CS208: Applied Privacy for Data Science
Privacy Law & DP

James Honaker, Priyanka Nanayakkara, Salil Vadhan

School of Engineering & Applied Sciences
Harvard University

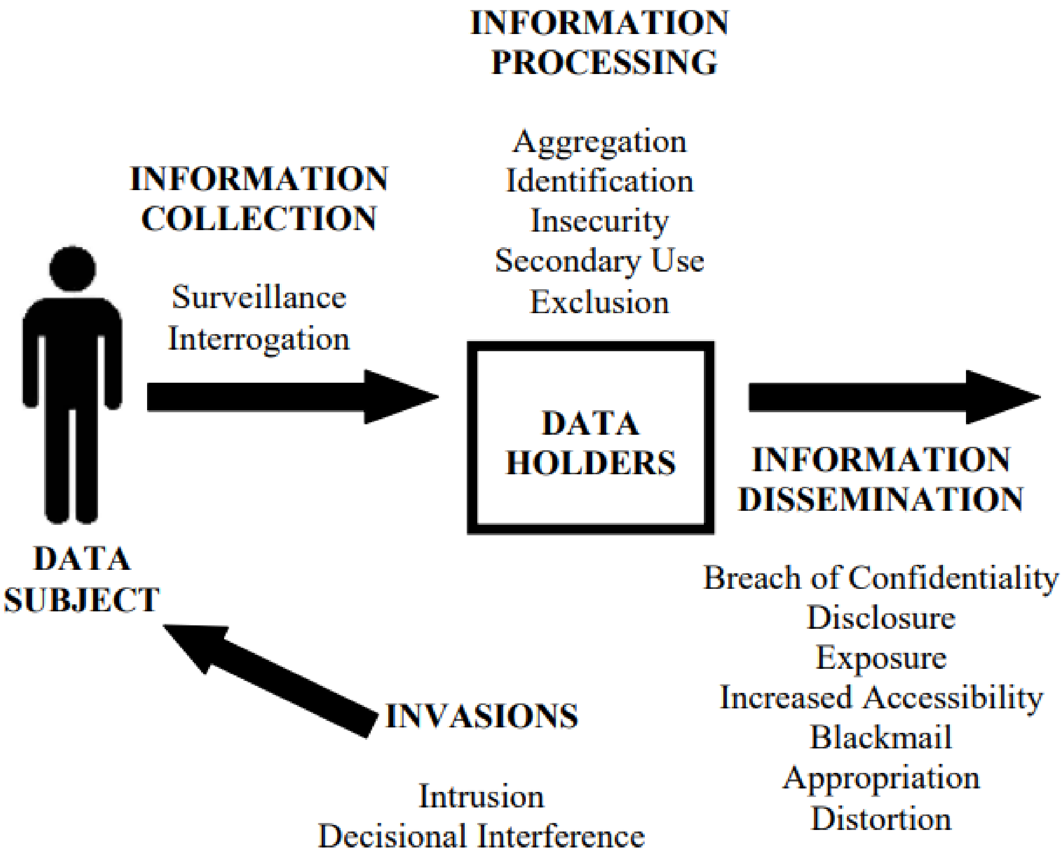
April 21, 2025

Announcements

Remaining class schedule

- Wed 4/23: Science, Technology, & Society (STS) & DP
- Mon 4/28: Industry & government panel
  - Jack Fitzsimons (Oblivious)
  - Badih Ghazi (Google)
  - Simson Garfinkel (Basis, formerly Census & NIST)
  - Wanrong Zhang (Tiktok)
- Wed 4/30: Conclusions, draft papers due
- Thu 5/8: Poster session, revised papers due

A Taxonomy of Privacy [Solove]



INFORMATION PROCESSING

Aggregation
Identification
Insecurity
Secondary Use
Exclusion

INFORMATION COLLECTION

Surveillance
Interrogation

DATA HOLDERS

INFORMATION DISSEMINATION

Breach of Confidentiality
Disclosure
Exposure
Increased Accessibility
Blackmail
Appropriation
Distortion

