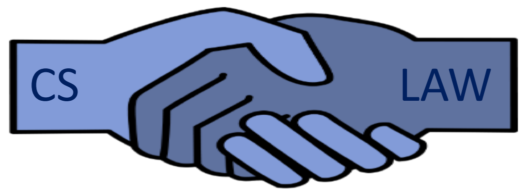
Legal theorems of privacy

Kobbi Nissim
Georgetown University



32nd IEEE Computer Security Foundations Symposium
[Hoboken 2019]

November 2019, two emails

From CSF 2019 Co-Chairs, Stéphanie Delaune and Limin Jia:

- On behalf of the program committee, we would like to invite you to be a keynote speaker at CSF.
- Your work on privacy will certainly be of great interest to the CSF audience. But of course you are welcome to speak on whatever topic you like.

November 2019, two emails

From CSF 2019 Co-Chairs, Stéphanie Delaune and Limin Jia:

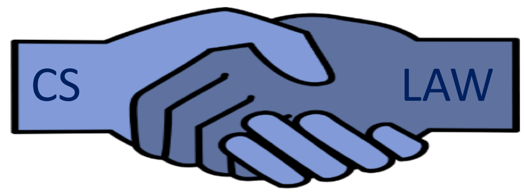
- On behalf of the program committee, we would like

From ####:

- On the behalf of International Association of Advanced Materials, it gives me a great pleasure to invite you to deliver a keynote talk in the thematic event on The Carbon Materials and Technology Conference.

Legal theorems of privacy

Kobbi Nissim
Georgetown University

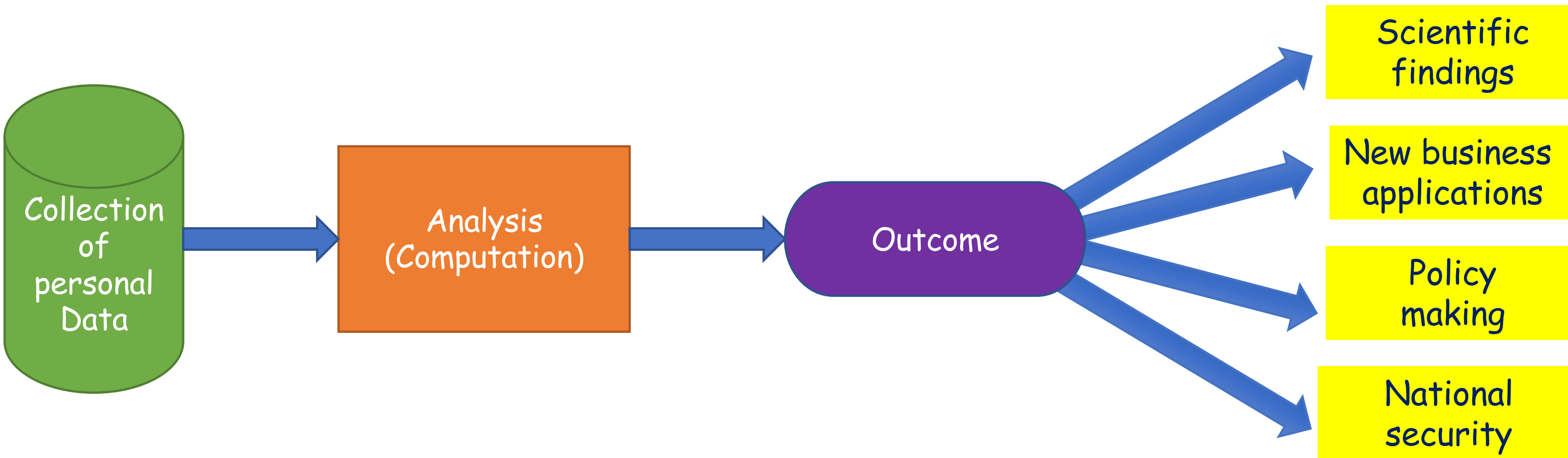


The presenter is not a lawyer

Based on collaborations with with:
M. Altman, A. Bembenek, M. Bun, A. Cohen, M. Gaboardi,
U. Gasser, D. O'Brien, T. Steinke, S. Vadhan, & A. Wood

32nd IEEE Computer Security Foundations Symposium
[Hoboken 2019]

