Legal Standards for De-identification

Alexandra Wood

Berkman Klein Center for Internet & Society at Harvard University

Presentation before the Commission on Evidence-Based Policymaking
February 24, 2017

These opinions are my own. They are not the opinions of the Berkman Klein Center, any of our funders, nor (with the exception of co-authorship on previously published work) my collaborators.

Evolving landscape for government data releases

- Government agencies are making efforts to release more information to the public for a wide range of purposes from transparency and accountability to scientific research and innovation.
- Releasing data about individuals inherently carries privacy risks.
- De-identification has long been used to enable the release of data while addressing privacy concerns.
- However, scientific understanding of privacy is evolving and traditional approaches to de-identification are increasingly shown to be inadequate.

Overview of US legal framework for de-identification

- De-identification standards are highly sector- and context-specific and vary widely depending on the setting. For example, some standards provide an objective for de-identification, while others prescribe a method for de-identification.

- Applicability is typically a binary determination that turns on the interpretation of terminology such as personal information, personally identifiable information, or individually identifiable information.

- Practices also vary, but generally are heuristic and focus on withholding, removing, or coarsening pieces of information considered to be identifying.

Variations in standards: Selected laws

- HIPAA Privacy Rule
- Family Educational Rights and Privacy Act
- Confidential Information Protection and Statistical Efficiency Act
- Massachusetts data security regulation

HIPAA Privacy Rule

Method #1 for de-identifying data: Expert determination

A person with appropriate knowledge of and experience with generally accepted statistical and scientific principles and methods for rendering information not individually identifiable:

(i) Applying such principles and methods, determines that the risk is very small that the information could be used, alone or in combination with other reasonably available information, by an anticipated recipient to identify an individual who is a subject of the information; and

(ii) Documents the methods and results of the analysis that justify such determination

HIPAA Privacy Rule

Method #2 for de-identifying data: Safe harbor

(i) Categories of information from a list of 18 identifiers (e.g., names, geographic units containing 20,000 or fewer people, dates (except year), telephone numbers, Social Security numbers, etc.) are removed, and

(ii) The covered entity does not have actual knowledge that the information could be used alone or in combination with other information to identify an individual who is a subject of the information.

(45 C.F.R. § 164.514)

Family Educational Rights and Privacy Act

Permits the release of de-identified information, without consent, “after the removal of all personally identifiable information provided that the educational agency or institution or other party has made a reasonable determination that a student’s identity is not personally identifiable, whether through single or multiple releases, and taking into account other reasonably available information.” *(20 C.F.R. § 99.31(b)(1))*

Personally identifiable information includes, but is not limited to, names, addresses, personal identifiers (e.g., SSNs, student numbers, biometric records), indirect identifiers (e.g., date of birth, place of birth, mother’s maiden name), other information that, alone or in combination, is linked or linkable to a specific student that would allow a reasonable person in the school community, who does not have personal knowledge of the relevant circumstances, to identify the student with reasonable certainty, or information requested by a person who the educational agency or institution reasonably believes knows the identity of the student [in the requested record]. *(20 C.F.R. § 99.3)*

Confidential Information Protection and Statistical Efficiency Act

CIPSEA protects data in identifiable form, meaning “any representation of information that permits the identity of the respondent to whom the information applies to be reasonably inferred by either direct or indirect means.”

Confidential Information Protection and Statistical Efficiency Act of 2002, Pub. L. No. 107-347, tit. V, § 502 (4) (2002).

Massachusetts data security regulation

Personal information is defined as the combination of

(1) *a Massachusetts resident's first name (or first initial) and last name, and*

(2) any one or more of the following:

(a) Social Security number,

(b) Driver's license number or state-issued identification card number, or

(c) Financial account number, or credit or debit card number.

This definition explicitly excludes publicly available information.

(201 Mass. Code Regs.§ 17.00)

Gaps in the current framework

- De-identification standards in the US often rely on concepts such as personally identifiable information that are not precisely defined.
- Guidance on selecting among and applying privacy measures is limited.
- Standards focus on releases of data in microdata (individual-level) formats.
- Standards and guidance encourage use of a narrow subset of the privacy measures available and hinder adoption of stronger techniques.
- Lack of clear guidance leads to inconsistent practices and uncertainty. As a result, similar privacy risks (or even identical data) are sometimes treated differently by different actors.

Limitations of de-identification more generally

- Advances in the scientific understanding of privacy have demonstrated that privacy approaches relying exclusively on de-identification fail to provide reasonable protection.

