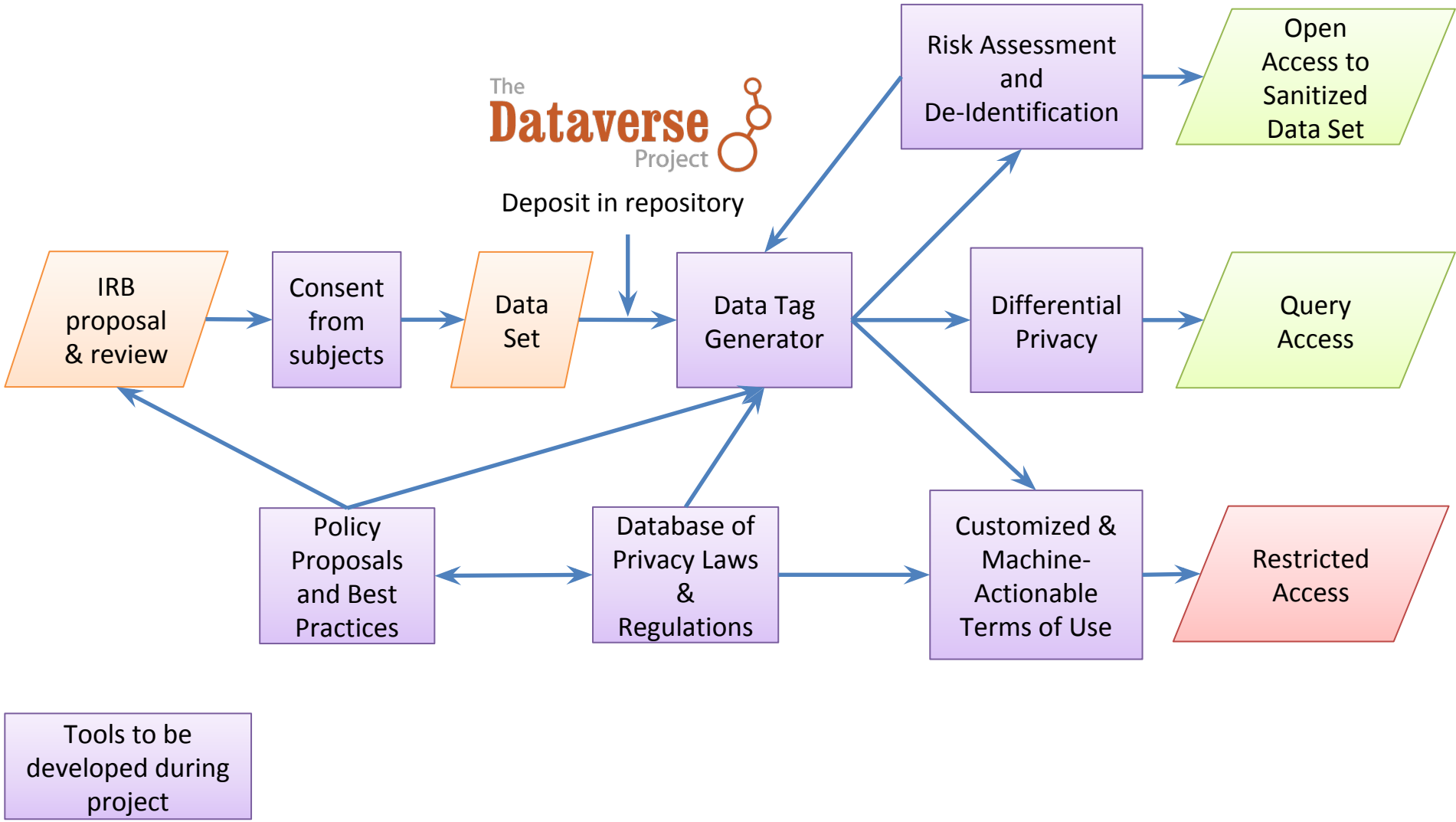


Sweeney `97

Inefficient Process for S[illegible]ng Restricted Data

- Can involve months of negotiation between institutions, original researchers

Vision: Integrated Privacy Tools



Elements of Our Approach

Defining & Measuring Privacy

- Mathematical, legal, statistical, experimental

Defining & Measuring Data Utility

- Enable social scientists to do their work

Tools for Privacy

- Algorithms (e.g. differential privacy)
- Legal instruments & best practices (e.g. data tags)

Education & Outreach

- Open resources
- Multidisciplinary training

Schedule for Next Few Days

- **Tuesday June 9th:** Knafel 262, 1737 Cambridge St
- 1:30-1:45 Introductions
- 1:45-2:15 Project Overview (Salil Vadhan)
- 2:15-2:45 Dataverse (James Honaker)
- 2:45-3:00 Coffee Break
- 3:00-3:30 Legal Overview (Alexandra Wood)
- 3:30-4:00 DataTags & DataTags Demo (Alexandra Wood & Michael Bar-Sinai)
-
- **Wednesday June 10th:** 10:30-1:00 (Knafel 354); 1:00 to 3:30 (Knafel 262,) 1737 Cambridge St
- 10:45-11:15 An overview of key concepts in Confidential Data Management (Micah Altman)
- 11:15-11:45 Differential Privacy for a non-technical audience (Salil Vadhan)
- 12:00-1:00 Lunch
- 1:00 Room Change
- 1:00-2:00 Differential Privacy for a technical audience (Kobbi Nissim)
- 2:00-2:30 Break
- 2:30-3:30 R tutorial (James Honaker)
-
- **Thursday June 11th:** CGIS South, Tsai Auditorium, 1730 Cambridge Street. Dataverse Community Meeting. Please register.
- 9:00-10:00 Lightning Talks: Dataverse Use Cases
- 10:00-10:15 Data Exploration & Analysis: TwoRavens (James Honaker, Vito D'Orazio)
- 10:15-11:30 Break
- 11:30-12:00 Sharing Privacy Sensitive Data: DataTags (Latanya Sweeney, Michael Bar-Sinai)
- 12:00-1:30 Break/opportunity for mentors to meet with mentees.
- 1:30-4:30: Breakout Session (5) on Sharing Sensitive Data (led by Latanya Sweeney & Jonathan Crabtree)
-
- **Friday June 12th**: Maxwell Dworkin 221, 33 Oxford Street
- 9:00- 11:00 Continuation of Differential Privacy (Kobbi Nissim)
- 9:00- 11:00 Differential Privacy Tools and API(Jack Murtagh)

Participating this Summer

- All-hands meetings: Wed 10-11
- Technical tutorials & research meetings on differential privacy, statistics, R: Mon 10-12, Wed 11-1 (also Fri 6/12 9-11, 6/19 10-12)
- “De-identification” working group (bridging CS+law)
- Data Tags working group
- Meetings with your mentors
- Meetings/activities with fellow SEAS/IQSS REU students, Berkman interns
- Group social activities (e.g. group hike)
- Feel free to suggest/organize others!

We look forward to working with all of you!