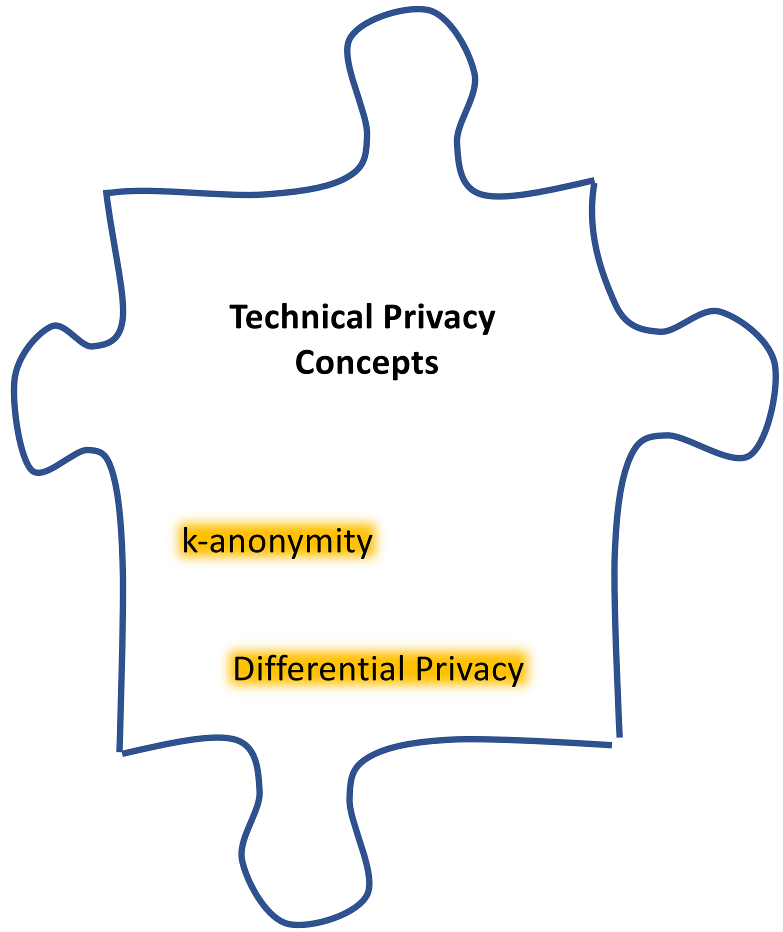
Data privacy: The problem

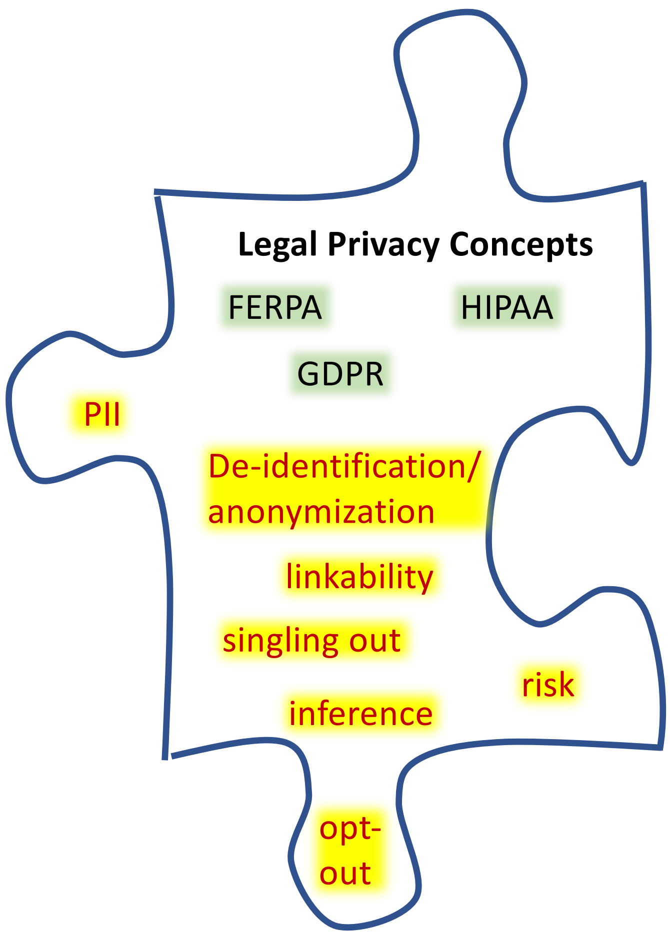


How to compute and release functions of datasets containing sensitive personal information while protecting individual privacy?

What does this mean?