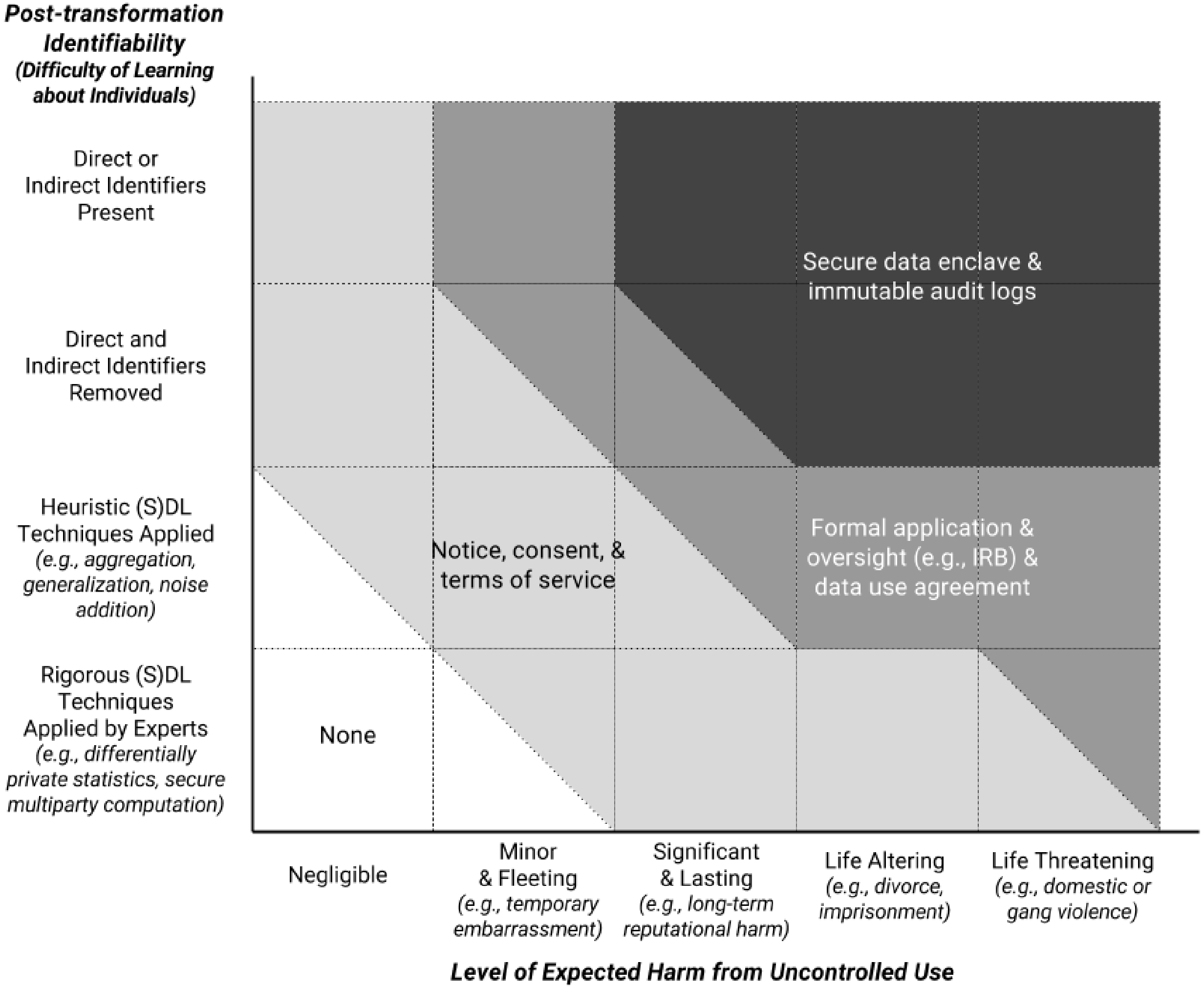
- While they may reduce some risks, traditional de-identification approaches
  - Do not prevent all disclosures or protect information in the manner that most individual subjects would expect,
  - Address only a subset of privacy attacks and attackers,
  - Are not readily scalable for use by non-experts, and
  - Often result in the redaction or withholding of useful information.

A modern approach

- In light of the limitations of de-identification, practitioners may consider:
  - Conducting a systematic analysis of informational risks and intended uses, and
  - Implementing a combination of privacy and security controls rather than relying solely on de-identification.
- For example, a **tiered access model** can
  - Closely match combinations of privacy controls to different risks and intended uses at each stage of the information lifecycle, and
  - Bring gains in both privacy and utility for a broad range of uses across different types of data.

A modern approach

Selecting combinations of appropriate privacy and security controls based on informational risks



Post-transformation Identifiability
(Difficulty of Learning about Individuals)

Direct or Indirect Identifiers Present

Direct and Indirect Identifiers Removed

Heuristic (S)DL Techniques Applied *(e.g., aggregation, generalization, noise addition)*

Rigorous (S)DL Techniques Applied by Experts *(e.g., differentially private statistics, secure multiparty computation)*

