Hybrid Legal-Technical Concepts of Privacy

Alexandra Wood
Berkman Klein Center for Internet & Society at Harvard University

Data Privacy: From Foundations to Applications (March 4, 2019)

Simons Institute for the Theory of Computing, UC Berkeley

An Interdisciplinary Collaboration

Computer Science

Kobbi Nissim, Aaron Bembenek, Mark Bun, Aloni Cohen, Marco Gaboardi, Thomas Steinke, Salil Vadhan

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

Georgetown University

Law & Policy

Urs Gasser, David O'Brien, Alexandra Wood, Aaron Fluitt

BERKMAN KLEIN CENTER
FOR INTERNET & SOCIETY
AT HARVARD UNIVERSITY

Georgetown University

Social Science & Information Science

Micah Altman

MITLibraries

Motivation:

Inadequacy of Current Privacy Frameworks

Inadequacy of Current Regulatory Framework

The current framework is ill-suited to protect privacy in the digital age.

- **Regulatory requirements are ambiguous and open to interpretation.**
    - It is difficult (if not impossible) to understand what they are designed to protect.
        - Particularly as technological developments challenge traditional understandings.
    - Practical approaches may require time-consuming, case-by-case review and negotiation.
    - This often results in markedly different treatment.

Inadequacy of Current Regulatory Framework

The current framework is ill-suited to protect privacy in the digital age.

- **Underlying concepts (e.g., PII) seem fundamentally unable to keep up with the pace of technological change.**
    - Endorse ad hoc approaches that fail to provide adequate protection (e.g., HIPAA Privacy Rule safe harbor method)
    - Recognize only a limited set of privacy failures (e.g., record linkage)
    - Require frequent amendment to regulatory standards

Weaknesses of Common Concepts

Privacy concepts in common use (e.g., personally identifiable information, de-identification, anonymization) are ill-defined.

- **Lack of generality**
    - Legal standards typically do not establish a general privacy goal to be achieved, but rather specify a technique to be used (e.g., redaction of identifiers)
    - Determinations often must be based on a case-by-case, contextual analysis

Weaknesses of Common Concepts

Privacy concepts in common use (e.g., personally identifiable information, de-identification, anonymization) are ill-defined.

- **Lack of precision**
  - Definitions of de-identification and anonymization are often tautological, or set forth a standard that is impossible to meet
  - Satisfaction of a legal standard cannot be demonstrated with certainty

Privacy’s Hybrid Nature

The Inherent Hybrid Nature of Data Privacy

- **Privacy regulations often rely on concepts that have both technical and legal meaning.**
    - E.g., identifiability, distinguishability, linkability, inference, and risk
    - These concepts play a central role in regulatory standards and criteria
        - They also have technical interpretations where they appear in the statistical disclosure limitation literature.
- **Yet legal and technical understandings of privacy are often not in harmony.**
    - This is particularly true when comparing new formal privacy models like differential privacy to legal approaches to protecting privacy.

Gaps between Technical & Normative Concepts

1. Generality of protection afforded

- Regulatory requirements vary according to industry sector, jurisdiction, institution, types of information, and other contextual factors
    - e.g., FERPA protects only certain information from education records maintained by educational agencies and institutions.
- Challenges: In practice, privacy risks are not limited to the information categories and contexts contemplated by regulations.
    - Further, analysts may combine information from different contexts.
- In contrast, formal privacy models like differential privacy can be applied wherever statistical or machine learning analysis is performed, regardless of context, and protect all information specific to an individual.

Gaps between Technical & Normative Concepts

2. Scope of attacks contemplated

- Regulations often contemplate a limited set of specific attacks and failures.
    - e.g., **record linkage** (the re-identification of one or more records in a de-identified dataset by uniquely linking those records with identified records in a publicly available dataset) is often the primary or sole failure mode.
- Challenges: Other privacy attacks are identified over time.
    - e.g., confirming an individual's presence in a dataset, singling out an individual (even if not fully identified), inferring information specific to an individual with less than absolute certainty.
- Formal privacy models provide protection against a wide collection of privacy attacks, even those that are not currently known.