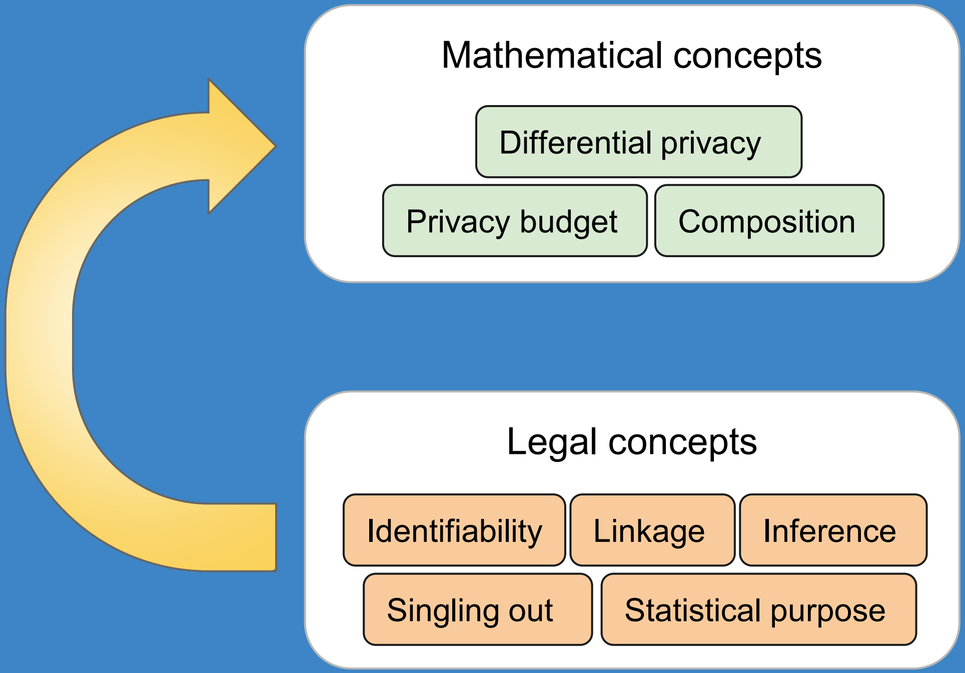
DATA SUBJECT

INVASIONS

Intrusion
Decisional Interference

Group Exercise 1: for which of these harms is DP relevant?

Formally reasoning about legal requirements



Mathematical concepts

Differential privacy

Privacy budget

Composition

Legal concepts

Identifiability

Linkage

Inference

Singling out

Statistical purpose

Remaining slides adapted from
Alexandra Wood [Harvard Berkman-Klein Center]
& Aloni Cohen [University of Chicago]

Formally Modeling Legal Privacy Requirements

Goal: Rigorously analyze the relationship between legal and mathematical concepts.

Approach:

1. Identify a fundamental concept used in legal standards (e.g., singling out, linkability, inference).
2. Perform a thorough legal analysis of the concept.
3. Construct a mathematical model of the concept.
4. Check whether the modeling agrees with the legal analysis.
5. Compare the model with the mathematical definition (e.g., differential privacy).

FERPA de-identification

GDPR singling out

FERPA: Family Educational Rights and Privacy Act

Kobbi Nissim, Aaron Bembenek, Alexandra Wood, Mark Bun, Marco Gaboardi, Urs Gasser, David O'Brien, Thomas Steinke, and Salil Vadhan. 2018. "Bridging the Gap Between Computer Science and Legal Approaches to Privacy ". Harvard Journal of Law & Technology, 31:2 pp. 687-780, 2018.

FERPA: Family Educational Rights and Privacy Act

Protects personally identifiable information in education records maintained by educational agencies and institutions, including:

"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)*

FERPA: 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)

FERPA: Family Educational Rights and Privacy Act

Permits the release of **directory information** , as long as the students (or, if minors, their parents) have received notice and an opportunity to opt out.

"'Directory information' means information contained in an education record of a student that would not generally be considered harmful or an invasion of privacy if disclosed."

