Matching Uses and Protections for Government Data Releases

Micah Altman
<micah_altman@alumni.brown.edu>
Center for Research in Equitable and Open Scholarship, MIT

Prepared for:

Data Privacy: From Foundations to Applications

Simons Institute

U.C. Berkeley

March 2019

Abstract

In this talk we describe work-in progress that aims to align emerging methods of data protections with research uses. We use the *American Community Survey* as an exemplar case for examining the range of ways that government data is used for research. We identify the range of research uses by combining evidence of use from multiple sources including research articles; national and local media coverage; social media; and research proposals. We then employ human and computer-assisted coding methods to characterize the range of data analysis methodologies that researchers employ. Then, building on previous work cataloging that surveys and characterizes computational and technical controls for privacy, we match these methods to available and emerging privacy and data security controls. Our preliminary analysis suggests that tiered-access to government data will be necessary to support current and new research in the social and health sciences.

Attribution Statement

Co-Conspirators:

- Cavan Capps, U.S. Census
- Zachary Lizee, U. Mass Boston
- Dylan Sam, Brown U.

Project Collaborators:

- Urs Gasser, David O'Brien, Ron Prevost, Salil Vadhan, and the Harvard University Privacy Tools Project

Sponsors:

- Supported in part by the Sloan Foundation

Disclaimer

These opinions are my own, they are not the opinions of my employers, collaborators, or project funders.

Secondary disclaimer:

"It's tough to make predictions, especially about the future!"

- Attributed to Woody Allen, Yogi Berra, Niels Bohr, Vint Cerf, Winston Churchill, Confucius, Disreali [sic], Freeman Dyson, Cecil B. Demille, Albert Einstein, Enrico Fermi, Edgar R. Fiedler, Bob Fourer, Sam Goldwyn, Allan Lamport, Groucho Marx, Dan Quayle, George Bernard Shaw, Casey Stengel, Will Rogers, M. Taub, Mark Twain, Kerr L. White, etc.

Related Work

- Altman, M., Wood, A., O'Brien, D. R., Vadhan, S., & Gasser, U. (2015). Towards a modern approach to privacy-aware government data releases. *Berkeley Technology Law Journal*, 30(3), 1967-2072. https://doi.org/10.15779/Z38FG17

- Altman, Micah and Capps, Cavan and Prevost, Ronald, *Location Confidentiality and Official Surveys* (March 31, 2016). Available at SSRN: https://ssrn.com/abstract=2757737 or http://dx.doi.org/10.2139/ssrn.2757737

- Altman, M., Wood, A., O'Brien, D. R.,, & Gasser, U., Practical approaches to big data privacy over time, *International Data Privacy Law*, Volume 8, Issue 1,, Pages 29–51, https://doi.org/10.1093/idpl/ipx027

