THE PRIVACY TOOLS PROJECT

December 11, 2017

Salil Vadhan
Harvard University

with support from:

NSF

Alfred P. Sloan FOUNDATION

Google

Motivation: Computational Social Science

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

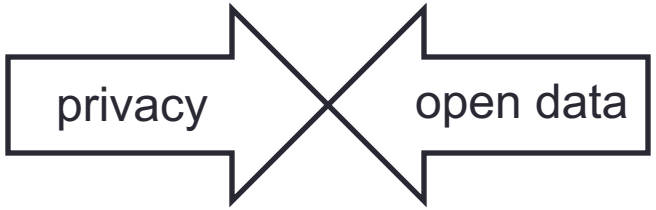
Google

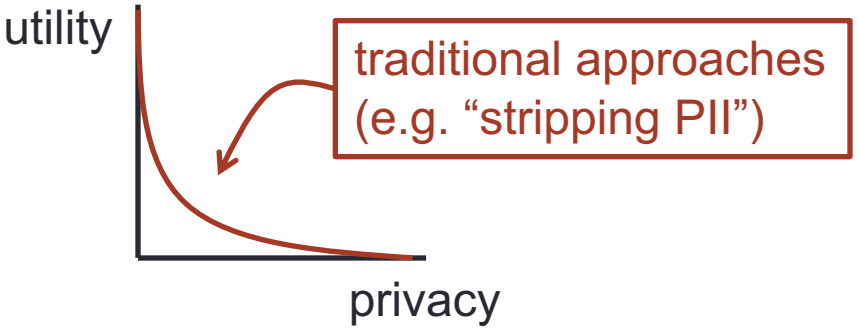
Google Search I'm Feeling Lucky

THE HUFFINGTON POST

amazon mechanical turk
Artificial Artificial Intelligence

The problem: protecting the privacy of individual subjects



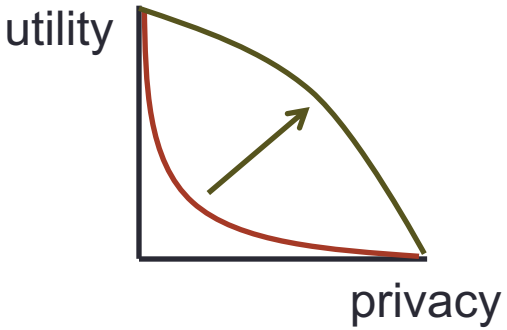


e.g. NYT 5/21/12 “Troves of Personal Data, Forbidden to Researchers”

Our Goal

Achieve:

&



Via:

Program on Information Science
MIT Libraries

social science

CRCS Center for Research on Computation and Society

IQSS
The Institute for Quantitative Social Science
HARVARD UNIVERSITY

computer science

data science

BERKMAN KLEIN CENTER
FOR INTERNET & SOCIETY AT HARVARD UNIVERSITY

law & policy

DATA PRIVACY LAB

Target: Data Repositories

Dataverse

About Guides Support Sign Up Log In

Harvard Datavers

Dataverse Repositories around the world:
27 installations

Harvard Dataverse Repository:
2400 dataverses with 75,000 datasets
and 2.9 million downloads
Largest social science repository in the world

Metrics 2,104,594 Downloads

Share, archive, and get elds.

Search this dataverse...

Add Data

Dataverses (1,979)
Datasets (63,985)
Files (365,114)

Dataverse Category
Research Project (543)
Researcher (543)
Organization or Institution (168)
Journal (134)
Research Group (21)

More...

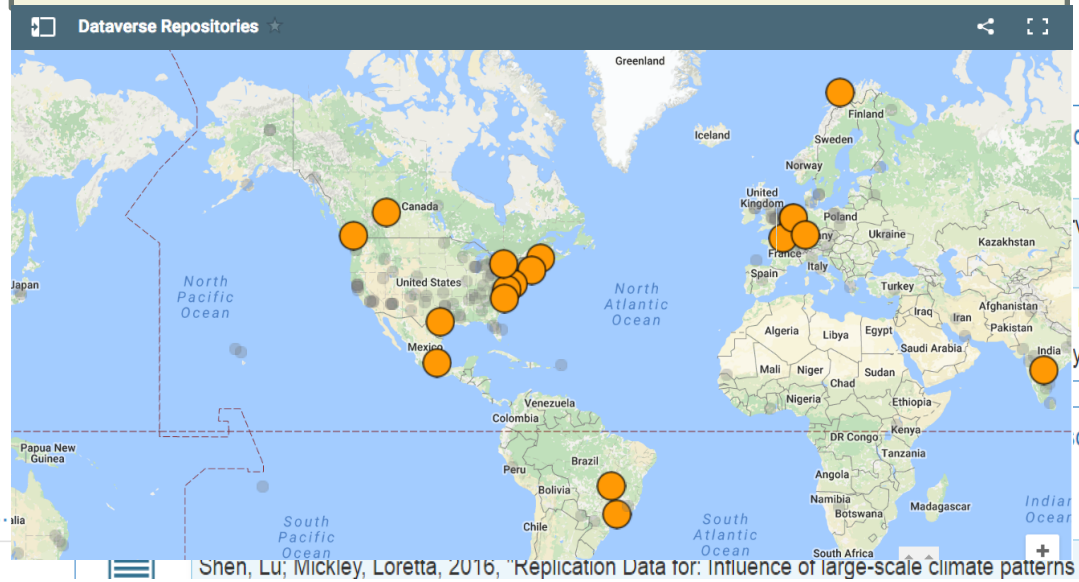
Sort

d Species

vation of a Threatened Species",

Resource for the Conservation of a
y RW, Kusumi K

onal predictive model for air



Shen, Lu; Mickley, Loretta, 2016, "Replication Data for: Influence of large-scale climate patterns on summertime U.S. ozone: A