Gaps between Technical & Normative Concepts

3. Expectations vs. the scientific understanding

- Regulations that rely on the concept of de-identification or anonymization are often not in agreement with the current scientific understanding privacy.
    - They may be limited in scope, may not provide an adequate level of privacy in practice, and may not withstand rigorous, formal mathematical scrutiny.
    - e.g., HIPAA Privacy Rule safe harbor method
        - Redaction of identifiers can fail to protect privacy, especially when applied to detailed information such as medical records.
    - Any information, even information not traditionally considered identifying, has the potential to leak information specific to individuals.

Gaps between Technical & Normative Concepts

3. Expectations vs. the scientific understanding

- Statutes may be interpreted to require something that is not technically feasible (i.e., absolute privacy protection when sharing personal data).
    - e.g., Title 13, U.S. Code protects the confidentiality of respondent information protected by the US Census Bureau, prohibiting any publication whereby the data furnished by an individual "can be identified."
    - If this concept were interpreted very conservatively, Title 13 would disallow any leakage of information about individuals.

Gaps between Technical & Normative Concepts

3. Expectations vs. the scientific understanding

- Binary view of privacy found in Title 13—whereby information is either identifiable or not—is common to many regulations.
- Issues with this approach:
  - (1) Information can never be made completely non-identifiable and
  - (2) Fails to recognize that privacy loss accumulates with successive releases of information about the same individuals (and can
eventually amount to a significant disclosure of personal information).
- In contrast, formal privacy models bound the privacy leakage of each release, and bound the total privacy leakage across multiple releases.

Gaps between Technical & Normative Concepts

4. (In)-stability over time

- Notions of privacy embedded in regulations are continually evolving.
    - e.g., OMB guidance updated over time to address new ways de-identified data may be vulnerable to potential attacks.
        - New 2017 guidance advises that non-PII may become PII in the future.
    - As hard-wired techniques (e.g., HIPAA safe harbor) are shown to be inadequate to protect privacy, it is unclear how to satisfy regulatory standards that are out of step with best practice.
- In contrast, differential privacy is the subject of ongoing scientific research and, regardless of implementation, there is a strong assurance that it provides a sufficient level of privacy in a wide variety of settings.

Gaps between Technical & Normative Concepts

5. Relationship to normative expectations

- Comparing the protection afforded by a particular privacy technology to a normative standard must be done with care.
- E.g., it may be difficult to make a sufficiency claim with respect to differential privacy and regulatory requirements.
    - Formal privacy models are “privacy-first” definitions, and real-world uses of data may demand a compromise between privacy and accuracy.
    - Regulations express requirements that can be interpreted to exceed the protection provided by formal privacy models.
        - E.g., Title 13 (especially with respect to protecting establishments)

Approach #1:

Integrating Modern Privacy Approaches Across the Information Lifecycle

Framework for Privacy-Aware Data Releases

Modeled on information security and lifecycle frameworks:

1. Developing a catalog of privacy controls
2. Identifying information uses, threats, and vulnerabilities
3. Designing data releases by aligning uses and risks with controls—at each stage of the information lifecycle