Broader Context

Of

Use and Risk

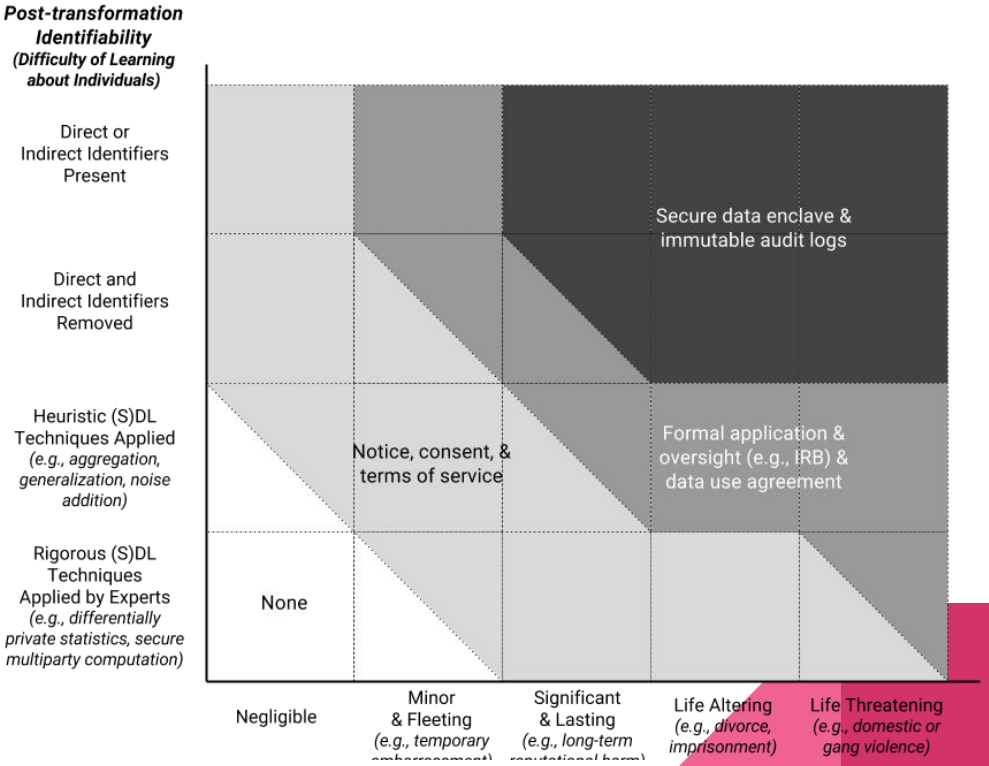
for Government Information

Functions of Government Information

- Official decision & communications
- Broader social benefit (research and business uses)
- Public transparency and accountability

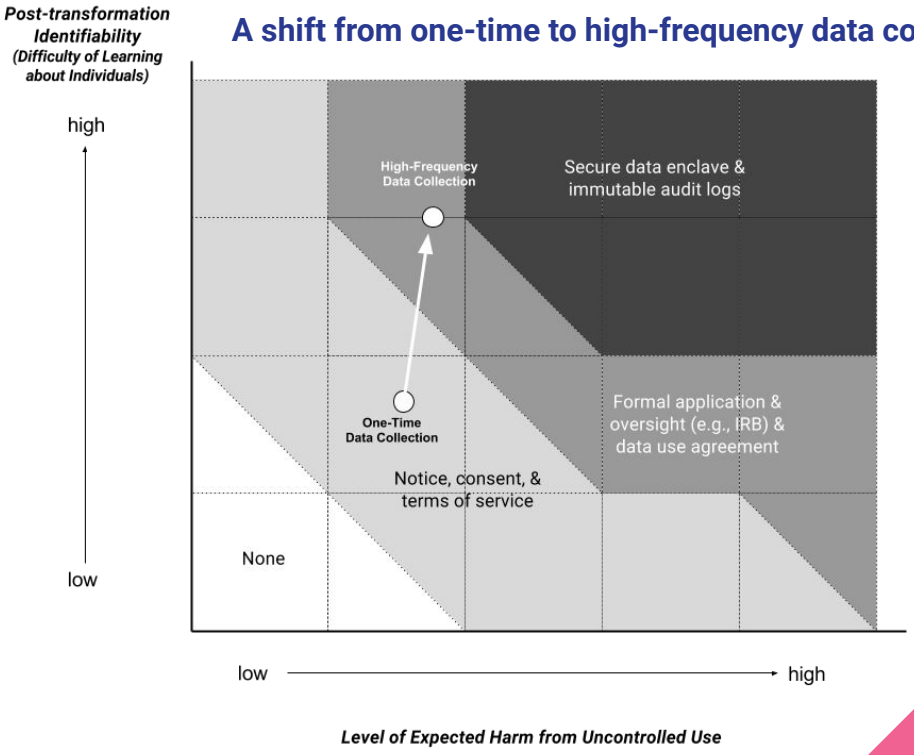
Harm Depends on Privacy & Sensitivity

Illustrating how to choose privacy controls that are consistent with the uses, threats, and vulnerabilities at each lifecycle stage



Changes in Data Collection, Environment Change Identifiability, Threats and Vulnerabilities

A shift from one-time to high-frequency data collection.



Example temporal risk factors for big data

	Identifiability	Threats (sensitivity)	Vulnerabilities (sensitivity)
Age	Small decrease	Moderate increase	Moderate decrease
Period	Small increase	Moderate increase	No substantial evidence of effect
Frequency	Large increase	Small increase	No substantial evidence of effect

Approaches to Managing Informational Harm

Informational controls

- Procedural, technical, educational, economic, and legal means for enhancing privacy can be applied at different stages