Metadata Source

Dataverse

About Guides Support Sign Up Log In

Henry A. Murray Research Archive at Harvard University

Murray Research Archive Dataverse (Harvard University) Home

Harvard Dataverse > Murray Research Archive Dataverse > **Two Generations of College-Educated Women: The Post-parental Phase of the Life Cycle, 1957-1979**

View Dataset Versions | Metrics | 60 Downloads

Two Generations of College-Educated Women: The Post-parental Phase of the Life Cycle, 1957-1979

Ida Davidoff; Marjorie Platt, 2007, "Two Generations of College-Educated Women: The Post-parental Phase of the Life Cycle, 1957-1979", http://hdl.handle.net/1902.1/00511, Harvard Dataverse, V2

Download Citation

If you use these data, please add this citation to your scholarly resources. Learn about Data Citation Standards.

Download | Request Access

7 Files

00511Davidoff-Platt-BoxCoverSheets.pdf
Adobe PDF - 406.7 KB - Nov 27, 2007 - 12 Downloads
Describes contents of each box of a paper data set
3. Supplementary Documentation
Download

00511Davidoff-Platt-Codebook.pdf
Adobe PDF - 27.7 MB - Nov 27, 2007 - 12 Downloads
Description of coded data variables
1. Documentation
Download

00511Davidoff-Platt-Data.por
SPSS Portable - 221.2 KB - Nov 27, 2007 - 0 Downloads
Data for Study in SPSS Portable Format
2. Data

00511Davidoff-Platt-Data.tab

00511Davidoff-Platt-Measures.pdf

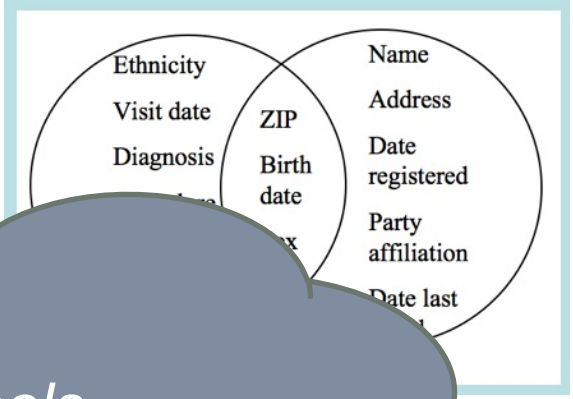
Datasets are restricted due to privacy concerns

