Monday May 23, 2016
*Via electronic filing*

Rachel Seeger, J.D.
HHS Office of Science and Data Policy
Assistant Secretary for Planning and Evaluation
< rachel.seeger2@hhs.gov >

Re: *Hearing, Subcommittee on Privacy, Confidentiality & Security; National Committee on Vital and Health Statistics*

Dear Assistant Secretary Seeger,

This comment is informed by research with collaborators through the *Privacy Tools for Sharing Research Data* project at Harvard University.[1] In this broad, multidisciplinary project, we are exploring the privacy 1 issues that arise when collecting, analyzing, and disseminating research datasets containing personal information. Our efforts are focused on translating the theoretical promise of new measures for privacy protection and data utility into practical tools and approaches. In particular, our work aims to help realize the tremendous potential from social science research data by making it easier for researchers to share their data using privacy protective tools.

Academic research in cryptography theory, statistics and information science has demonstrated a number of challenges related to managing information privacy in the modern world. A recent research workshop illustrated three challenges, paraphrased below:[2]

The first challenge is that many human behaviors leave behind distinct behavioral fingerprints in the data -- even in the complete absence of traditional identifiers (or quasi-identifiers) This creates a problem for most traditional statistical disclosure limitation methods. A second challenge is that when data is released that is protected by traditional statistical disclosure control methods, such as identifier-based redaction or aggregation to prevent record-linkage, informational risks to individuals from that data release continue to grow in the future as new external data is released. This is because traditional methods whatever modification they make to the data (e.g. swapping. While these methods may be sufficient for controlling what can be learned about an individual from a specific data set – modern privacy research shows that such approaches cannot provide any strict bounds on the amount that can be learned from composing with independent auxiliary information. A third challenge revealed by modern privacy research is that every release of data, if it has any utility, no matter how it is protected, inevitably leaks some private information, and this leakage increases with each release. In other words – there is no free lunch with respect to information privacy, you always have to buy it with utility.

---

[1] The Privacy Tools for Sharing Research Data project is supported by a National Science Foundation Secure and Trustworthy Cyberspace Frontier grant and a grant from the Alfred P. Sloan Foundation. See Privacy Tools for Sharing Research Data, http://privacytools.seas.harvard.edu.

[2] Altman M, Capps C, Prevost R. Location Confidentiality and Official Surveys. Social Science Research Network [Internet]. 2016.

In previous publications an and regulatory comments my collaborators and I have offered a number of recommendation that we believe would enable the wider sharing of research data while providing strong privacy protection for individuals. [3]

Although most of the writings above do not comment directly on HIPAA regulations and controls. It is my judgement that the risks discussed apply to protected health information, and that the broad findings and recommendations are readily applicable to improving HIPAA. do Thus my comments summarize these prior recommendations. And I recommend that the committee read and incorporate these previous recommendations.

As a general framework, myself and collaborators have recommended the development of rules and guidance based on the following principles of a modern approach to privacy:[4]

- Calibrating privacy and security controls to the intended uses and privacy risks associated with
  the data;
- When conceptualizing informational risks, considering not just reidentification risks but also inference risks, or the potential for others to learn about individuals from the inclusion of their information in the data;
- Addressing informational risks using a combination of privacy and security controls rather than
  relying on a single control such as consent or deidentification;
- Anticipating, regulating, monitoring, and reviewing interactions with data across all stages of the lifecycle (including the post-access stages), as risks and methods will evolve over time; and
- In efforts to harmonize approaches across regulations and institutional policies, emphasizing the
  need to provide similar levels of protection to research activities that pose similar risks.

(We note in prior writings that terms above, such as privacy, confidentiality, security, and sensitivity are used in multiple communities of practice in somewhat different ways, and they are

[3]See: Altman M, Wood A, O'Brien D, Vadhan S, Gasser U. Towards a Modern Approach to Privacy-Aware Government Data Releases. Berkeley Journal of Technology Law 30(3) 1967-2072. 2016; Wood A, Airoldi E, Altman M, de Montandre Y, Gasser U, O'Brien D, Vadhan S. Privacy Tools project response to Common Rule Notice of Proposed Rule Making. Comments on Regulation.Gov . 2016. (Copy available here: http://informatics.mit.edu/publications/privacy-tools-project-response-common-rule-notice-proposed-rule-making); and
Vayena E, Gasser U, Wood A, O'Brien D, Altman M. Elements of a New Ethical and Regulatory Framework for Big Data Research. Washington and Lee Law Review. 2016;72(3):420-442.