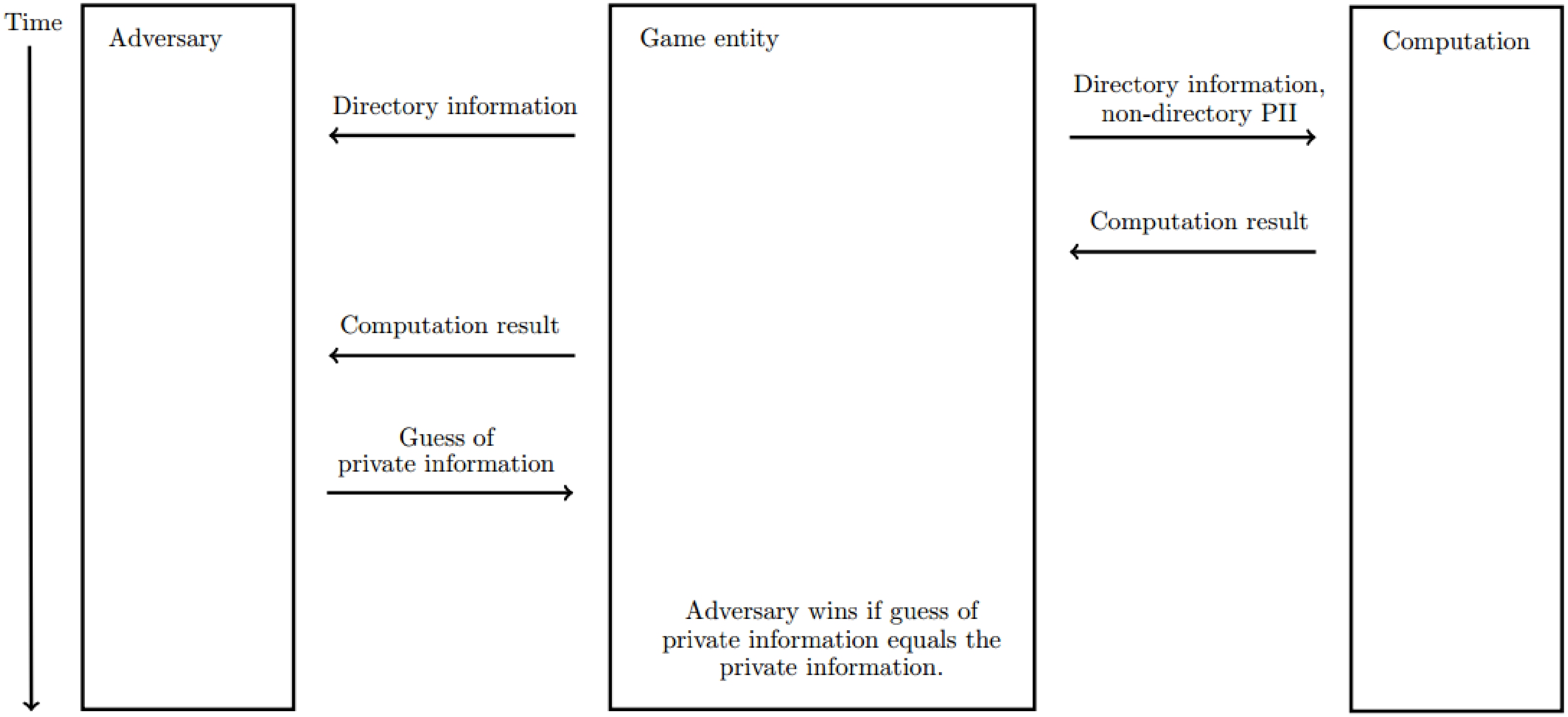
20 C.F.R. § 99.3

Extracting a Formal Definition from FERPA

FERPA allows the release of de-identified information and directory information from education records.

- We can model de-identification as a computation.

Components of a FERPA Privacy Game



Modeling FERPA: Directory Information

- Regulatory language ambiguous

- We interpret it conservatively, err on the side of safety
    - e.g. Allow the adversary to decide what constitutes directory information.
    - Gives a sufficient condition for satisfying the regulation.
    - Next example (singling out) takes the opposite approach, identifying a necessary condition.

Modeling FERPA: The Adversary

FERPA's def of PII : "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."

Gives us an **implicit adversary** .

- The adversary can have both high-level knowledge (e.g., demographics of school) and "insider" knowledge about specific individuals in local community.
- Model lack of personal knowledge via probability distributions.

Proving Differential Privacy Satisfies FERPA

- Formal model of FERPA requirements allows us to reason, with high confidence, about whether the use of a privacy technology satisfies FERPA.

- We can prove mathematically that any computation that is differentially private meets this definition, and (since the requirements of this definition are likely stricter than that of FERPA) thus satisfies the privacy requirements of FERPA.

Singling Out

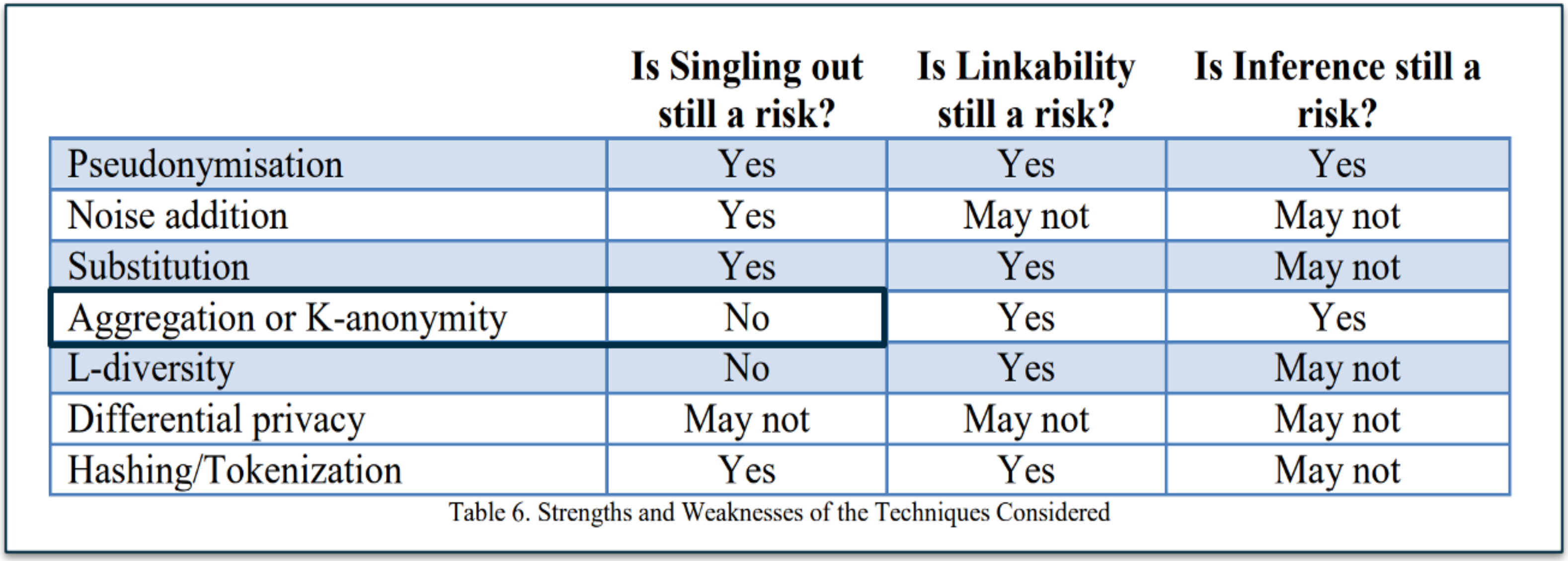
Aloni Cohen & Kobbi Nissim, Towards Formalizing the GDPR's Notion of Singling Out, 117 *Proceedings of the National Academy of Sciences* 8344 (2020).

Kobbi Nissim, Alexandra Wood, Micah Altman & Aloni Cohen, What a Hybrid Legal-Technical Analysis Teaches Us About Privacy Regulation: The Case of Singling Out, 27 *B.U. J. Sci. & Tech. L.* 1 (2021).