Secure data enclave & immutable audit logs

Formal application & oversight (e.g., IRB) & data use agreement

Notice, consent, & terms of service

None

Negligible

Minor & Fleeting *(e.g., temporary embarrassment)*

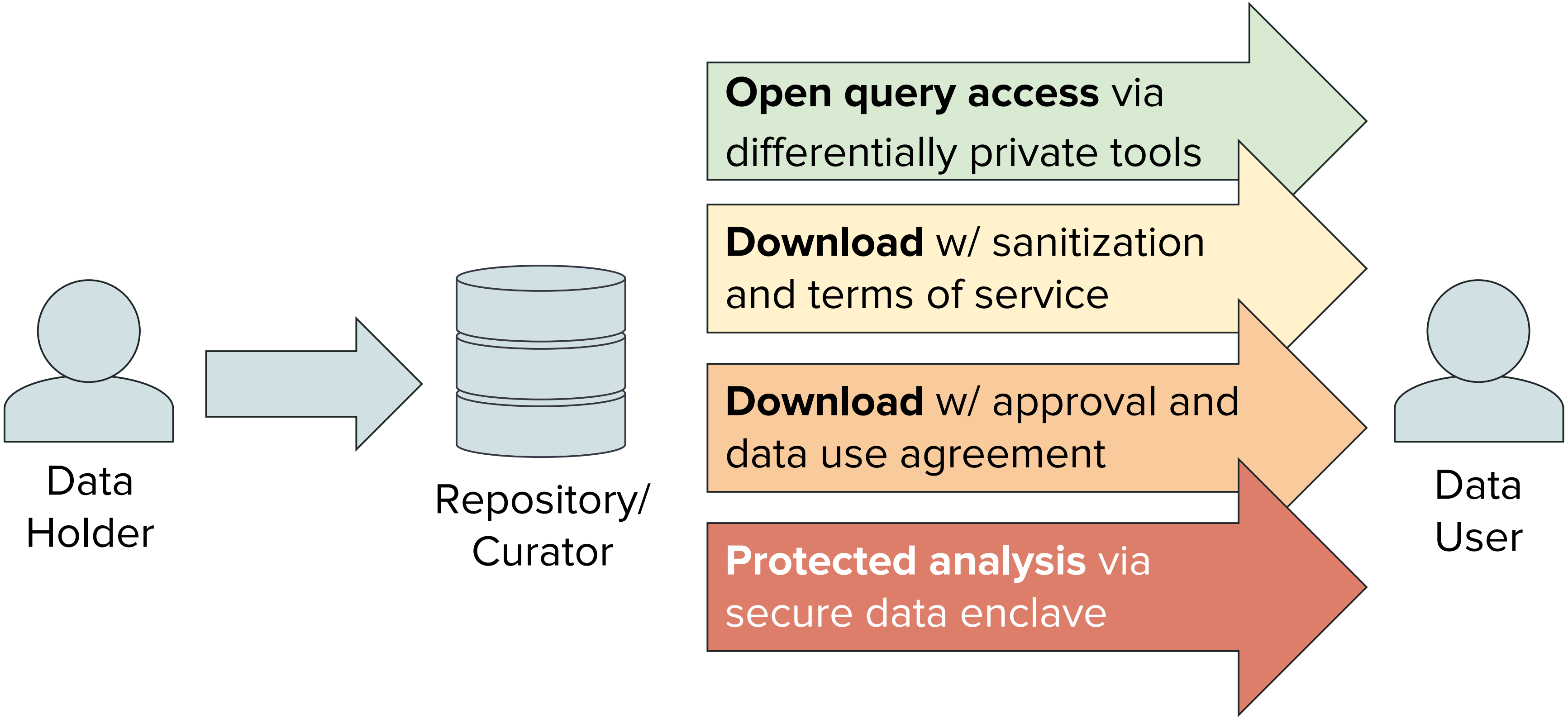
Significant & Lasting *(e.g., long-term reputational harm)*

Life Altering *(e.g., divorce, imprisonment)*

Life Threatening *(e.g., domestic or gang violence)*

Level of Expected Harm from Uncontrolled Use

A modern approach: Example tiered access model



Data Holder

Repository/ Curator

Open query access via differentially private tools

Download w/ sanitization and terms of service

Download w/ approval and data use agreement

Protected analysis via secure data enclave

Data User

References

Micah Altman, Alexandra Wood, David R. O’Brien, Salil Vadhan, and Urs Gasser, **Towards a Modern Approach to Privacy-Aware Government Data Releases**, 30 *Berkeley Technology Law Journal* 1967 (2015).

Alexandra Wood, Edo Airoldi, Micah Altman, Yves-Alexandre de Montjoye, Urs Gasser, David O’Brien, and Salil Vadhan, **Comments on the Proposed Rules to Revise the Federal Policy for the Protection of Human Subjects** (2016).

Available from http://privacytools.seas.harvard.edu

Thank you

ALFRED P. SLOAN FOUNDATION

BERKELEY CENTER FOR LAW & TECHNOLOGY

Microsoft

Commission on Evidence-Based Policymaking

National Science Foundation

Alfred P. Sloan Foundation

Berkeley Center for Law & Technology

Microsoft Corporation

and our collaborators through the Privacy Tools for Sharing Research Data project

IQSS
The Institute for Quantitative Social Science

CRCS Center for Research on Computation and Society
at Harvard John A. Paulson School of Engineering and Applied Sciences

BERKMAN KLEIN CENTER
FOR INTERNET & SOCIETY AT HARVARD UNIVERSITY

Program on Information Science
MIT Libraries

DATA PRIVACY LAB