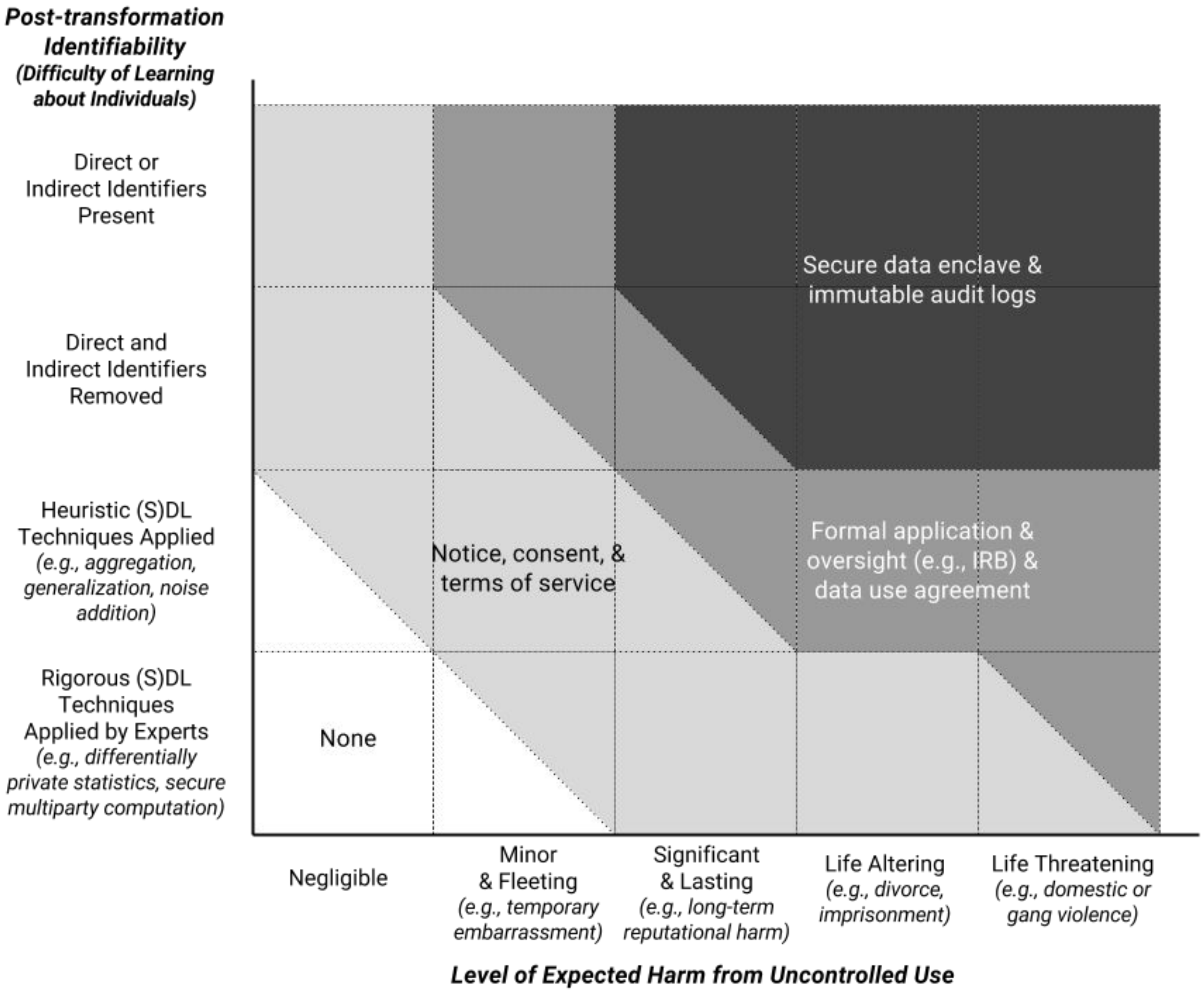
Catalog of Privacy Controls

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

Guide to Selecting Appropriate Privacy Controls