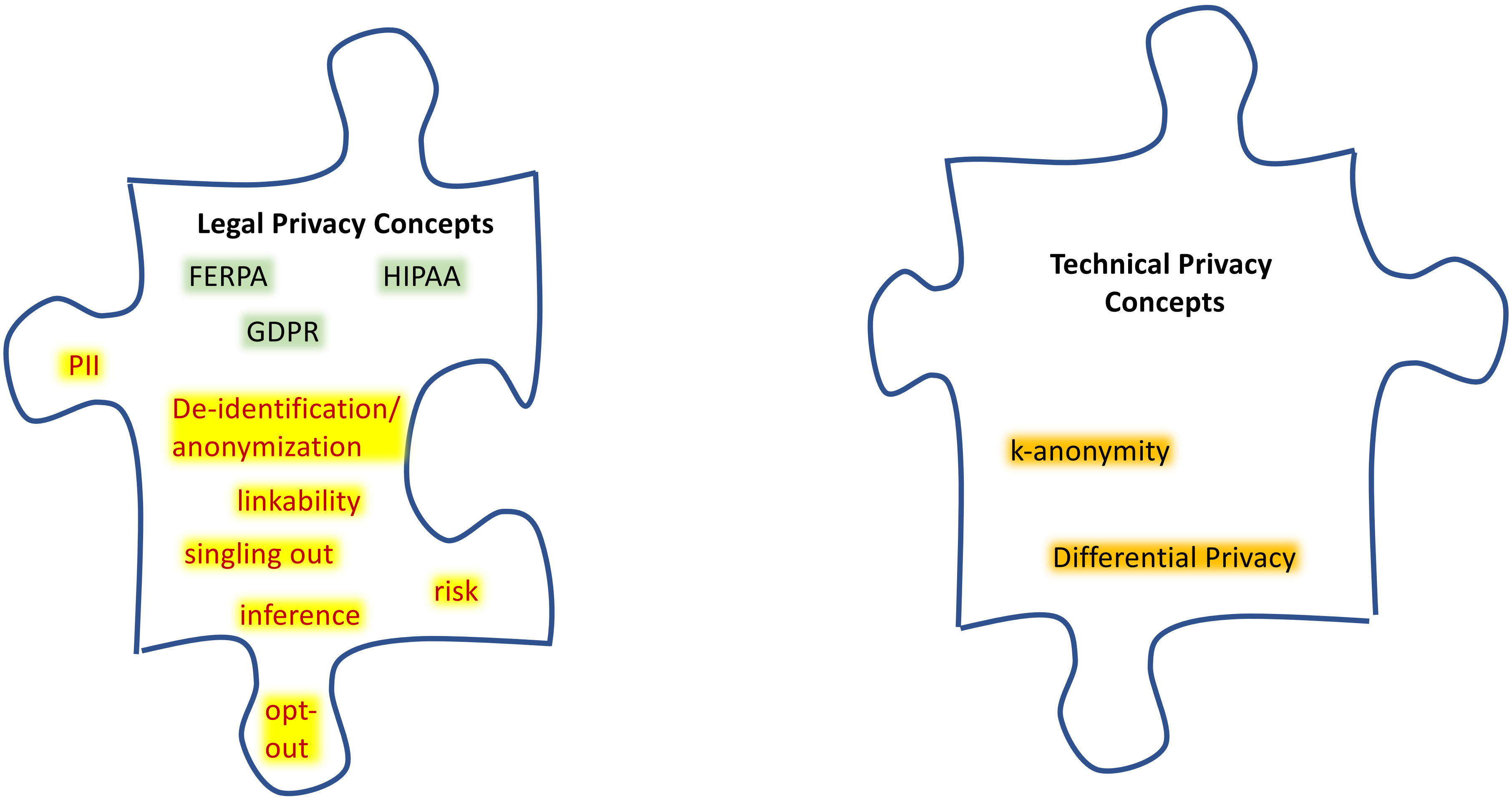
- Attempt to offer general privacy protection
- Uses mathematical language
- Seek to provide provable privacy guarantees