Goal: enable wider sharing while protecting privacy

Challenges for Sharing Sensitive Data

Difficulty of Deidentification

- Stripping "PII" usually provides weak protections



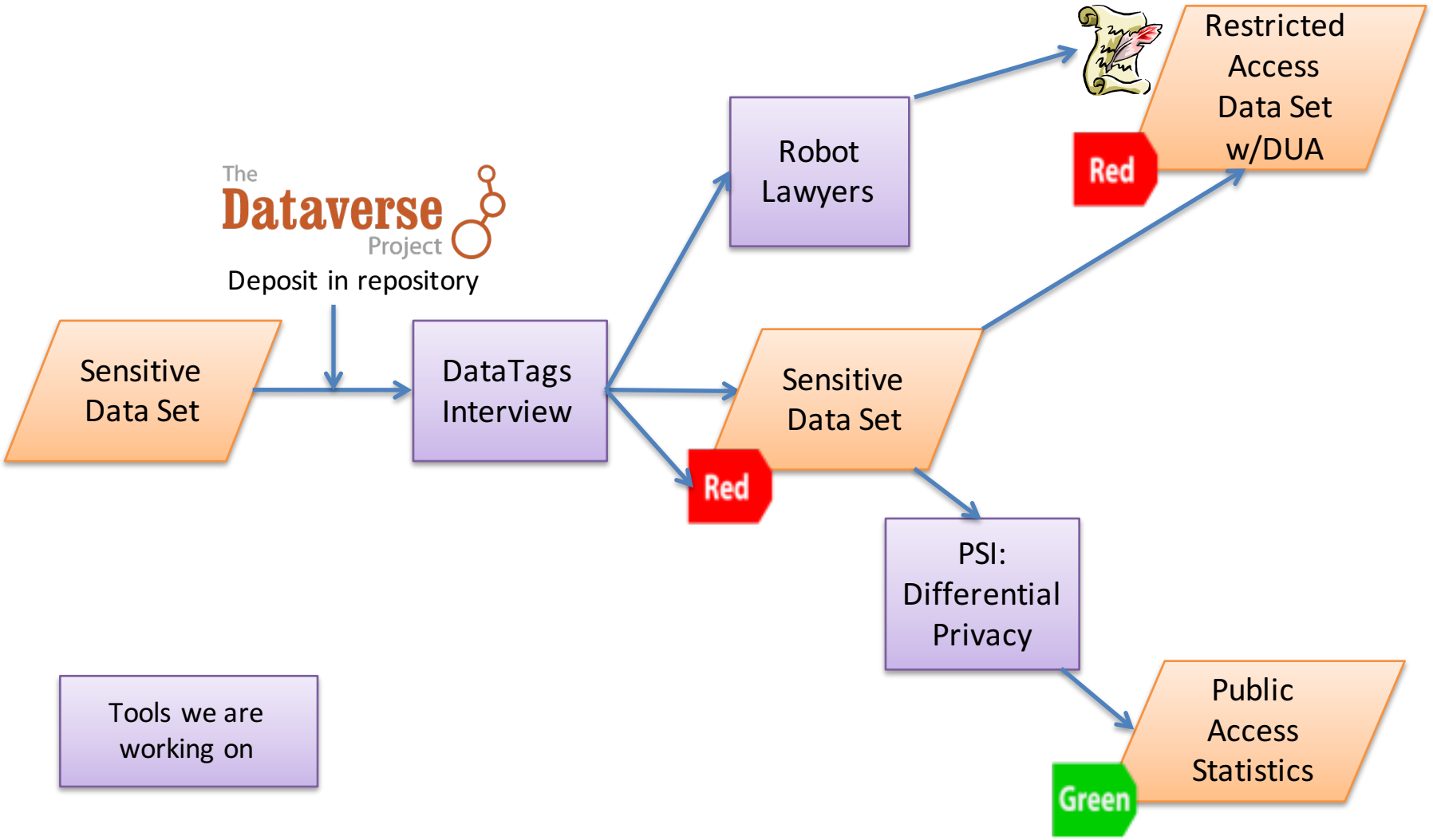
Sweeney `97

Vision: array of *computational, legal, policy tools* that make *privacy-protective data-sharing* easier for researchers without expertise in privacy law/cs/stats.