Singling Out

GDPR, Recital 26: “To determine whether a natural person is identifiable, account should be taken of all the means reasonably likely to be used, such as **singling out**, either by the controller or by another person to identify the natural person directly or indirectly.”

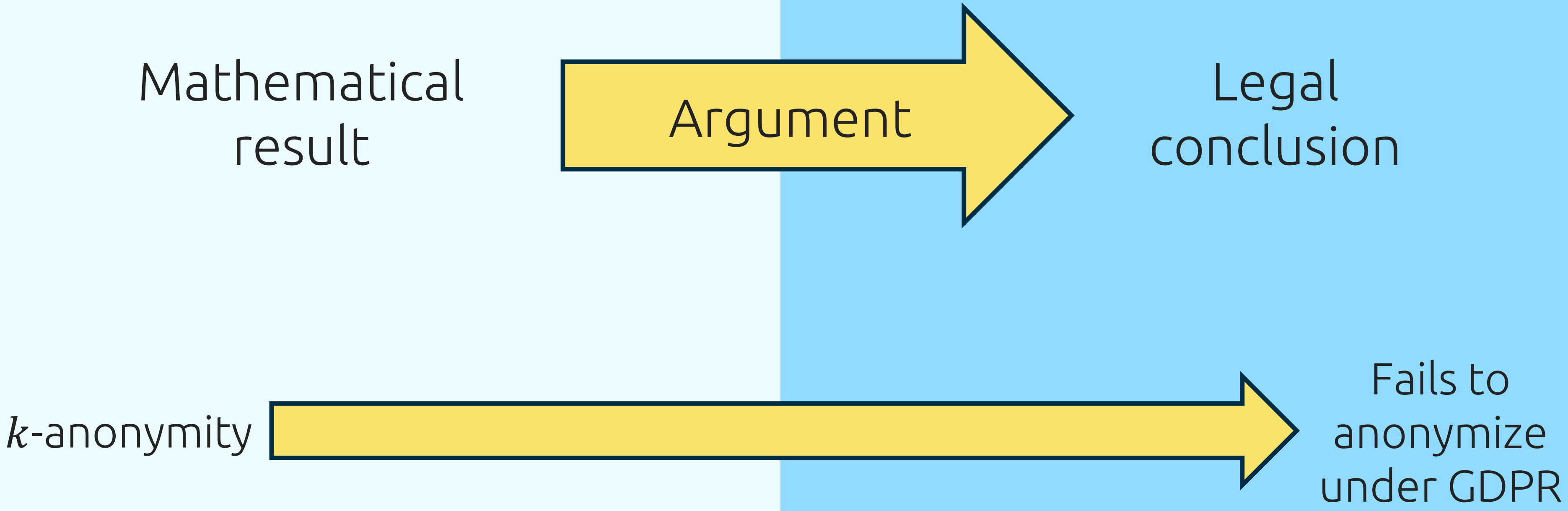
Article 29 Data Protection Working Party defines singling out as “the possibility to isolate some or all records which identify an individual in the dataset”



	Is Singling out still a risk?	Is Linkability still a risk?	Is Inference still a risk?
Pseudonymisation	Yes	Yes	Yes
Noise addition	Yes	May not	May not
Substitution	Yes	Yes	May not
Aggregation or K-anonymity	No	Yes	Yes
L-diversity	No	Yes	May not
Differential privacy	May not	May not	May not
Hashing/Tokenization	Yes	Yes	May not