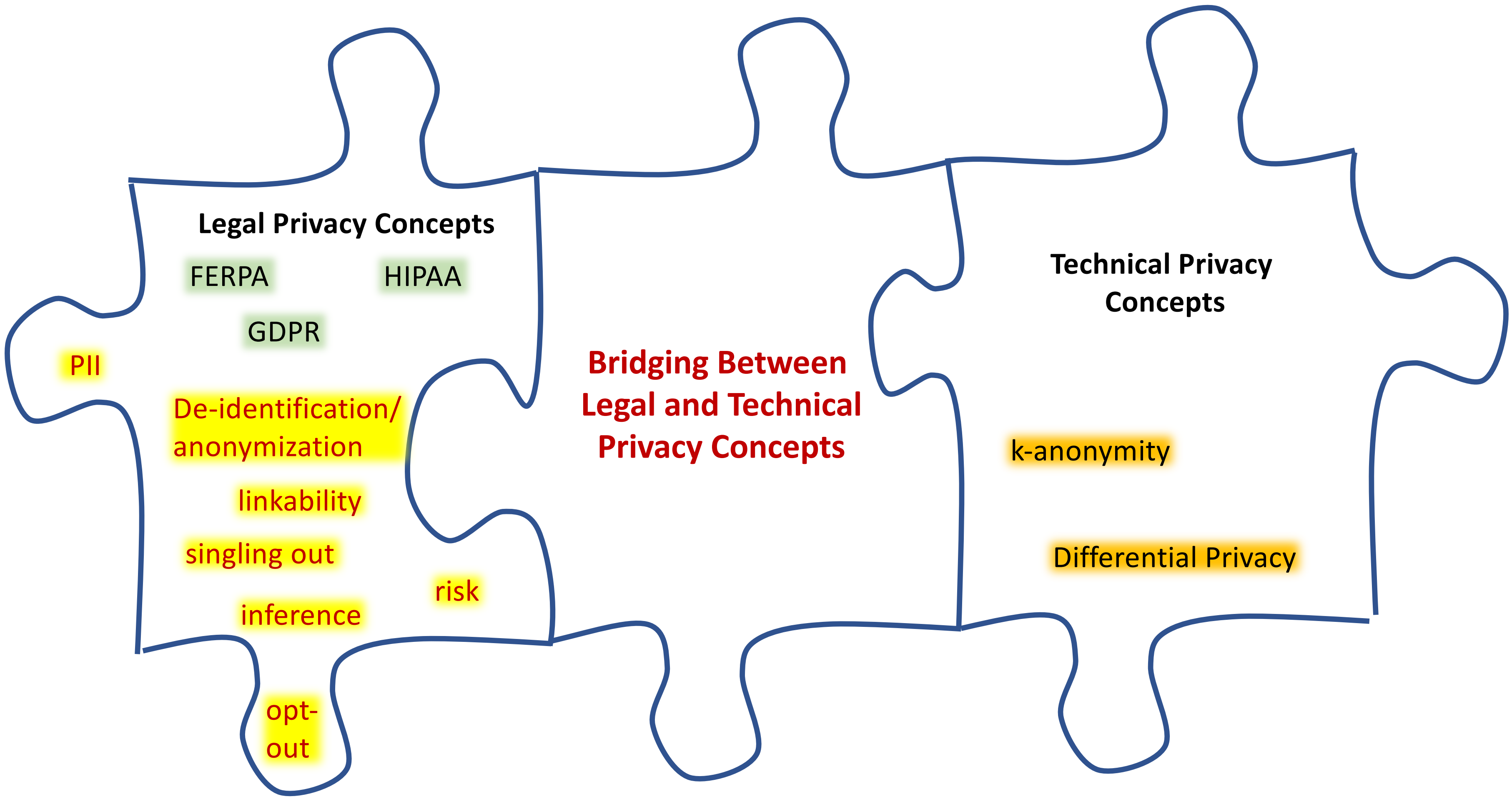


- Intuitive, not formal/accurate from a mathematical standpoint
- Often sector-based and non-general
- Leaves significant "gray areas", uncertainty
- Sometimes in disagreement with up-to-date scientific knowledge

FERPA: Family Educational Rights and Privacy Act;
HIPAA: Health Insurance Portability and Accountability Act;
GDPR: EU General Data Protection Regulation
PII: Personal Identifiable Information



Legal Privacy Concepts
FERPA
HIPAA
GDPR
PII
De-identification/
anonymization
linkability
singling out
risk
inference
opt-
out
Technical Privacy
Concepts
k-anonymity
Differential Privacy


Legal Privacy Concepts
FERPA
HIPAA
GDPR
PII
De-identification/
anonymization
linkability
singling out
inference
risk
opt-
out
Bridging Between
Legal and Technical
Privacy Concepts
Technical Privacy
Concepts
k-anonymity
Differential Privacy

This talk

Background:

- Privacy failures
- k-anonymity
- Differential privacy

CS and privacy law:

- Prior work
- Example: formalizing and reasoning about the GDPR's singling out

Summary, questions

Bridging Between Legal and Technical Privacy Concepts

GIC Linkage Attack [Sweeney '00]

GIC
Group Insurance
Commission

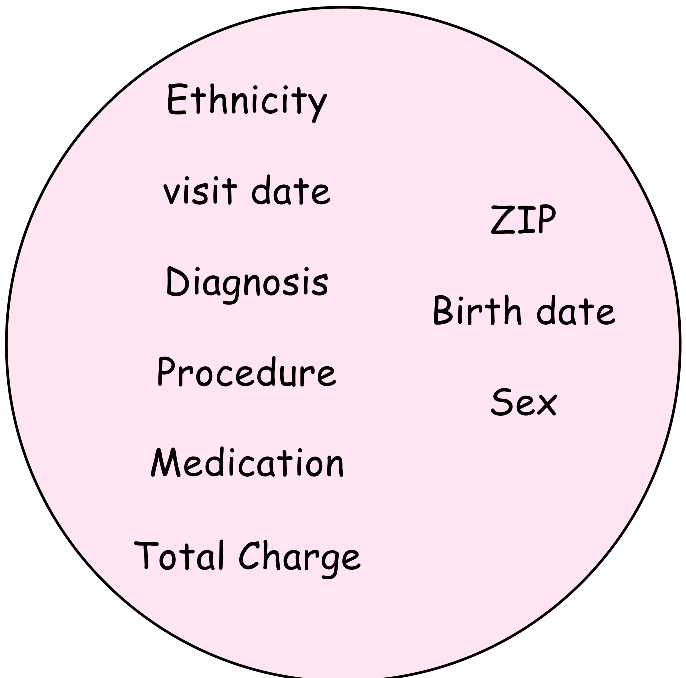
patient specific data
(≈ 135,000 patients)

≈100 attributes
per encounter

Anonymized

Common
Group Insurance Commission

GIC Linkage Attack [Sweeney ‘00]



GIC Linkage Attack [Sweeney '00]