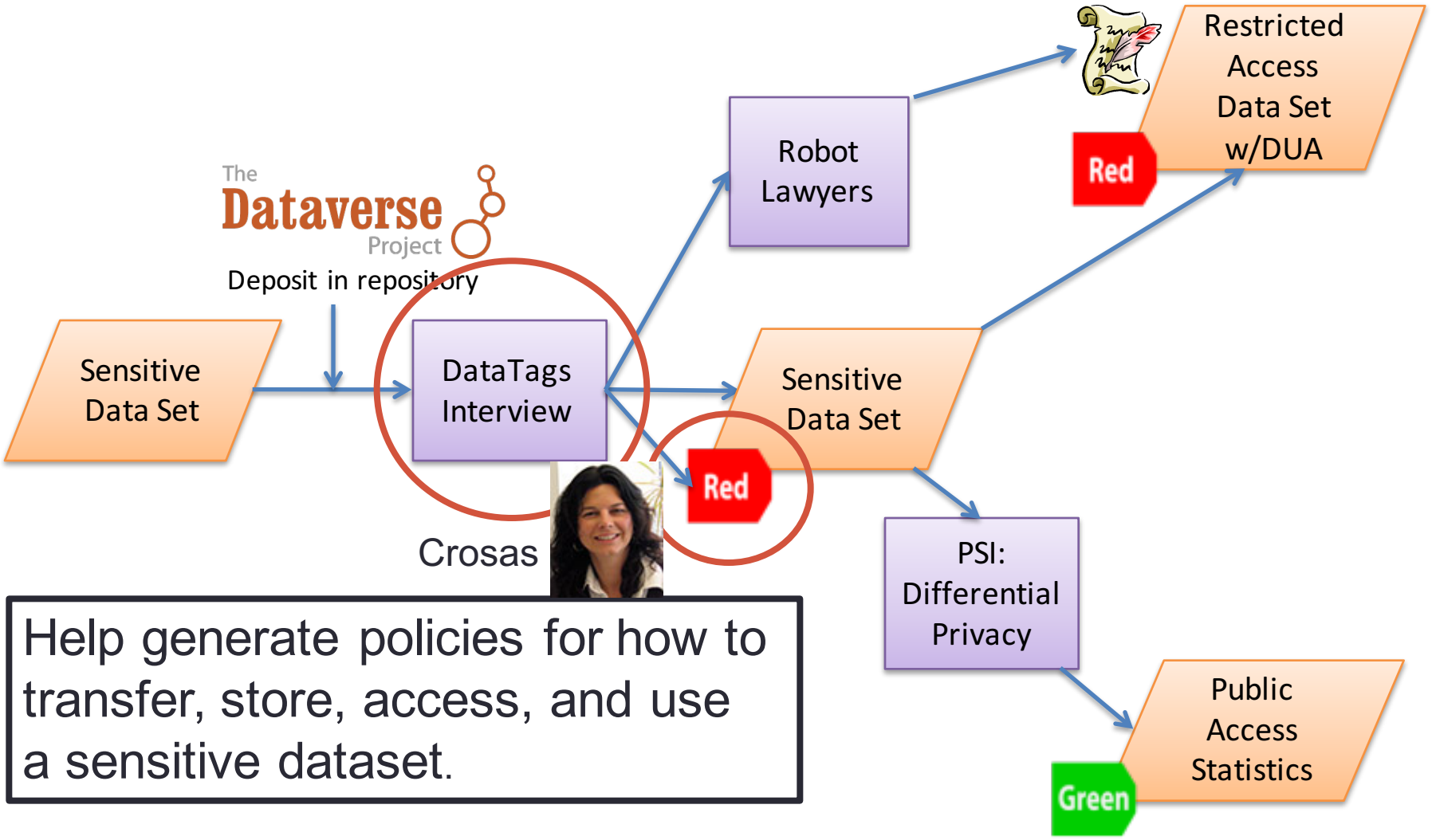
Complexity

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

Approach: Integrated Privacy Tools