[4] Altman M, Wood A, O'Brien D, Vadhan S, Gasser U. Towards a Modern Approach to Privacy-Aware Government Data Releases. Berkeley Journal of Technology Law 30(3) 1967-2072. 2016; Wood A, Airoldi E, Altman M, de Montandre Y, Gasser U, O'Brien D, Vadhan S. Privacy Tools project response to Common Rule Notice of Proposed Rule Making. Comments on Regulation.Gov . 2016.

defined inconsistently throughout the literature. We suggest a vocabulary for these terms in the works cited above. And I recommend that any regulation refer to explicit and definitions of these terms.)

And in related research we have argued for the need for comprehensive regulatory protection against information privacy harms from research. Protections for people whose information is used in research should be based on the risks and benefits to the subject – and not on ethically irrelevant elements of the research context such as the institution conducting the research, its commercial status, or sources of funding.[5]

The research cited above find that generally addressing privacy risks requires a sophisticated approach, and the privacy protections currently used in government releases of data do not take advantage of advances in data privacy research. We note that there are wide range of technical, procedural, legal, educational, and economic controls; but that most government data releases rely almost on redaction and binary access control. This focus on a small set of controls likely fails to address the nuances of data privacy risks.

The research above notes (as paraphrased) that advances in science and technology enable the increasingly sophisticated characterization of privacy risks and harms and new interventions for protecting data subjects. We describe a lifecycle approach that supports a systematic decomposition of the factors relevant to data management at each information stage, including the collection, transformation, retention, access/release, and post-access stages. And we propose a framework for developing guidance on selecting appropriate privacy and security measures that are calibrated to the context, intended uses, threats, harms, and vulnerabilities associated with a specific research activity.

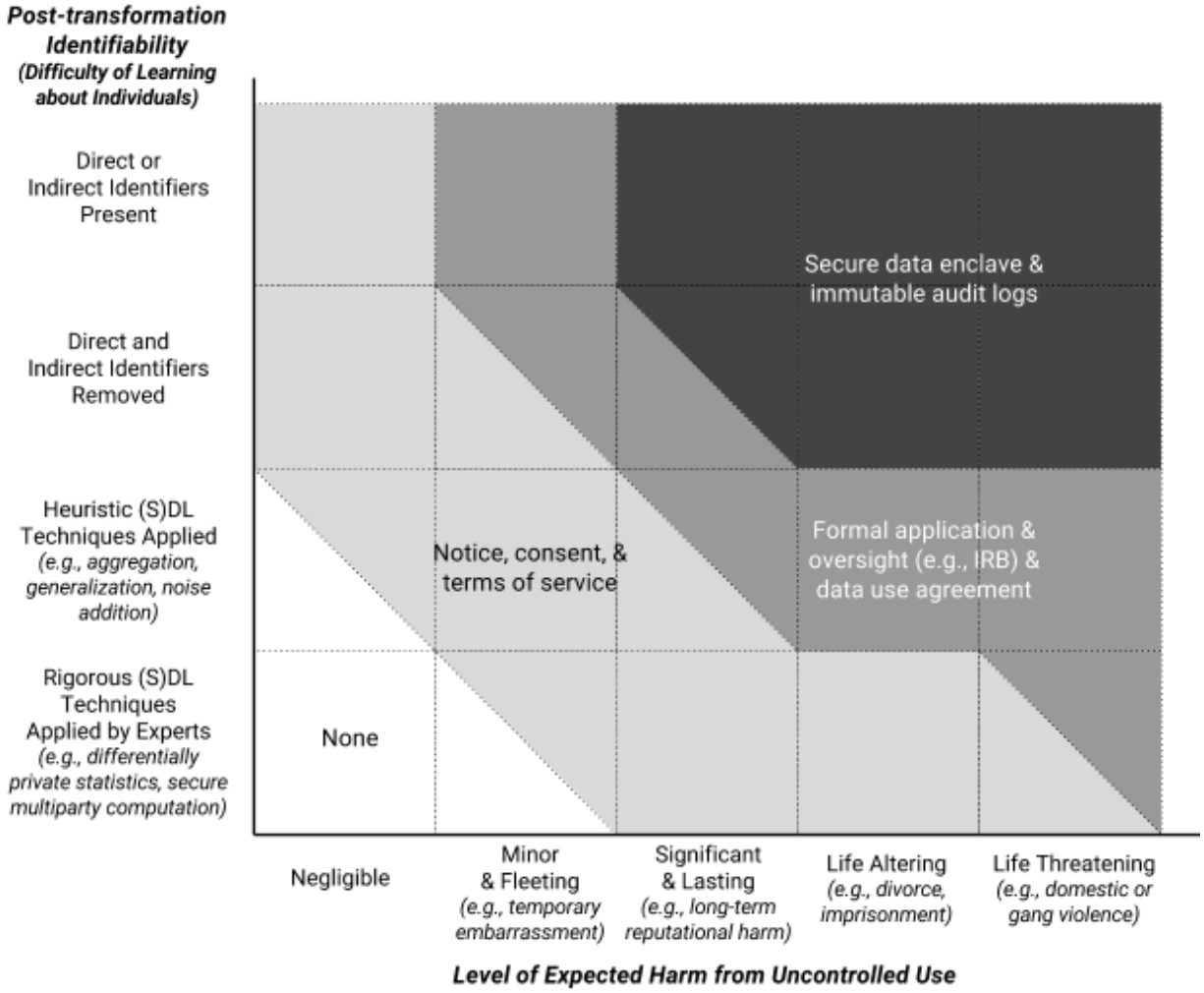
Figure 1 provides a partial conceptualization of this framework. In this diagram, the x-axis provides a scale for the level of expected harm from uncontrolled use of the data, meaning the maximum harm the release could cause to some individual in the data based on the sensitivity of the information. This scale ranges from low to high levels of expected harm, with harm defined to capture the magnitude and duration of the impact a misuse of the data would have on an affected individual's life, and we have placed examples as reference points along this axis. The y-axis provides a scale for the post-transformation identifiability, the potential for others to learn about individuals based on the inclusion of their information in the data, and a number of examples are provided as anchor points, ranging from data sets containing direct or indirect identifiers, to data shared using expertly applied rigorous disclosure limitation techniques backed by a formal mathematical proof of privacy.