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)*

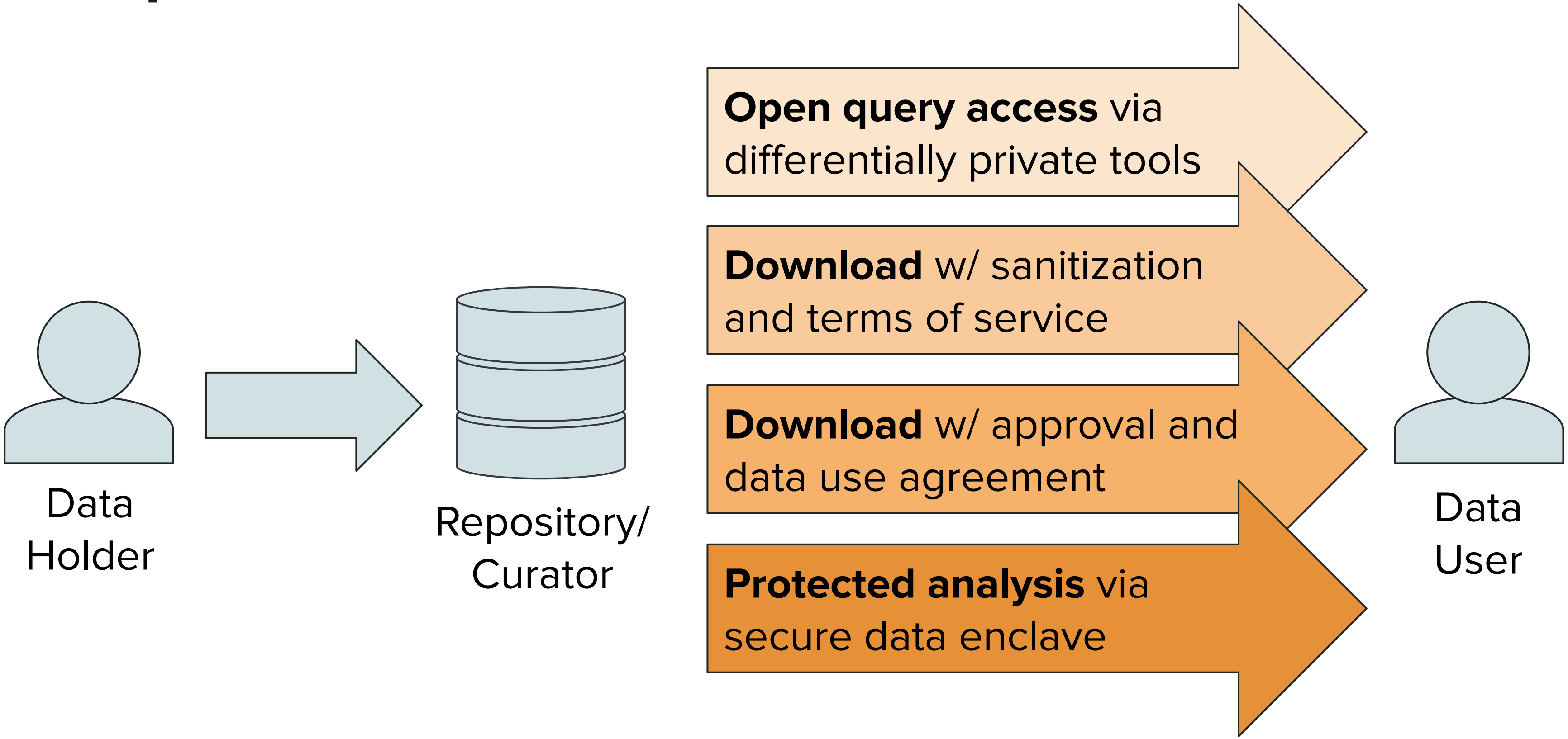
Significant & Lasting *(e.g., long-term reputational harm)*