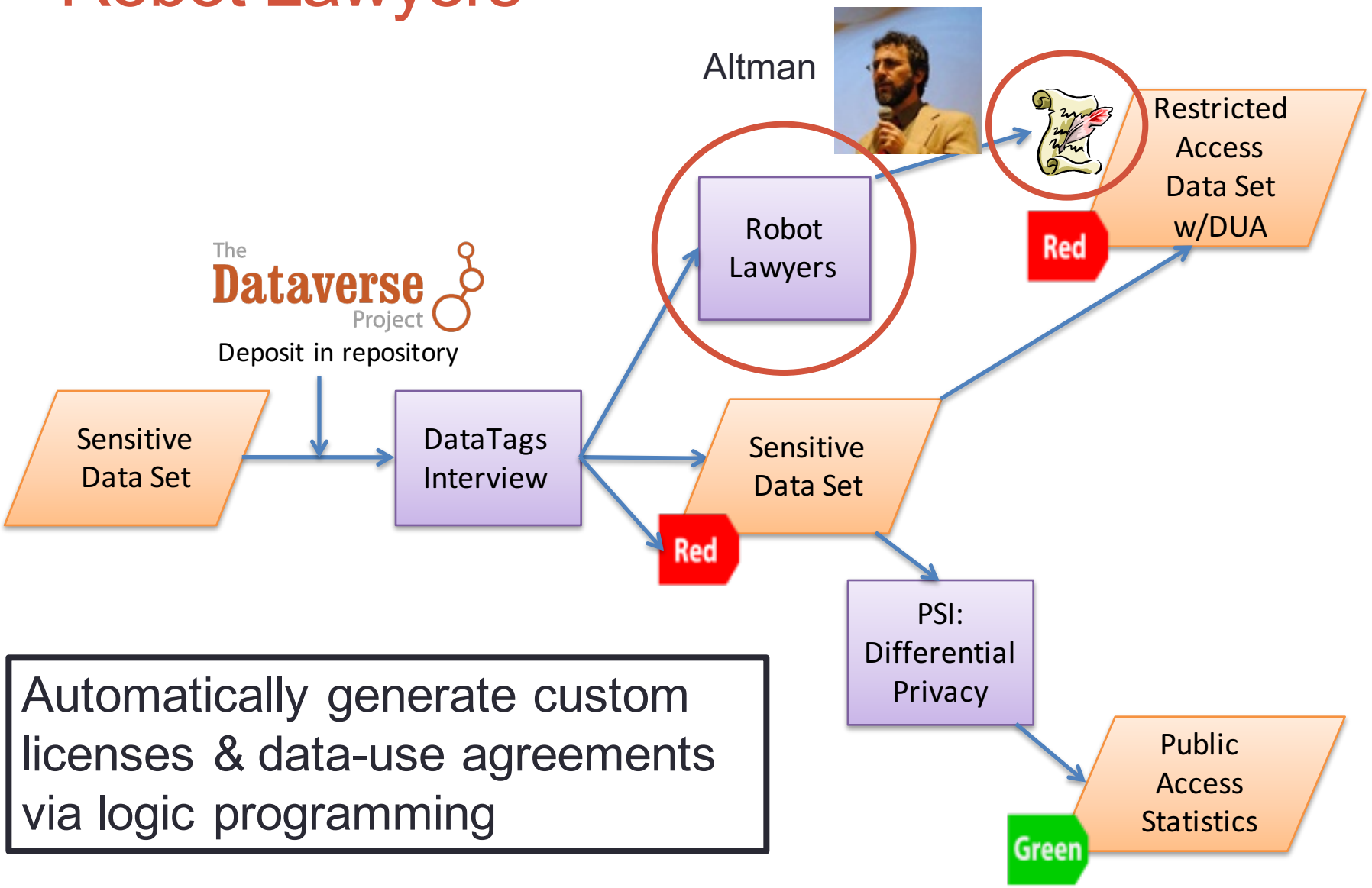


DataTags



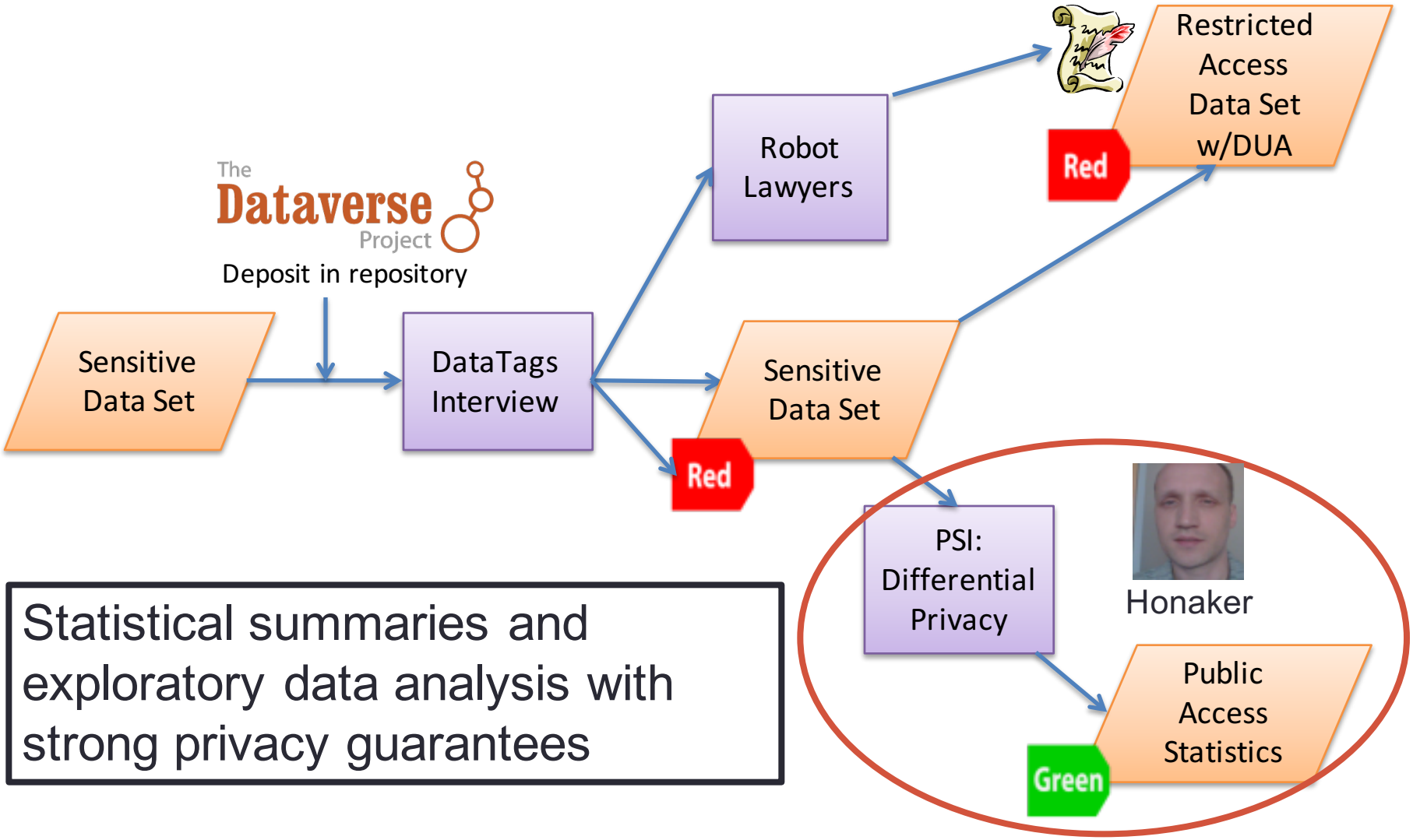
Help generate policies for how to transfer, store, access, and use a sensitive dataset.