The level of expected harm from uncontrolled use and the post-transformation identifiability of the data, taken together, point to minimum privacy and security controls that are appropriate in a given case, as shown by the shaded regions in the diagram. Regions divided by a diagonal line correspond to categories of information for which an actor could reach different conclusions based on the intended uses of the data or privacy standards that vary based on the applicability of

[5] Vayena E, Gasser U, Wood A, O'Brien D, Altman M. Elements of a New Ethical and Regulatory Framework for Big Data Research. Washington and Lee Law Review. 2016;72(3):420-442.

a regulation, contract, institutional policy, or best practice. The sets of controls within the shaded regions focus on a subset of controls from the more comprehensive set of procedural, economic, educational, legal, and technical controls we catalog below in Table 1 in Section. In practice, the design of a data management plan should draw from the wide range of available interventions and incorporate controls at each stage of the lifecycle, including the post-access stage. Also note there are regions of this diagram that deviate from current practice in some domains. For example, we argue that data that have been de-identified using simple redaction or other heuristic techniques should in many cases be protected using additional controls.

**Figure 1.** Calibrating privacy and security controls.




For many activities, implementing a single set of privacy and security controls may not be appropriate for all intended uses of the information. For this reason, we generally recommend that regulators and data controllers implement a tiered access model. A tiered access model is one in which data are made available to different categories of data users through different mechanisms.

Figure 1 illustrates the relationship between transformation and release controls, and suggests how controls could be selected for different access tiers. For example, an investigator could provide public access to some data without restriction after robust disclosure limitation

techniques have transformed the data into differentially private statistics. Data users who intend to perform analyses that require the full dataset, including direct and indirect identifiers, could be instructed to submit an application to an IRB, and their use of the data would be restricted by the terms of a data use agreement. We argue that this framework, implemented through a data management plan and tiered access model, would help IRBs and investigators calibrate the privacy and security controls to the contexts, threats, harms, and vulnerabilities associated with a research activity, as well as the purposes desired by different categories of data users.

Table 1 below provides an example catalog illustrating the wide range of procedural, economic, educational, legal, and technical controls that are available at each lifecycle stage that should be considered for inclusion in an appropriate set of controls.