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

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

Level of Expected Harm from Uncontrolled Use

Example Tiered Access Model



Data Holder

Repository/ Curator

Open query access via differentially private tools

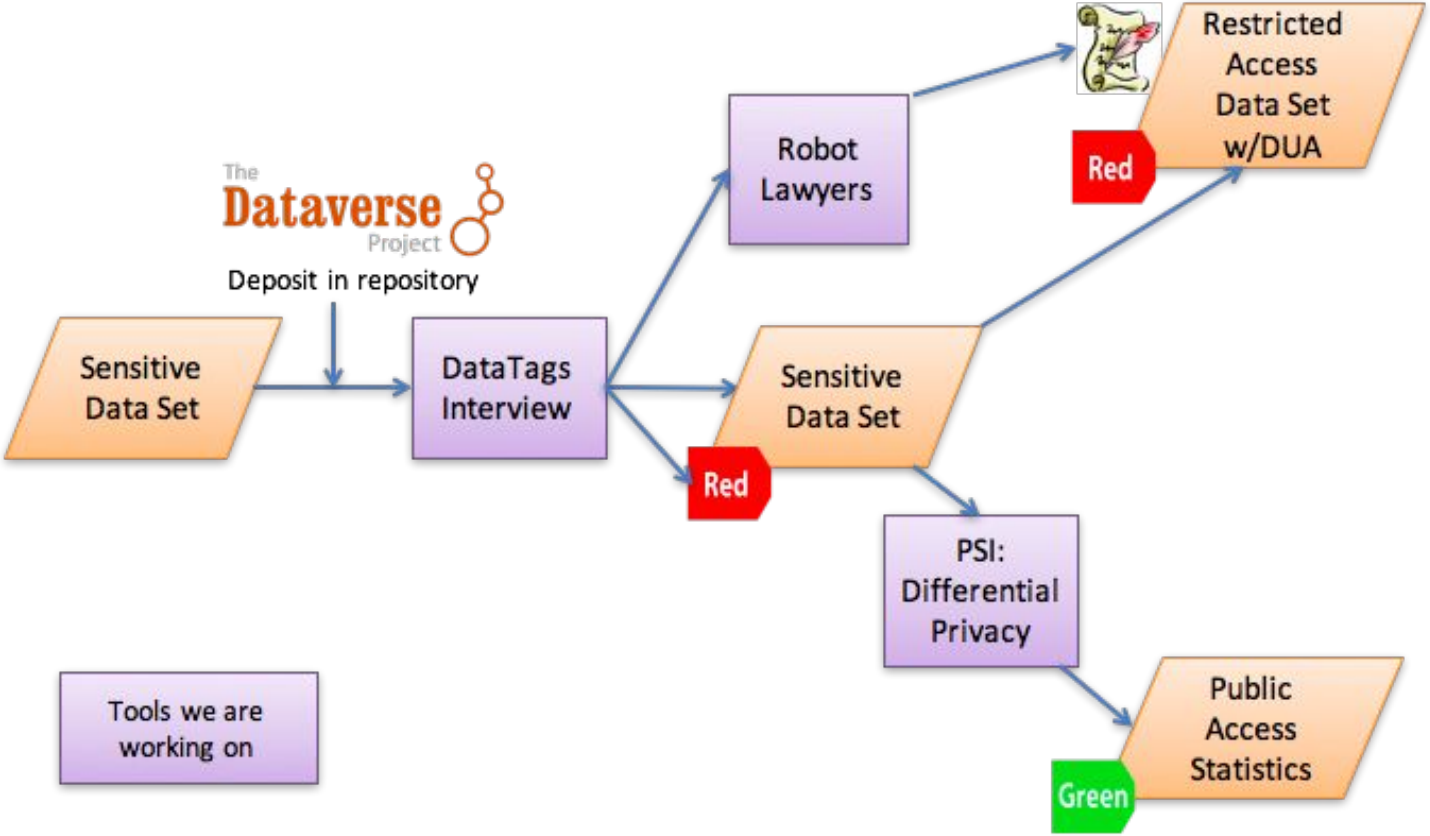
Download w/ sanitization and terms of service

Download w/ approval and data use agreement

Protected analysis via secure data enclave

Data User

Inspired by Privacy Tools Project Model



Approach #2:

Formally Modeling Legal Requirements

Is it possible to bridge these very different languages?

$M\colon X^n \to T$ satisfies ϵ-differential privacy if

$\forall x, x' \in X^n$ s.t. $dist_H(x, x') = 1\ \forall S \subseteq T$,

$$\Pr_{\mathrm{M}}[M(x) \in S] \le e^{\epsilon} \Pr_{\mathrm{M}}[M(x') \in S].$$

Approach to Formal Modeling

Goal: Rigorously arguing that a technological privacy solution satisfies the requirements of a particular law.

Proposed approach:

1. Extracting a formal mathematical requirement from the law.
2. Proving mathematically that a technological privacy solution satisfies the requirement derived from the law.

Illustration: Modeling FERPA

We extracted a formal model of the privacy desiderata for the Family Educational Rights & Privacy Act (FERPA).

We used a **game-based privacy definition**:

- This provides a concise and fairly intuitive abstraction of FERPA's requirements.
- If a formal model, such as differential privacy, satisfies the definition, then we have a strong argument that it satisfies the requirements of FERPA.

Modeling FERPA: The Adversary

- **Personally identifiable information**: “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.”

- This is FERPA’s **implicit adversary**.