Ethnicity

visit date

Diagnosis

Procedure

Medication

Total Charge

ZIP

Birth date

Sex

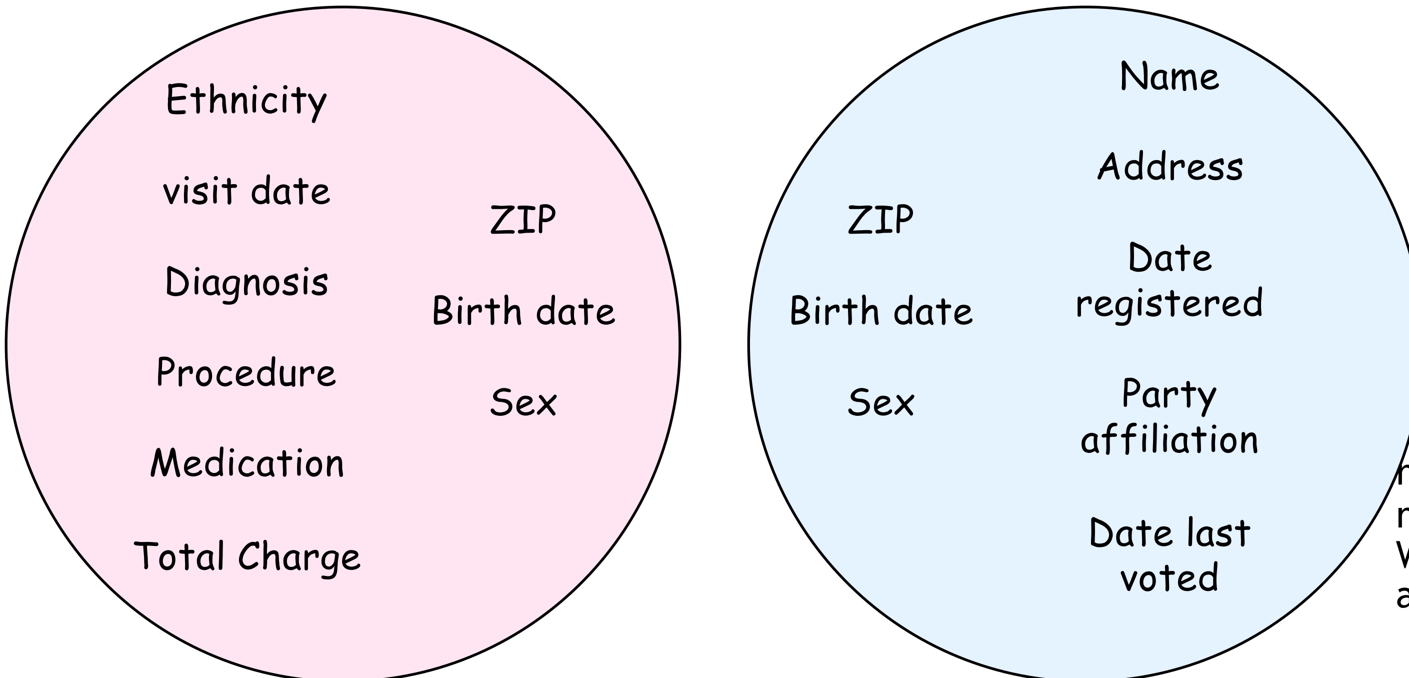
Voter registration
of Cambridge MA

"Public records"
open for inspection by
anyone

Common
Group Insurance Commission

REGISTER
TO VOTE!

GIC Linkage Attack [Sweeney '00]



Common Group Insurance Commission

REGISTER TO VOTE!

re-identified med records of William Weld (gov. of MA at the time)

*

Some privacy failures

- Re-identification [Sweeney '00, …]
  - GIC data, health data, clinical trial data, DNA, Pharmacy data, text data, registry information, …
- Blatant non-privacy [Dinur, Nissim '03], …
- Auditors [Kenthapadi, Mishra, Nissim '05]
- AOL Debacle '06
- Genome-Wide association studies (GWAS) [Homer et al. '08]
- Netflix award [Narayanan, Shmatikov '08]
- Social networks [Backstrom, Dwork, Kleinberg '07]
- Genetic research studies [Gymrek, McGuire, Golan, Halperin, Erlich '13]
- Microtargeted advertising [Korolova 11]
- Recommendation Systems [Calandrino, Kilzer, Narayanan, Felten, Shmatikov '11]
- Israeli CBS [Mukatren, Nissim, Salman, Tromer '14]
- Attack on statistical aggregates [Homer et al.'08] [Dwork, Smith, Steinke, Vadhan '15]
- Reconstruction attack on 2010 Census data

Slide idea stolen shamelessly from Or Sheffet


It was fun while it lasted
PRIVACY

Takeaways from Privacy Failures

Lack of rigor leads to unanticipated privacy failures

In setting clear meaningful privacy goals

In understanding how normative and technical conceptions of privacy interact

In analyzing resilience to future attacks

In taking auxiliary knowledge into account

In accounting for privacy loss across multiple releases

In scrutiny of privacy technology

Can we do better?