**Table 1.** Example catalog of privacy and security controls.

| | **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 | Consent education; Transparency; Notice; Nutrition labels; Public education; Privacy icons | Data minimization; Notice and consent; Purpose specification | Computable policy |
| **Transformation** | Process for correction | | Metadata; Transparency | Right to correct or amend; Safe harbor de-identification standards | Aggregate statistics; Computable policy; Contingency tables; Data visualizations; Differentially private data |

| | | | | | |
|---|---|---|---|---|---|
| | | | | | summaries; Redaction; SDL techniques; Synthetic data |
| **Retention** | Audits; Controlled backups; Purpose specification; Security assessments; Tethering | | Data asset registers; Notice; Transparency | Breach reporting requirements; Data retention and destruction requirements; Integrity and accuracy requirements | Computable policy; Encryption; Key management (and Secret sharing); Federated databases; Personal data stores |
| **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 |
| **Post-Access (Audit, Review)** | Audit procedures; Ethical codes; Tethering | Fines | Privacy dashboard; Transparency | Civil and criminal penalties; Data use agreements/ Terms of service; | Computable policy; Immutable audit logs; Personal data stores |

| | | | | | |
|---|---|---|---|---|---|
| | | | | Private right of action | |

In the prior work noted above[6] we also call special attention to advanced data-sharing models and emerging formal approaches to privacy. We note that there are a number of privacy methods and data-sharing models that can provide stronger privacy protection than traditional de-identification techniques that are in wide use today – these include synthetic data, interactive query servers, and multiparty computation systems. We further note: "Many of these data-sharing models are also compatible with a formal privacy guarantee called differential privacy. Differential privacy is a strong, quantitative notion of privacy that is provably resilient to a very large class of potential misuses. As a robust privacy framework that addresses both known and unforeseeable attacks, differential privacy represents a solution that moves beyond the penetrate-and-patch approach that is characteristic of traditional de-identification approaches. We recommend that government regulations, through the proposed list of approved safeguards, encourage the use of stronger privacy measures, including measures that are compatible with formal privacy models."

Thank you for your consideration of these comments.

Respectfully,

Micah Altman
Director of Research, MIT Libraries
Nonresident
Senior Fellow, Brookings Institution

[6] Wood A, Airoldi E, Altman M, de Montandre Y, Gasser U, O'Brien D, Vadhan S. Privacy Tools project response to Common Rule Notice of Proposed Rule Making. Comments on Regulation.Gov . 2016

**Prepared for**

*Hearing, Subcommittee on Privacy, Confidentiality & Security; National Committee on Vital and Health Statistics*

***Washington DC***
**May 2016**

# Comments on Regulating Information Privacy -- A Modern Approach

Micah Altman
Director of Research
MIT Libraries

CC BY SA

# DISCLAIMER

These opinions are my own, they are not the opinions of MIT, Brookings, any of the project funders, nor (with the exception of co-authored previously published work) my collaborators

*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.

# Collaborators & Co-Conspirators

- Privacy Tools for Sharing Research Data Team (Salil Vadhan, P.I.)
  http://privacytools.seas.harvard.edu/people
- Research Support
  - Supported in part by NSF grant CNS-123723
  - Supported in part by the Sloan Foundation

# Related Work

Main Project:

- Privacy Tools for Sharing Research Data
  http://privacytools.seas.harvard.edu/

Related publications:


- Altman M, Wood A, O'Brien D, Vadhan S, Gasser U. Towards a Modern Approach to Privacy-Aware Government Data Releases. Berkeley Journal of Technology Law 30(3) 1967-2072. 2016
- Wood A, Airoldi E, Altman M, de Montandre Y, Gasser U, O'Brien D, Vadhan S. Privacy Tools project response to Common Rule Notice of Proposed Rule Making. Comments on Regulation.Gov . 2016. (Copy available here: http://informatics.mit.edu/publications/privacy-tools-project-response-common-rule-notice-proposed-rule-making); and
- Vayena E, Gasser U, Wood A, O'Brien D, Altman M. Elements of a New Ethical and Regulatory Framework for Big Data Research. Washington and Lee Law Review. 2016;72(3):420-442.
- Altman M, Capps C, Prevost R. Location Confidentiality and Official Surveys. Social Science Research Network [Internet]. 2016.