Modeling FERPA: Directory Information

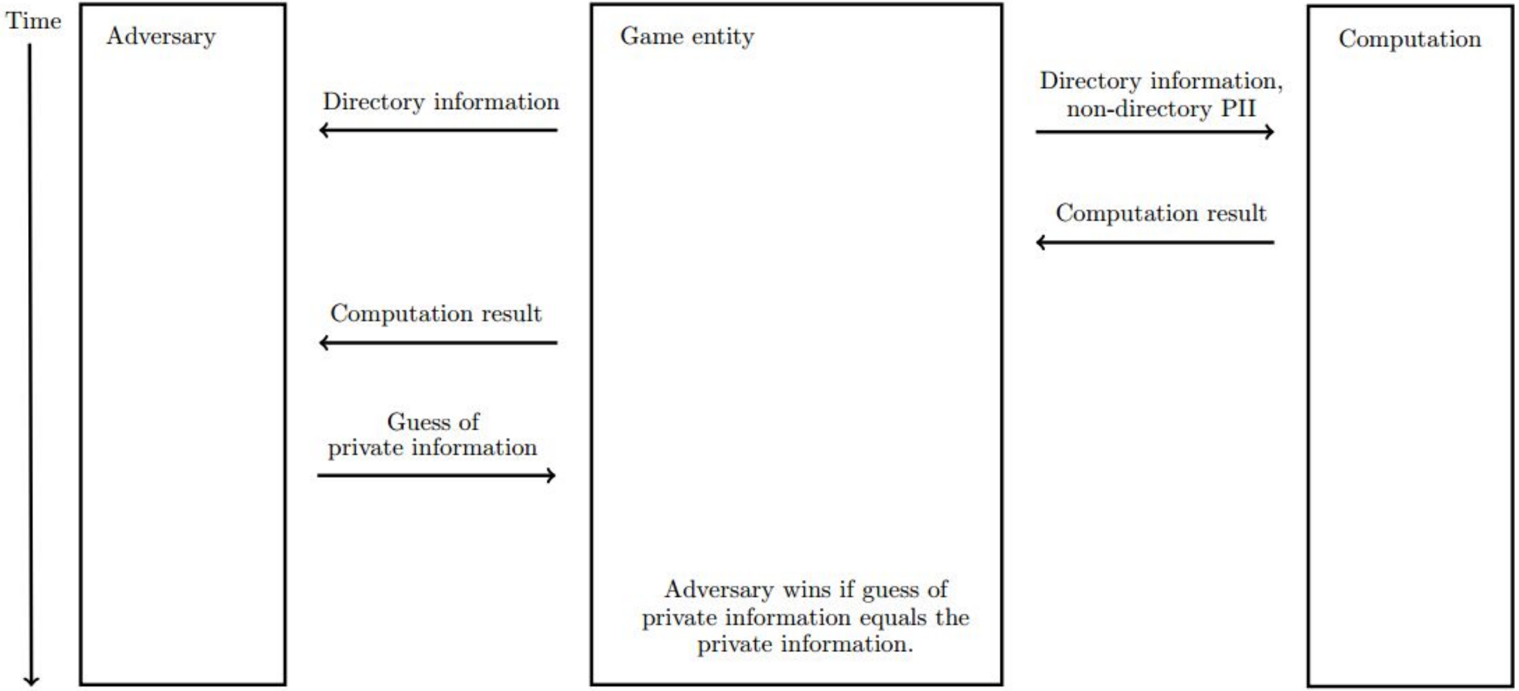
- Regulatory language interpreted conservatively, i.e., erring towards what is most beneficial for the adversary.

- E.g., the definition of **directory information** is ambiguous.

  - If we made assumptions about the definition, new interpretations could call these assumptions into question.

  - Instead, we let the attacker choose what constitutes directory information.

Components of a FERPA Privacy Game



Approach #3:

Interpreting Formal Privacy Guarantees

Common Privacy Concepts in the Law

- Personally identifiable information
- De-identification
- Linkage
- Inference
- Risk
- Consent and opting out
- Purpose and access restrictions

These concepts are interpreted differently across laws. They also appear in the technical literature, often with different definitions and interpretations.

Interpreting the Differential Privacy Guarantee

- Legal requirements relevant to issues of privacy in computation rely on an understanding of a range of different privacy concepts.

- None of the privacy concepts that appear in the law refer directly to differential privacy.
    - However, the differential privacy guarantee can be interpreted in reference to these concepts—while accommodating differences in how these concepts are defined across contexts.

Personally Identifiable Information

Personally identifiable information is a central concept appearing in privacy law.