Robot Lawyers



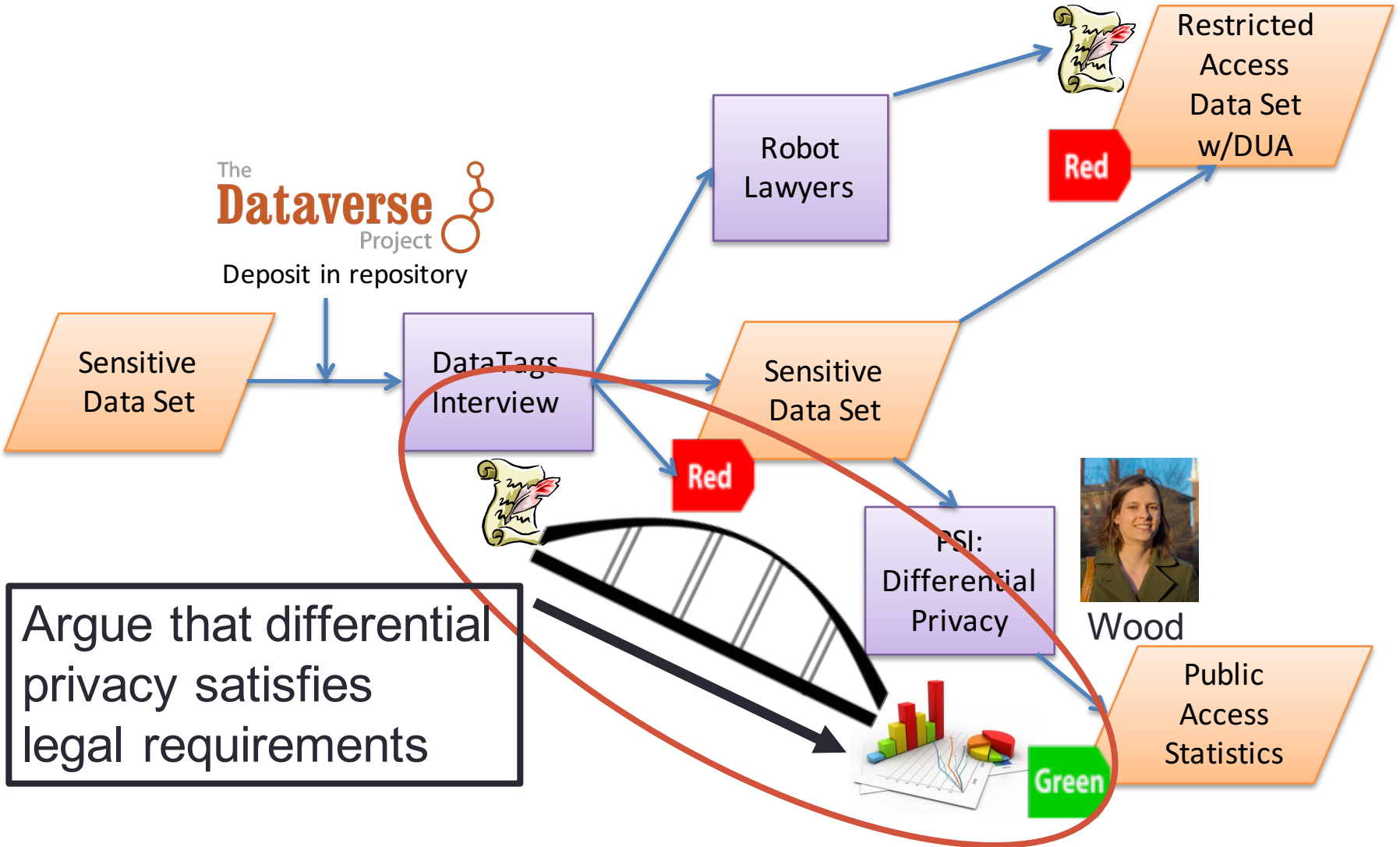
Automatically generate custom licenses & data-use agreements via logic programming