Slides and reprints available from:
informatics.mit.edu

# Privacy Core Concepts

**Privacy**
Control over extent and circumstances of sharing


**Confidentiality**
Control over disclosure of information


**Identifiability**
Potential for learning about individuals based on their inclusion in a data


**Sensitivity**
Potential for harm
if information disclosed and used to learn about indviduals

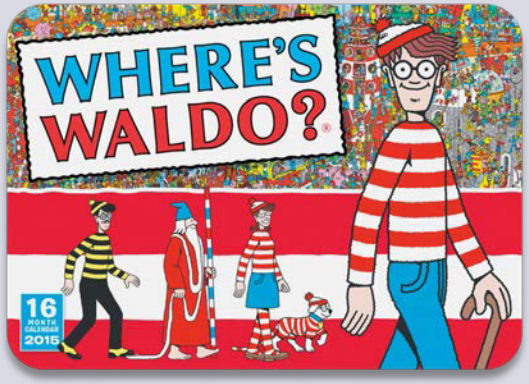
# Anonymization

- Anonymization / deidentification are legal concepts – typically without general rigorous formal definition
- Definition varies by law, may include …
  - Presence of specific attributes (e.g., PII, HIPAA identifiers)
  - Feasibility of record linkage ...
  - Evaluation of knowledge of data publisher (e.g. "no actual knowledge", "readily ascertainable")

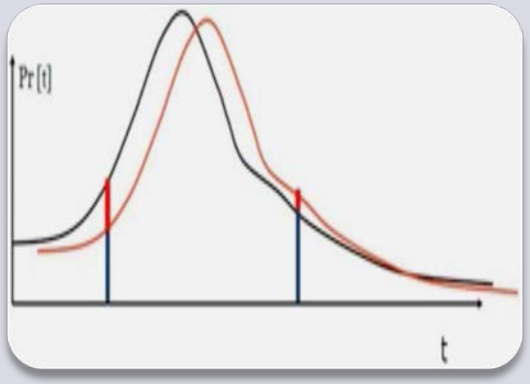
# Challenges to Inferential Confidentiality

- Many human behaviors leave behind distinct behavioral fingerprints
- When data is released that is protected by traditional statistical disclosure risks to individuals from that data release continue to grow in the future as new external data is released.
- Every release of data, if it has any utility, no matter how it is protected, inevitably leaks some private information.

# Different types of identifiability








## Record-linkage

*"where's waldo"*

- Match a real person to precise record in a database
- *Examples*: direct identifiers.
- *Caveats*: Satisfies compliance for specific laws, but not generally; substantial potential for harm remains

## Indistinguishability

*"hiding in the crowd"*

- Individuals can be linked only to a cluster of records (of known size)
- *Examples*: K-anonymity, attribute disclosure
- *Caveats*: Potential for substantial harms may remain, must specify what external information is observable, & need diversity for sensitive attributes

## Limited Adversarial Learning

*"confidentiality guaranteed"*

- Formally bounds the total learning about any individual that occurs from a data release
- *Examples*: differential privacy, zero-knowledge proofs
- *Caveats*: Challenging to implement, often requires interactive systems

Less Protection → More Protection

# Unpacking “sensitivity”

- **Threats** are defined broadly as potential adverse circumstances or events that could cause harm to a data subject as a result of inclusion of the subject’s data
- **Harms** are defined as injuries sustained by data subjects as a result of a threat being realized
- **Vulnerabilities** are defined as characteristics that increase the likelihood that threats will be realized

***“Sensitivity” measures should summarize the expected harm that would occur if specified private information was learned about an individual***

# Who might be harmed by information release?




# What use is it?

- **Utility** is defined broadly as the analytical value of the data
- No free lunch – no method can provide optimal maximum privacy and utility simultaneously…
- However, new methods can sometimes do better than traditional anonymization on both fronts.