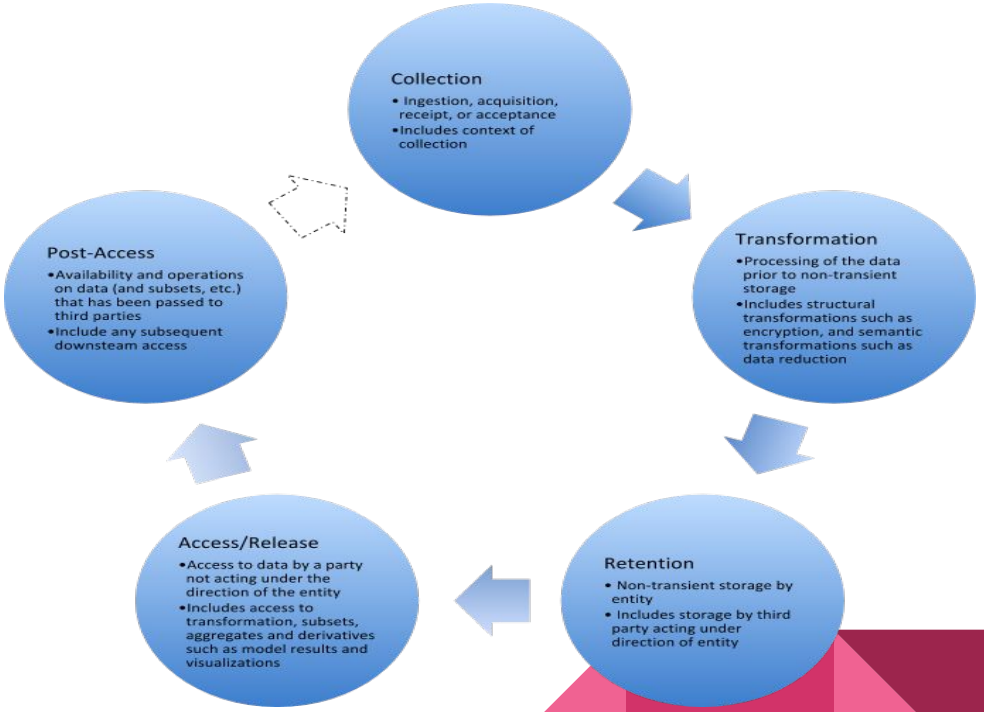
Table 1: Example categorization of privacy controls and interventions.

		Procedural	Economic	Educational	Legal	Technical
Collection/ Acceptance		• Collection Limitation; Data Minimization; Data Protection Officer; • Institutional Review Boards; Notice and Consent Procedures; • Purpose Specification; • Privacy Impact Assessments	• Collection Fees; • Markets for Personal Data; • Property Rights Assignment	• Consumer Education; • Transparency; • Notice; • Nutrition Labels; • Public Education; • Privacy Icons	• Data Minimization; • Notice and Consent Requirements; • Purpose Specification	• Computable Policy
Transformation		• Process for Correction		• Mistakes; • Transparency	• Right To Correct Information; • Safe Harbor De-Identification Standards	• Aggregate Statistics; • Computable Policy; • Contingency Plans; • Data Visualizations; • Differentially Private Data Releases; • Reduction; • SDL Techniques; • Synthetic Data
Retention		• Audits; • Controlled Backups; • Purpose Specification; • Security Assessments; • Tethering		• Data Asset Registers; • Notice; • Transparency	• Breach Reporting; • Data Retention Requirements; • Requirements; • Integrity and Accuracy Requirements	• Computable Policy; • Encryption; • Key Management (and Recur Hashing); • Federated Databases; • Personal Data Stores
Access/ Release		• Access Controls; • Consent; • Expert Panels; • Individual Privacy Settings; • Presumption of Openness vs. Privacy; • Purpose Specification; • Registration; • Risk Assessments	• Access/Use Fees (for Data Controller or Subjects); • Property Rights Assignment	• Data Asset Registers; • Notice; • Transparency	• Integrity and Accuracy Requirements; • Data Use Agreements (Contract with Data Recipient); • Terms of Service	• Authentication; • Computable Policy; • Differential Privacy; • Encryption (Incl. Functional, Homomorphic); • Interactive Query Systems; • Secure Multiparty Computation
Post-Access/ Retention		• Audit Procedures; • Ethical Codes; • Tethering	• Fines	• Privacy Dashboard; • Transparency	• Civil/Criminal Penalties; • Data Use Agreements; • Terms of Service; • Private Right of Action	• Computable Policy; • Computable Audit Log; • Personal Data Stores

	Procedural	Economic	Educational	Legal	Technical
Access/Release	Access controls; Consent; Expert panels; Individual privacy settings; Presumption of openness vs. privacy; Purpose specification; Registration; Restrictions on use by data controller; Risk assessments	Access/Use fees (for data controller or subjects); Property rights assignment	Data asset registers; Notice; Transparency	Integrity and accuracy requirements; Data use agreements (contract with data recipient)/ Terms of service	Authentication; Computable policy; Differential privacy; Encryption (incl. functional, homomorphic); Interactive query systems; Secure multiparty computation

Sequencing controls

- Review of uses, threats, and vulnerabilities as information is used over time
-
- Select appropriate controls at each stage



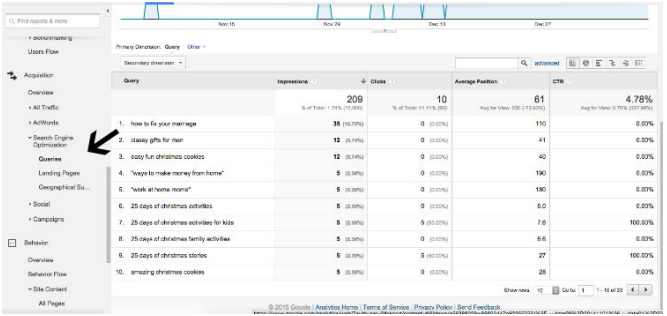
What do technical controls control?