PSI: Differential Privacy Tool

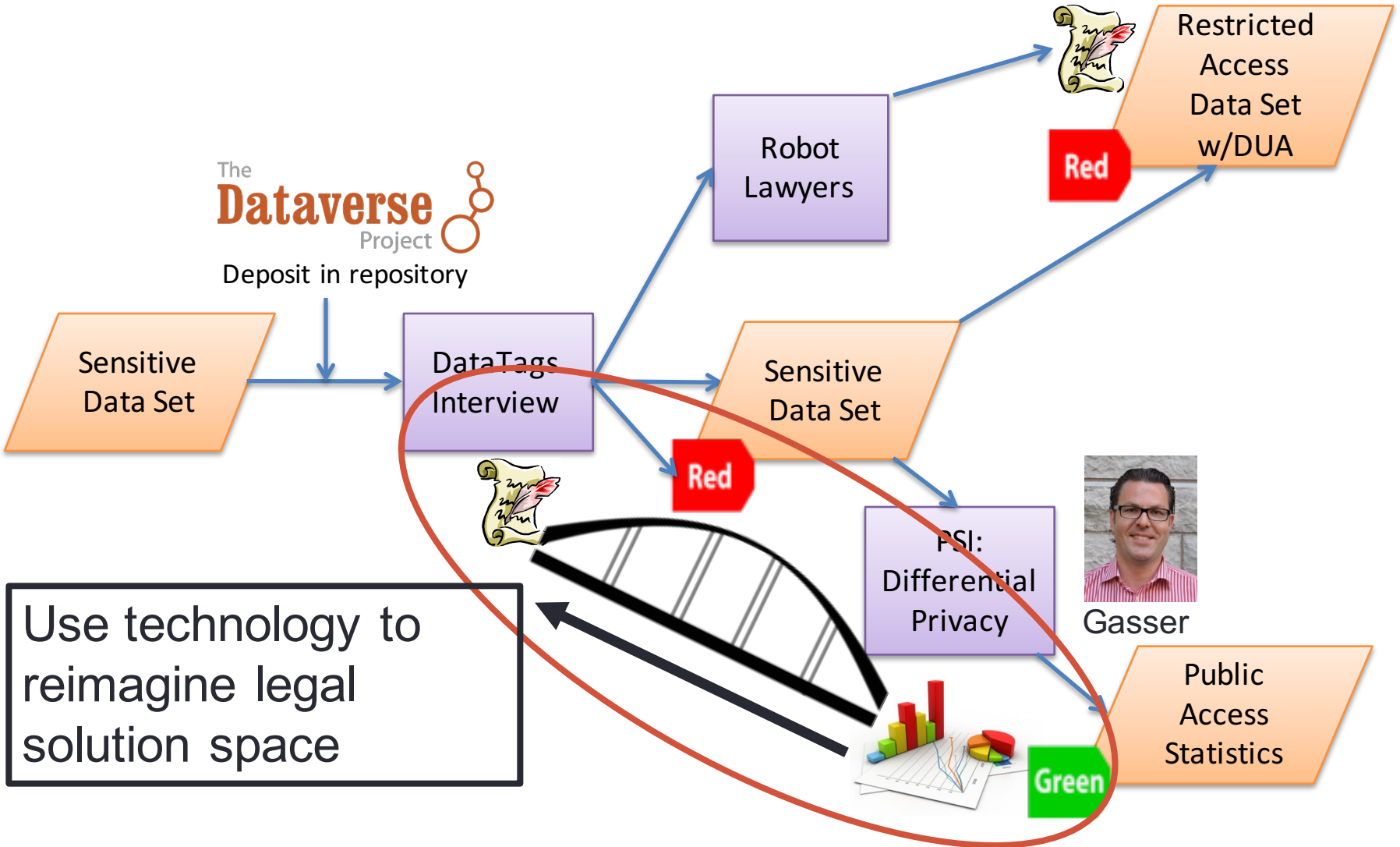


Statistical summaries and exploratory data analysis with strong privacy guarantees

Bridging Definitions of Privacy



Recoding Privacy Law



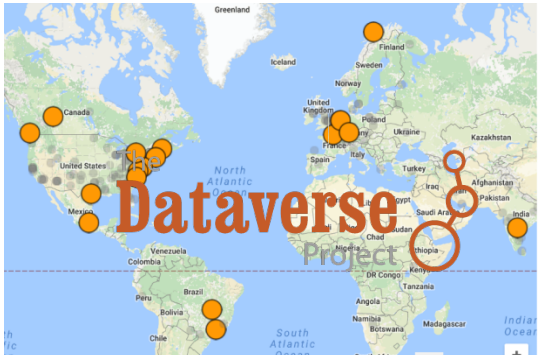
Broader Impacts: Overarching Goals

- Exposing a multidisciplinary understanding of data privacy to a wide range of audiences (students, policymakers, public)

- Bringing integrated solutions to data privacy problems to practice (focusing on data repositories and computational social science)

Broader Impacts

Infrastructure for research in social science and other human subjects research fields

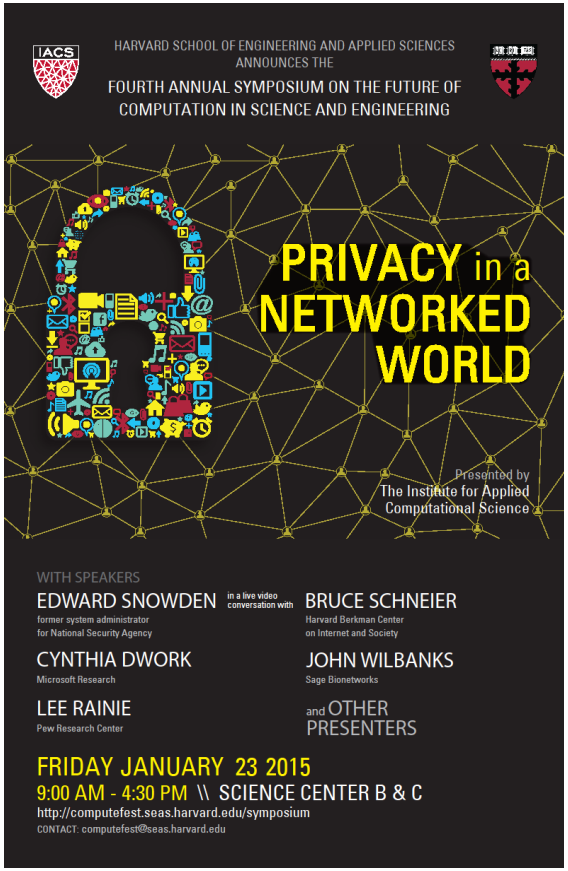


Training in multidisciplinary research: ≈120 students, postdocs, interns from law, computer science, social science, stats

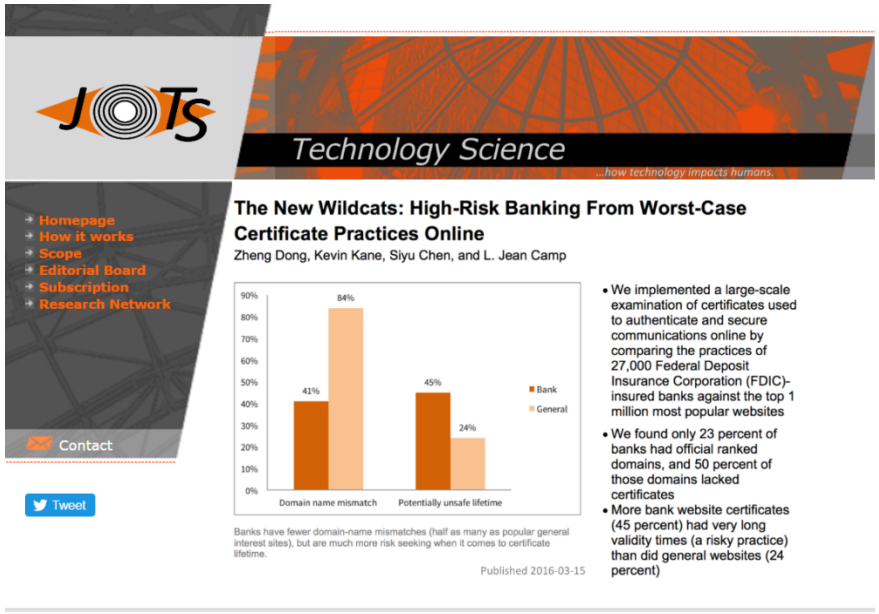
Policy impact: White House Big Data Privacy Study, National Privacy Research Strategy, NIST 800-188 Deidentifying Government Datasets, …