# Information Security Core Properties

Non-Repudiation

- all actions and changes provably sourced to a unique person

Authenticity

- authorized users can validate information source


Availability

- authorized users can access as needed

**Confidentiality (Secrecy)**

- control over disclosure

Integrity

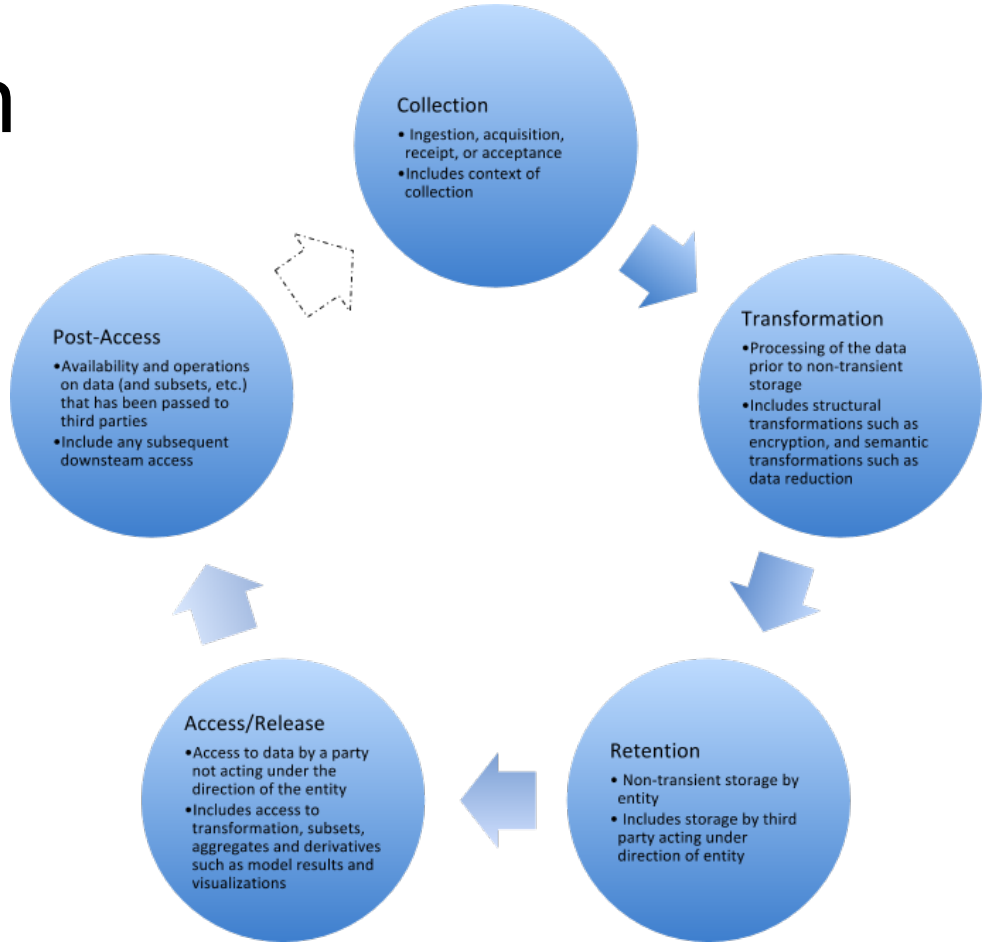
- control over modification

# Principles of a Modern Approach to Information Privacy & Confidentiality

- Calibrating privacy and security controls to the intended uses and privacy risks associated with the data;
- When conceptualizing informational risks, considering not just reidentification risks but also inference risks, or the potential for others to learn about individuals from the inclusion of their information in the data;
- Addressing informational risks using a combination of privacy and security controls rather than relying on a single control such as consent or deidentification;
- Anticipating, regulating, monitoring, and reviewing interactions with data across all stages of the lifecycle (including the post-access stages), as risks and methods will evolve over time; and
- In efforts to harmonize approaches across regulations and institutional policies, emphasizing theneed to provide similar levels of protection to research activities that pose similar risks.