Maybe: k-anonymity and differential privacy

k-anonymity [Samarati Sweeney 98,Sweeney 02]

A k-anonymous dataset is achieved via suppression to make every combination of potentially identifying attributes appear at least k times

potentially identifying

ZIP	Age	sex	Disease
23456	55	Female	Heart
12345	30	Male	Heart
12346	33	Male	Heart
13144	45	Female	Breast Cancer
13155	42	Male	Hepatitis
23456	42	Male	Viral

ZIP	Age	sex	Disease
23456	**	*	Heart
1234*	3*	Male	Heart
1234*	3*	Male	Heart
131**	4*	*	Breast Cancer
131**	4*	*	Hepatitis
23456	**	*	Viral

In use!

- E.g., EdX data [Angiuli Blitzstein Waldo '15]

Does k-anonymity provide privacy?

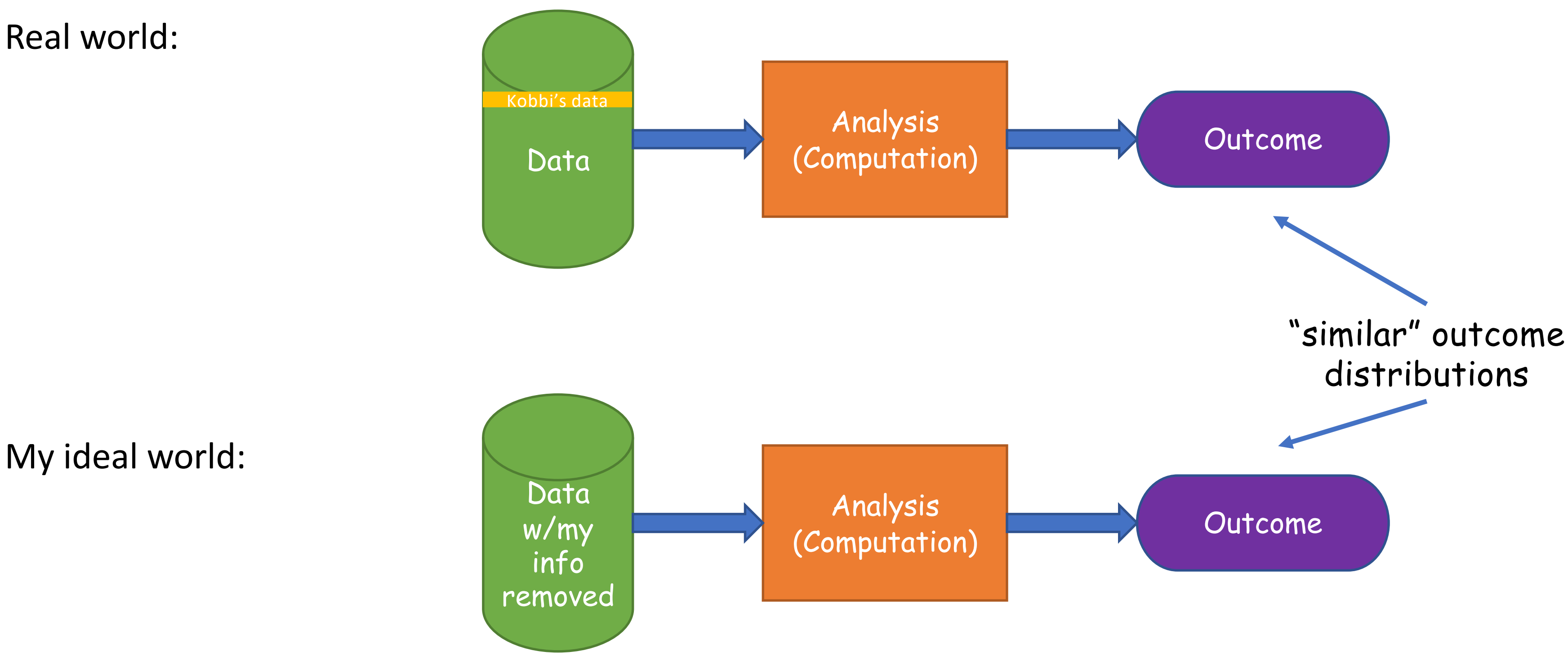
- k-anonymity is an intuitive syntactic condition on the outcome of an anonymization process, designed to foil Sweeney's linkage attack …
  - … but does not necessarily protect against other attacks
    - Homogeneity attacks, background attacks [Machanavajjhala et al 2007]
    - **Composition attacks** [Ganta et al 2008] [Cohen Nissim 2019, in preparation]
- Variants:
  - l-diversity [Machanavajjhala et al 2007]
  - t-closeness [Li et al 2007]
  - …

Differential Privacy [Dwork, McSherry, Nissim, Smith 2006]