Table 6. Strengths and Weaknesses of the Techniques Considered

Hybrid mathematical-legal theorem



Mathematical result

Argument

Legal conclusion

k-anonymity

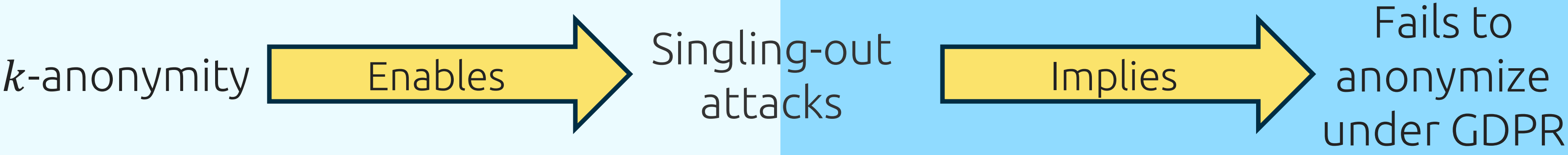
Fails to anonymize under GDPR

Math

Law

Singling-out attacks

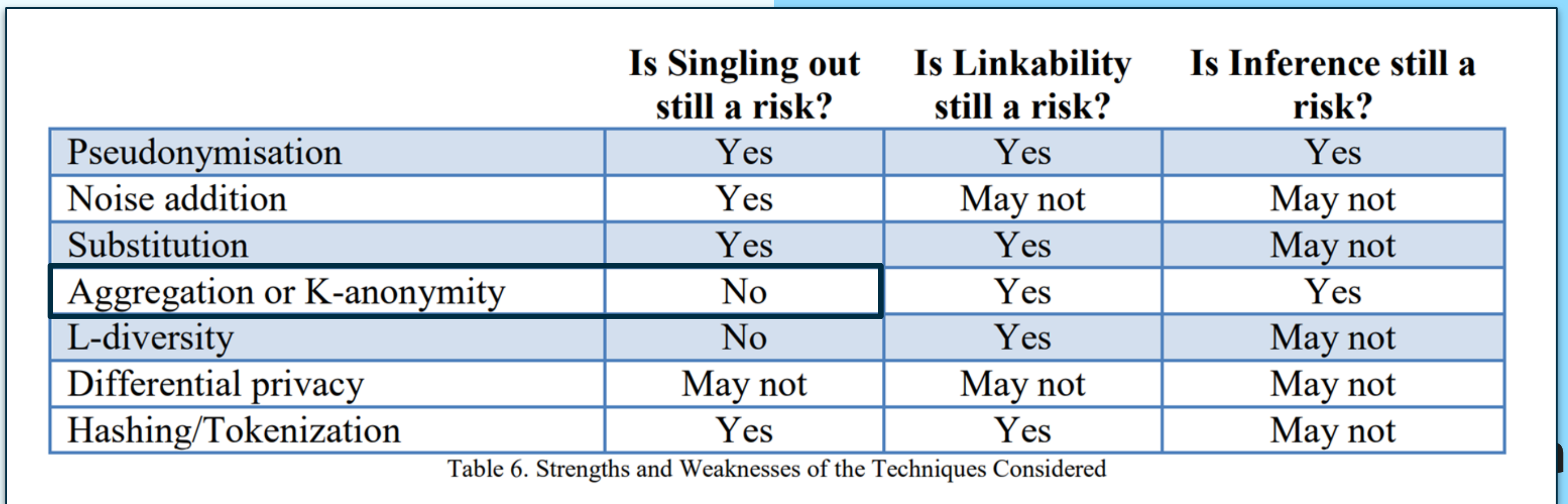
A new privacy notion



k-anonymity

Singling-out attacks

Fails to anonymize under GDPR



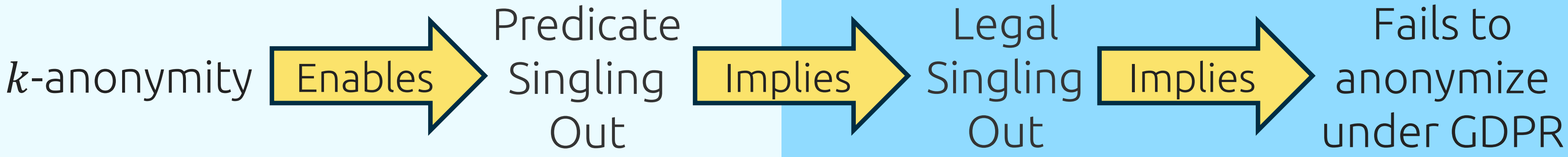
	Is Singling out still a risk?	Is Linkability still a risk?	Is Inference still a risk?
Pseudonymisation	Yes	Yes	Yes
Noise addition	Yes	May not	May not
Substitution	Yes	Yes	May not
Aggregation or K-anonymity	No	Yes	Yes
L-diversity	No	Yes	May not
Differential privacy	May not	May not	May not
Hashing/Tokenization	Yes	Yes	May not