- Legal protections typically extend only to PII, and information not considered personally identifiable is not protected (e.g., FERPA, HIPAA Privacy Rule).
- Although definitions vary significantly, they are generally understood to refer to the presence of pieces of information that are linkable to the identity of an individual or to an individual's personal attributes.
- PII is also related to the concept of *de-identification*, which refers to a collection of techniques, that if performed successfully, used as to remove PII, or transform PII into non-PII.

PII: Interpretation of DP Guarantee

PII does not have a precise technical meaning.

In practice it can be difficult to determine whether information is personal, identifying, or likely to be considered identifying in the future.

Further, the meaning of PII in releases that are not in a microdata or tabular format, such as statistical models or outputs of a machine learning system, is unclear.

However, when differential privacy is used, it can be understood as ensuring that using an individual's data will not reveal essentially any personally identifiable information specific to her, regardless of the definition of PII that is used.

➢ Here, *specific* is used to refer to information that is unique to the individual and cannot be inferred unless the individual's information is used in the analysis.

Approach #4:

Introducing Hybrid Concepts

Introducing Hybrid Concepts

- **Privacy concepts are neither purely legal nor purely technical.**
  - They have an inherent “hybrid” legal-technical nature.
- Adopting an understanding of privacy that is consistent across its technical and normative dimensions will be critical to ensuring personal data are adequately safeguarded over the long term.
- Conceptual gaps between existing technical and normative concepts create challenges for arriving at a universal notion of privacy.
- Understanding the hybrid nature of these concepts and developing tools for implementing them is necessary to bridge the gaps between the current legal and technical understanding of privacy.

Hybrid Concepts

- **Example: Contextual integrity**
  - Aims to capture the dual technical-normative nature of privacy.
  - Where it is possible to encode norms formally and determine whether a particular information flow respects these norms (i.e., when the norms are unambiguous and effectively testable), the framework can be used.
  - Normative concepts are often not defined explicitly, and, when they are, they are not expressed in a formal language that enables a precise analysis.
- More research is needed to develop hybrid concepts.

Hybrid Concepts

- **Approach to analyzing legal and technical concepts:**

  1. Identify fundamental concepts used in legal standards of privacy and the statistical disclosure literature (e.g., singling out, linkability, inference).
  2. For each concept, perform a thorough legal analysis.
  3. Construct a mathematical model of the concept.
  4. Check whether the modeling agrees with the legal analysis.
  5. Compare the model with differential privacy.

For an in-depth example, see Aloni Cohen's presentation on Wednesday.

Opportunities for Future Privacy Regulations

- **Regulations should articulate clear goals for privacy protection.**
    - These goals should be line with the scientific understanding of privacy.
    - Regulations should move away from implicitly or explicitly endorsing ad hoc de-identification techniques.
- **Example: Guidance on European data protection law outlines goals that go beyond the traditional notion of de-identification.**
    - Protection from singling out, linking, or inferring an individual's personal data from a dataset.
    - These concepts have not yet been defined precisely and formally from a mathematical perspective, but they aim to describe a clearer goal.

References

Kobbi Nissim, Alexandra Wood, Micah Altman, and Aloni Cohen, "**Hybrid Legal-Technical Concepts of Privacy**," Working Paper prepared for the 11th Annual Privacy Law Scholars Conference (2018).

Kobbi Nissim and Alexandra Wood, "**Is Privacy *Privacy*?**," 376 *Philosophical Transactions of the Royal Society A* 20170358 (2018).

Alexandra Wood, Micah Altman, Aaron Bembenek, Mark Bun, Marco Gaboardi, James Honaker, Kobbi Nissim, David O'Brien, Thomas Steinke, and Salil Vadhan, "**Differential Privacy: A Primer for a Non-technical Audience**," 21 *Vanderbilt Journal of Entertainment and Technology Law* 209 (2018).

Kobbi Nissim, Aaron Bembenek, Alexandra Wood, Mark Bun, Marco Gaboardi, Urs Gasser, David R. O'Brien, and Salil Vadhan, "**Bridging the Gap between Computer Science and Legal Approaches to Privacy**," 31 *Harvard Journal of Law & Technology* 687 (2018).

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