Broader Impacts

Numerous workshops and symposia including public symposium with 700+ registrants.



New journal "Technology Science" utilizing DataTags



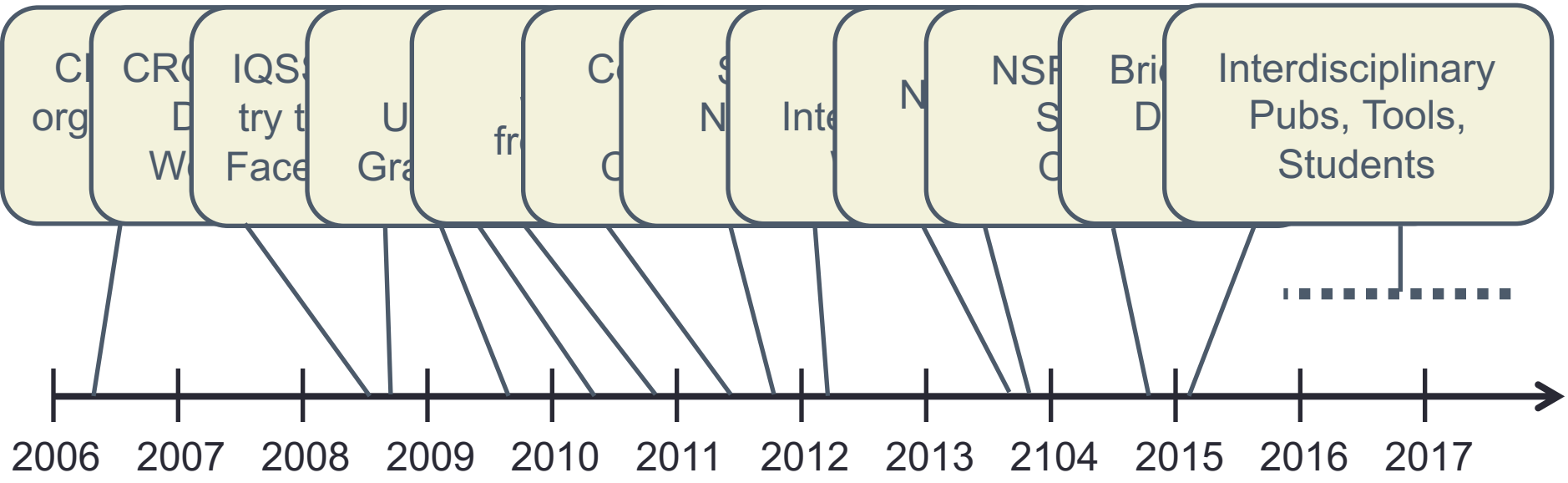
Open-access pedagogical materials on data privacy for many audiences

Other Accomplishments

- Many **theoretical results** illuminating the limits of differential privacy (lower bounds, algorithms, hardness results, attacks).
- **Bridging differential privacy & statistical inference** (confidence intervals, hypothesis testing, Bayesian 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.

Lessons Learned: Interdisciplinary Research

- Interdisciplinary centers to seed efforts
- Shared motivating problem
- Funding is hard
- Policy commentary as a collaboration vehicle
- Large & broad grant crucial
- Building a community
- Value of external critiques
- Creating safe environments
- It takes time!



Lessons Learned: Theory vs. Practice

(Caricature of) our initial proposal:

1. Solve biggest open theory problem in differential privacy literature.
2. Have summer interns implement our solution.
3. Privacy-protective data-sharing solved!

Reality:

- Asymptotic theoretical performance ≠ Practical performance
- Even simplest theory solutions introduce challenges in practice
  ⇒ more interesting theory problems!
- Can't rely solely on interns for tool development
- Long path from research prototypes to production software
- Institutional challenges to sharing sensitive data

Our Goals for Today

- Share where we've come in the Privacy Tools Project.
- Hear about related efforts & challenges.
- Find collaborators, users.

Discuss directions forward:

- **"Applying Theoretical Advances in Privacy to Computational Social Science Practice"**

Alfred P. Sloan FOUNDATION

- **"Computing over Distributed Sensitive Data"**

NSF

Chong

- **"Formal Privacy Models and Title 13"**

Nissim

United States Census Bureau

- **Production-level Tools** for long-term, wide use (in planning)