Table 6. Strengths and Weaknesses of the Techniques Considered

Ma

w

Singling-out attacks



k-anonymity

Enables

Predicate Singling Out

Implies

Legal Singling Out

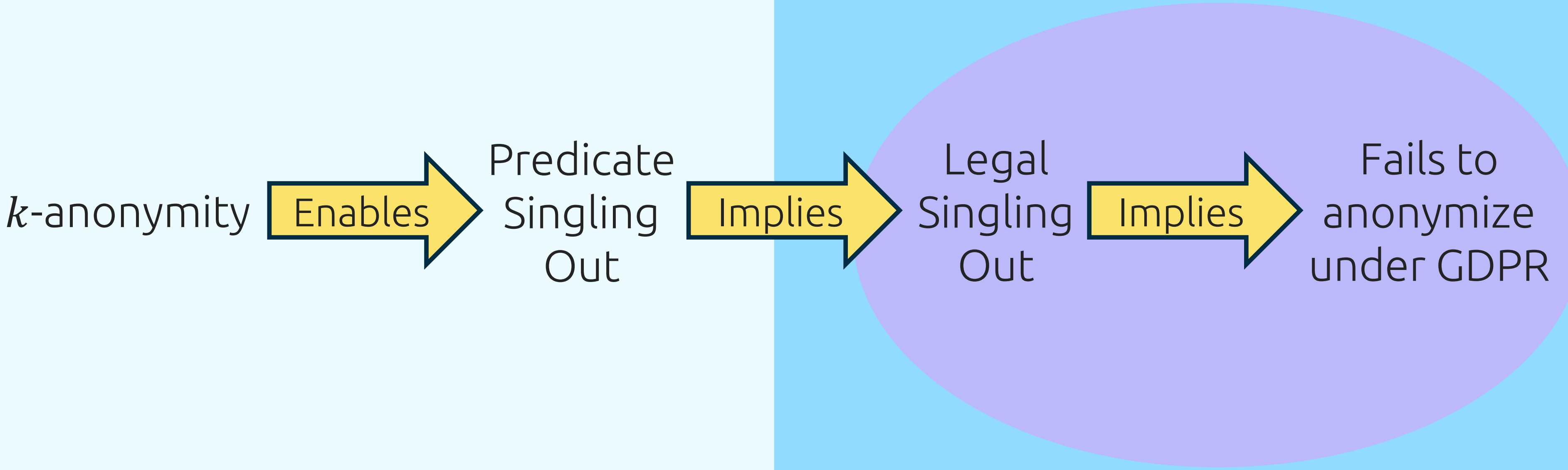
Implies

Fails to anonymize under GDPR

Math

Law

A hybrid theorem for singling-out



k-anonymity

Enables

Predicate Singling Out

Implies

Legal Singling Out

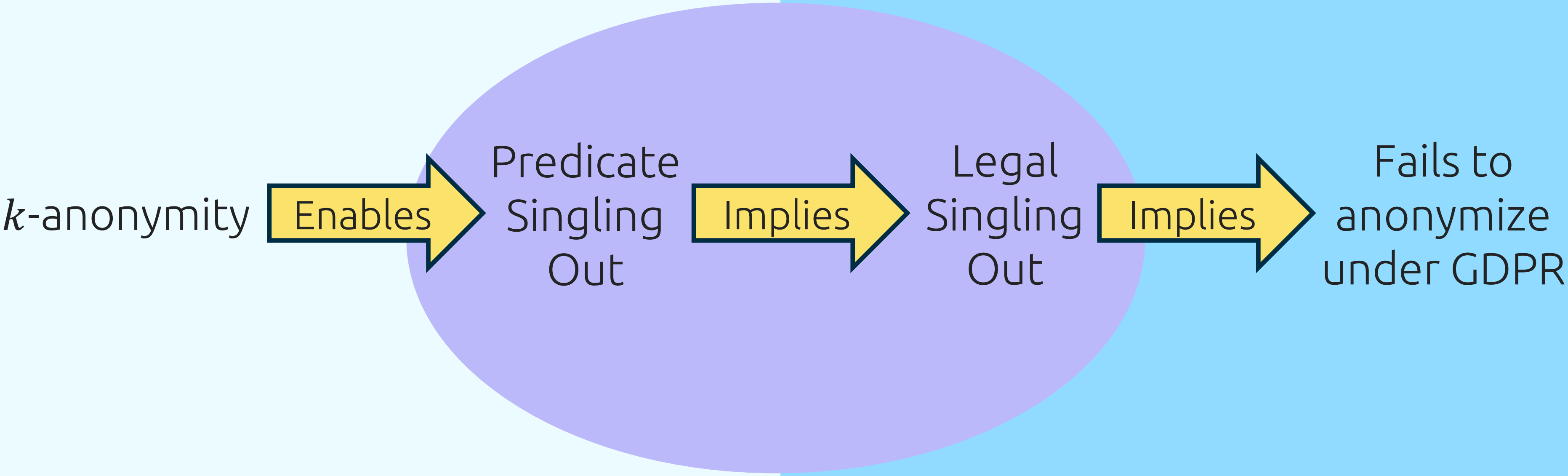
Implies

Fails to anonymize under GDPR

Math

Law

A hybrid theorem for singling-out



k-anonymity

Enables

Predicate Singling Out

Implies

Legal Singling Out

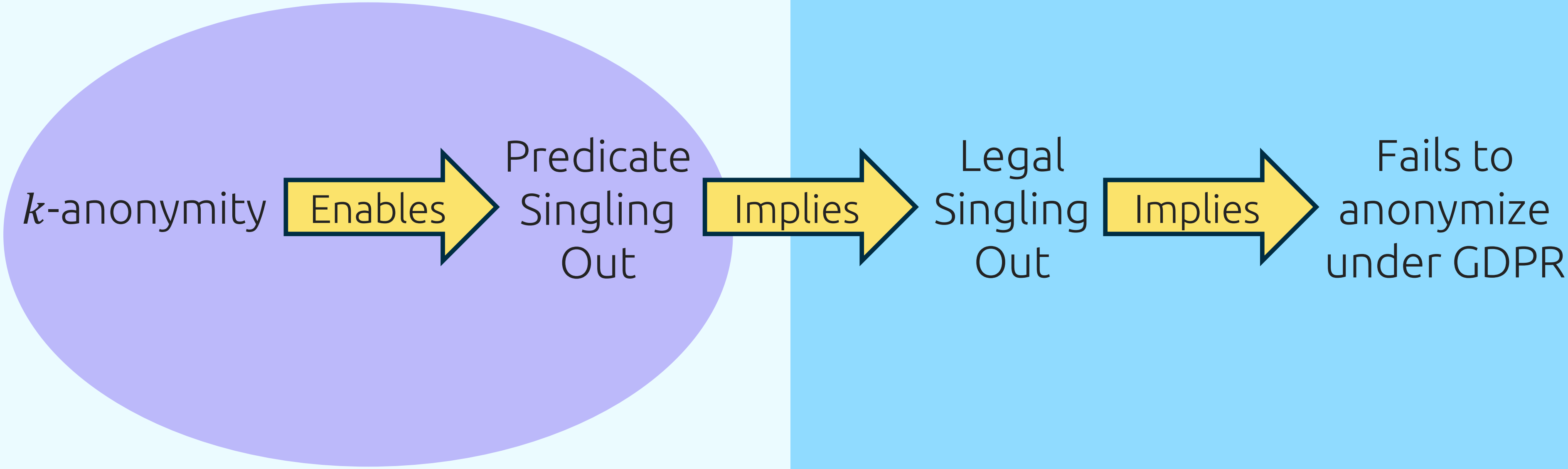
Implies

Fails to anonymize under GDPR

Math

Law

A hybrid theorem for singling-out



k-anonymity Enables Predicate Singling Out

Implies

Legal Singling Out

Implies

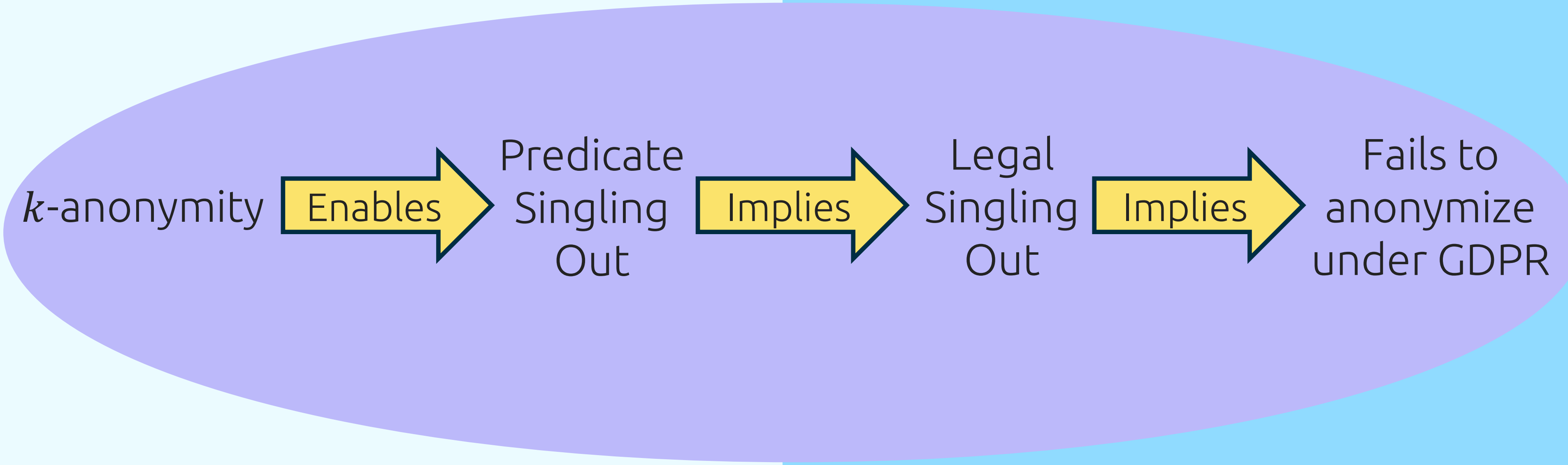
Fails to anonymize under GDPR

Differential privacy prevents predicate singling out

Math

Law

A hybrid theorem for singling-out



k-anonymity

Enables

Predicate Singling Out