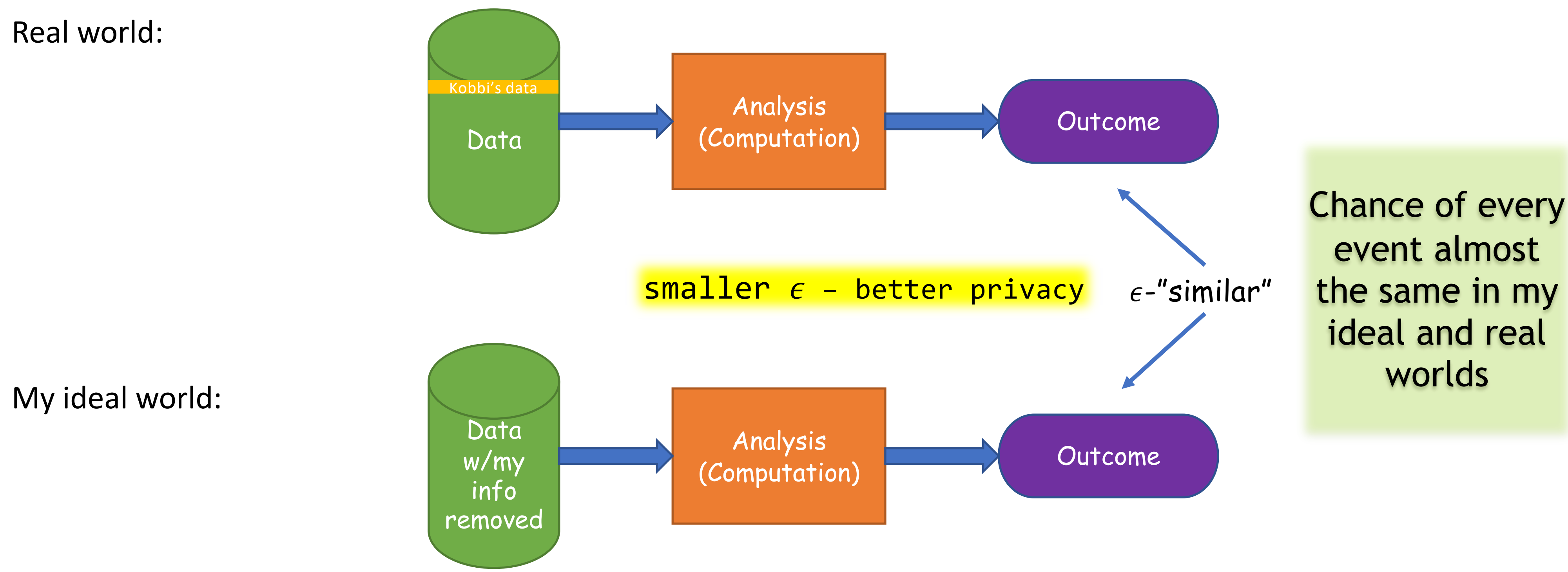
A mechanism is differentially private if:

Any information-related risk to a person should not change significantly as a result of that person's information being included, or not, in the analysis.

The differential privacy desiderata



The differential privacy desiderata



Differential privacy

A mechanism $M: X^n \rightarrow T$ satisfies ϵ-differential privacy if

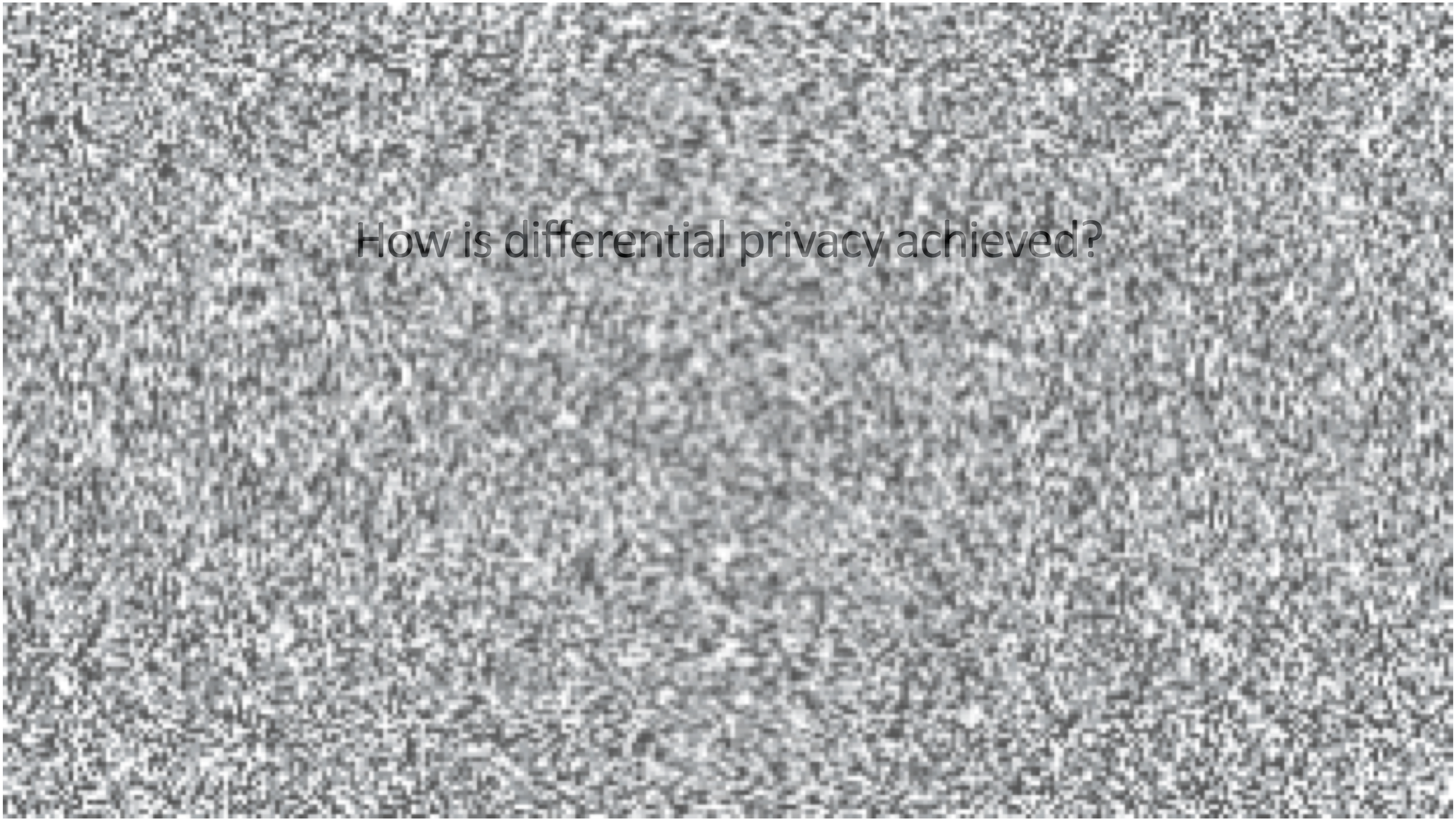
$$\forall x, x' \in X^n \text{ s.t. } dist_H(x, x') = 1 \ \forall S \subseteq T,$$