# Lifecycle approach to data management

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



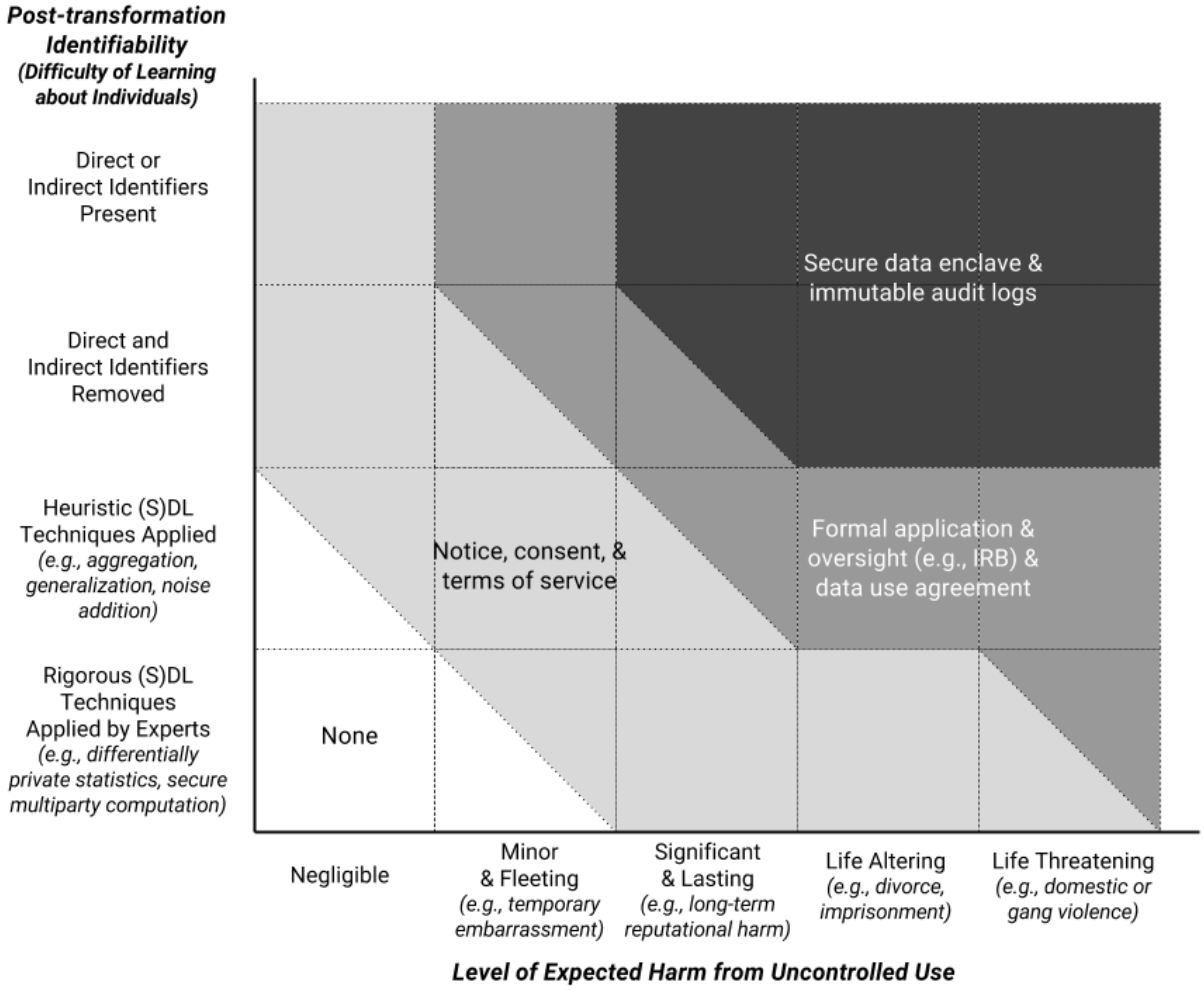
# Catalog of privacy controls

- Procedural, technical, educational, economic, and legal means for enhancing privacy—at each stage of the information lifecycle

| | 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 |

# Calibrating Controls

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



# References


- Altman M, Wood A, O'Brien D, Vadhan S, Gasser U. Towards a Modern Approach to Privacy-Aware Government Data Releases. Berkeley Journal of Technology Law 30(3) 1967-2072. 2016
- Wood A, Airoldi E, Altman M, de Montandre Y, Gasser U, O'Brien D, Vadhan S. Privacy Tools project response to Common Rule Notice of Proposed Rule Making. Comments on Regulation.Gov . 2016. (Copy available here: http://informatics.mit.edu/publications/privacy-tools-project-response-common-rule-notice-proposed-rule-making); and
- Vayena E, Gasser U, Wood A, O'Brien D, Altman M. Elements of a New Ethical and Regulatory Framework for Big Data Research. Washington and Lee Law Review. 2016;72(3):420-442.
- Altman M, Capps C, Prevost R. Location Confidentiality and Official Surveys. Social Science Research Network [Internet]. 2016.