Implies

Legal Singling Out

Implies

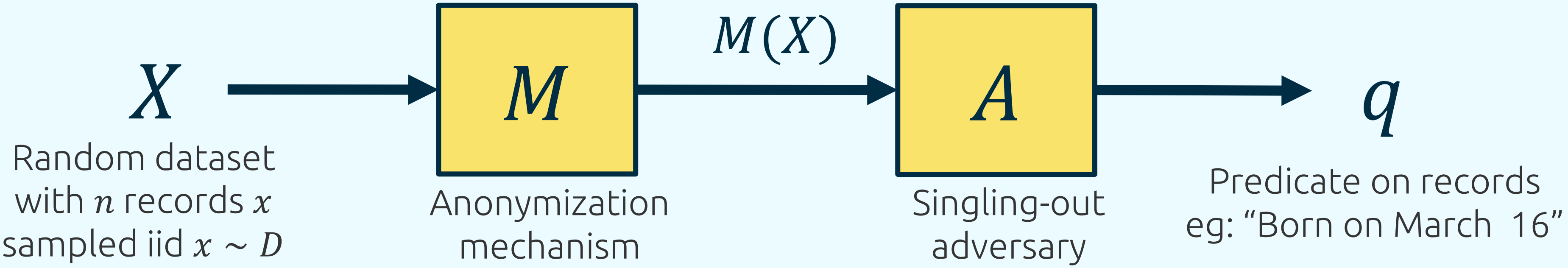
Fails to anonymize under GDPR

Differential privacy prevents predicate singling out

Math

Law

The setting



X

Random dataset with n records x sampled iid $x \sim D$

M

Anonymization mechanism

$M(X)$

A

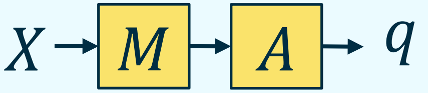
Singling-out adversary

q

Predicate on records eg: “Born on March 16”

Compare A’s ability to __________ before and after seeing the output M(X)

“q isolates in X” if it’s true on *exactly one* record in X



Predicate singling-out attacks by example

Isolation “q isolates in X” if it’s true on *exactly one* record in X

Example ($n = 365$)

q_1 = “Born on March 16th“
Attack: q_1 isolates ≫ 37% of the time

$\text{weight}(q_1) = \frac{1}{365} = \frac{1}{n}$

q_2 = “Vegan Colombian Jewish pilot fluent in Dutch”
Attack: q_2 isolates ≫ 0% of the time

$\text{weight}(q_2) \approx 0$

Baseline (informal)
How often A isolates before seeing $M(X)$. Depends on weight.