- **Controls on computation --**
  limit the direct computations that can be meaningfully performed
  - Common: file-level encryption, interactive-analysis systems (model servers)
  - Emerging: functional encryption, homomorphic encryption, secure public ledgers, personal data stores
- **Controls on inference**
  limit how learning from computations about the observed units
  - Common: redaction, SDL
  - Emerging: differentially private mechanisms
- **Controls on purpose**
  limit the domain of human activity to which inferences are applied.
  - Common: legal mechanisms
  - Emerging: executable policy languages; machine actionable taxonomies, personal data stores

Examples of appropriate privacy and security controls based on the risk drivers and intended mode of analysis identified a big data use case.

		Big Data Risk Drivers **Lower Risk → Higher Risk**			
		Age, Period, Sample Size, Population Diversity	*High Dimensional*	*High Frequency*	*High Dimensional & High Frequency*
Intended Mode of Analysis	*Population-level Statistical Analysis*	Notice, Consent, Terms of Service; Formal Oversight	Differential Privacy; Formal Oversight		Secure Data Enclave/Model Server; Restricted Access; Formal Oversight
	Individual Analytics		Personal Data Stores; Blockchain Audit Logs; Secure Multiparty Computation; Formal Oversight		

Characterizing Traces of Use



Access Logs

Published Analyses

Media and Social Media Mentions

Matching Uses and Protections

(Exploratory, Preliminary)

Identifying Current Modes of Dissemination

Published Estimates
- Official Indicators
- Pre-computed published tables

Quick Lookups
- Interactive queries to find a single number or table
- Based on pre-computed tables

Dynamic Tables & Maps
- Public interactive servers
- Based on public use tabulations or micro-data

Public Use Tabulations
- Aggregated to pre-defined geographical or logical units
- Processed statistical disclosure limitation methods
- Based on protected micro-data

Public Use Micro-data
- Processed with SDL: deidentification, sampling, synthetic data
- In rare cases, synthetic data used
- Based on protected micro-data

Protected Micro-data
- Possibly identified
- Available within Research Data Centers

What could this inform?

- Data prep
  - External data sources used
  - Cleaning - level
  - Linking - level
- Statistical computing approach
  - Sum queries/univariate method
  - Linear models/GLM
  - Likelihood
  - Bayesian estimates
- Diagnostics
  - Summary diagnostics
  - Sensitivity analysis
  - Individual outliers
- Desired purpose
  - Research
  - Policy
  - Commercial
  - Education
- Data Characteristics
  - What ACS measures used
  - ACS Unit of analysis
  - Study unity of analysis
  - Time dimensions
  - Other Structure
    - Network
    - Textual
    - Spatial
    - Video
- Presentation characteristics
  - Summary/regression
  - Individual cases/plots
- Replication
  - Results
  - Full

Conclusions

and/or

Provocations

Summary

- One size does not fit all
  -- anticipate that tiered access will be necessary to address major uses
- Government data supports several objectives
  -- government decision & communication; broader social benefit (research and economy); transparency and accountability
- Informational controls vary in compatibility --
  controls should be matched to objectives and modes of analysis

Provocations & Vigorous Hand Waving

- *Discovery research (currently) requires access beyond limits of formal protections*
  -- empirically guided exploratory research, theory generation, process tracing, novel syntheses (etc.) are incompletely understood and formalized

- *A representative use isn't*
  -- need to consider multiple uses and tensions between these to get substantial social benefit avoid substantial harms

- *Worst-case analyses aren't*
  -- some formal (DP) and legal analysis (Title 13) take worst case approach to inferential risk, but…
  -- apply average-case analysis to use/utility reduction
  -- are optimistic about operational/implementation risks

Non-temporal risk factors of big data also affect privacy risk components in different ways.

High-dimensional data pose challenges for traditional privacy approaches such as de-identification, and can support new uses of data that were unforeseen at the time of collection.

Broader analytic uses, such as the use of data for personalized classification, and both traditional and modern approaches to de-identification fail to protect against learning facts about populations that could be used to discriminate.

Increases in sample size and diversity lead to heightened risks that a target individual is included, vulnerable populations are included, and a wide range of threats are plausible.

Questions? Observations? Arguments?*

Now

(10 minutes)

or

Later

micah_altman@alumni.brown.edu

micahaltman.com

*£1 for a five minute argument, but only £8 for a course of ten.