Weight of q Probability of matching a random record

$$\text{weight}(q) \coloneqq \Pr_{x\sim D}[q(x)]$$

Predicate singling-out attacks (informal)
A outputs low-weight q that isolates much more often than the baseline

Calculation

$$\Pr_X[q_1 \text{ isolates in } X] = 365 \cdot \frac{1}{365} \cdot \left(1 - \frac{1}{365}\right)^{364} \approx e^{-1} \approx 0.37$$

$$\Pr_X[q_2 \text{ isolates in } X] \le n \cdot \Pr_{x\sim D}[q_2(x)] \approx 0$$

Case Study: Privacy and the 2020 Census

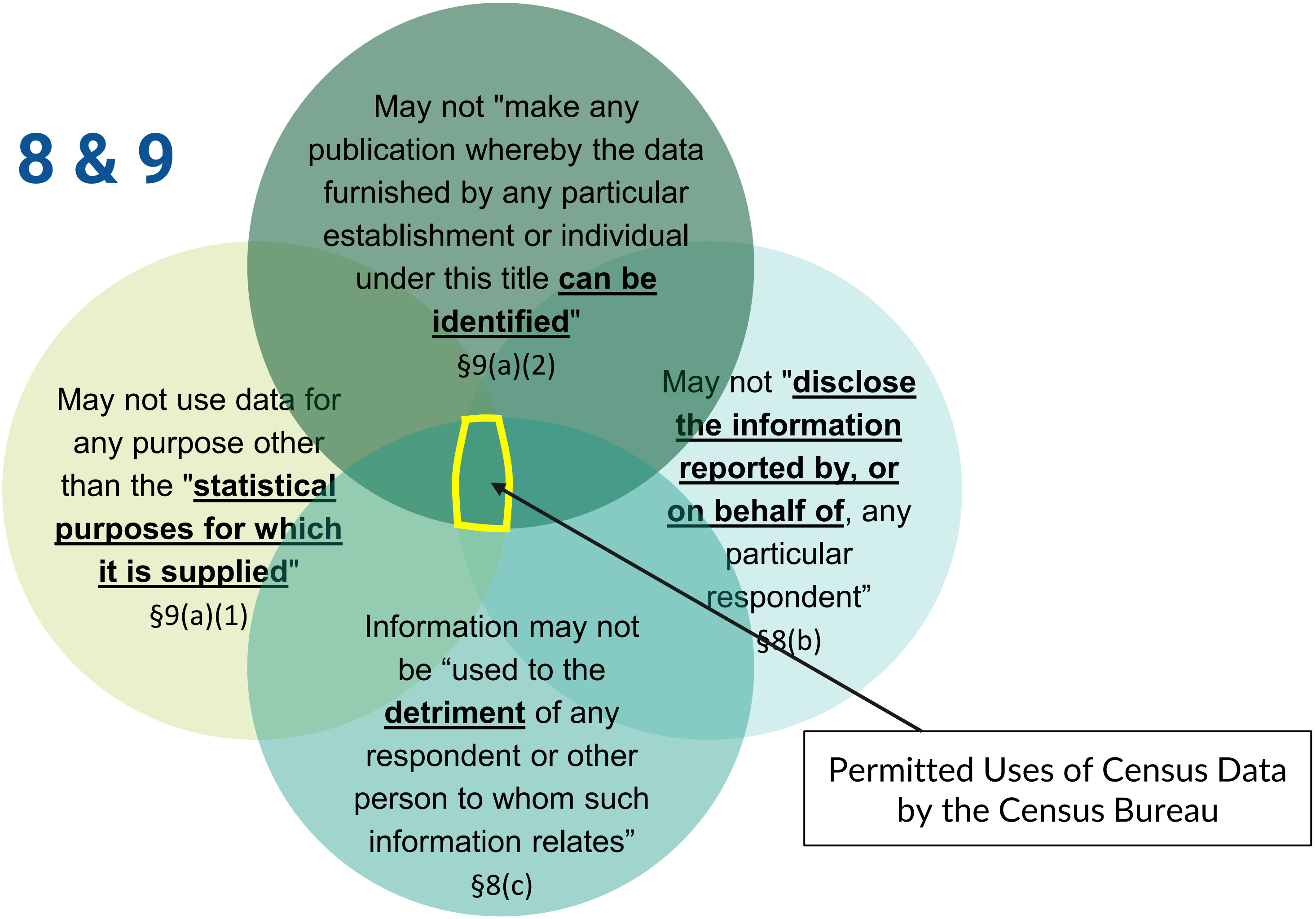
US Census Bureau’s dual statutory mandate

1. Collecting and publishing data necessary for democracy
2. Protecting the privacy of individuals to ensure trust and prevent harm

These goals are in conflict

- Fundamental law of information recovery: “overly accurate” estimates of “too many” statistics completely destroys privacy (Dinur & Nissim, Dwork & Roth)
- 2010 Census involved the release of approximately 7.8 billion statistics

Title 13, Sections 8 & 9



May not "make any publication whereby the data furnished by any particular establishment or individual under this title **can be identified**" §9(a)(2)

May not use data for any purpose other than the "**statistical purposes for which it is supplied**" §9(a)(1)

May not "**disclose the information reported by, or on behalf of**, any particular respondent" §8(b)

Information may not be "used to the **detriment** of any respondent or other person to whom such information relates" §8(c)

Permitted Uses of Census Data by the Census Bureau

Group Exercise 2: Privacy and the 2020 Census

Based on your read of the confidentiality mandate:

1. Does the Bureau’s confidentiality mandate require protection against reconstructing respondent data or inferring information specific to an individual with less than absolute certainty?

2. Does use of DP satisfy the Bureau’s confidentiality mandate? Under what conditions?

3. How are different stakeholder groups likely to interpret the Bureau’s duty to balance privacy and data utility differently?