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

Why Differential Privacy?

- DP: Strong, quantifiable, composable mathematical privacy guarantee
- Provably resilient to attacks!
- Natural interpretation: I am protected (almost) to the extent I'm protected in my privacy-ideal scenario
- Theoretically, DP enables many computations with personal data while preserving personal privacy
  - Experience in practicing DP beginning to accumulate

How is Differential Privacy Achieved?

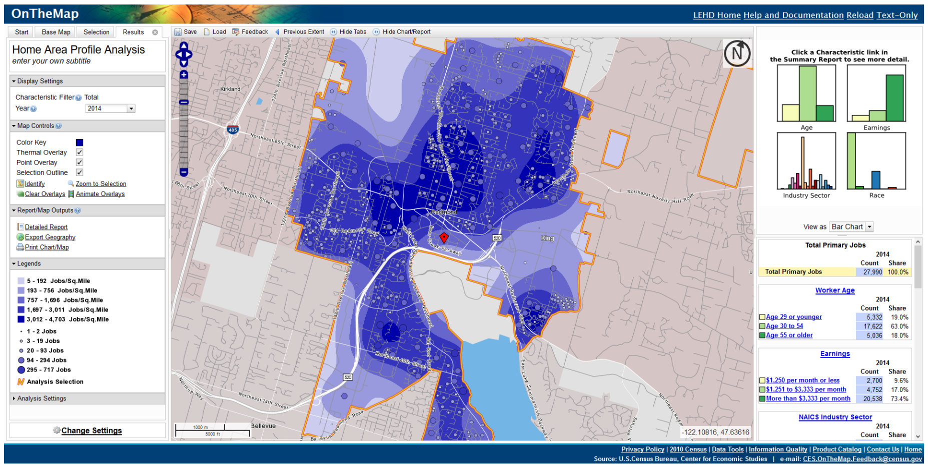

How is differential privacy achieved?

What can be Computed with Differential Privacy?

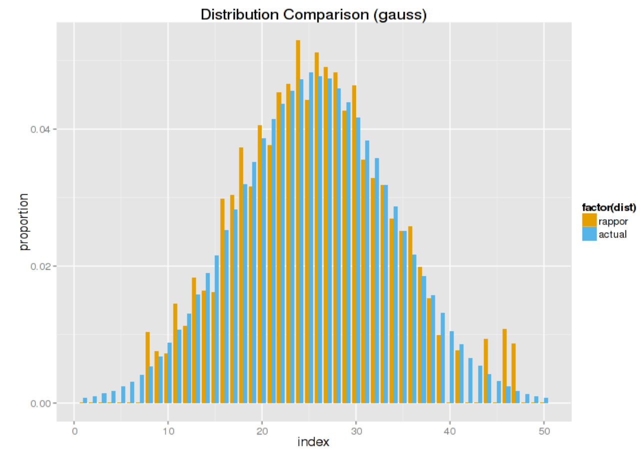
- Descriptive statistics: counts, mean, median, histograms, boxplots, etc.
- Supervised and unsupervised ML tasks: classification, regression, clustering, distribution learning, etc.
- Generation of synthetic data

Because of noise addition, differentially private algorithms work best when the number of data records is large

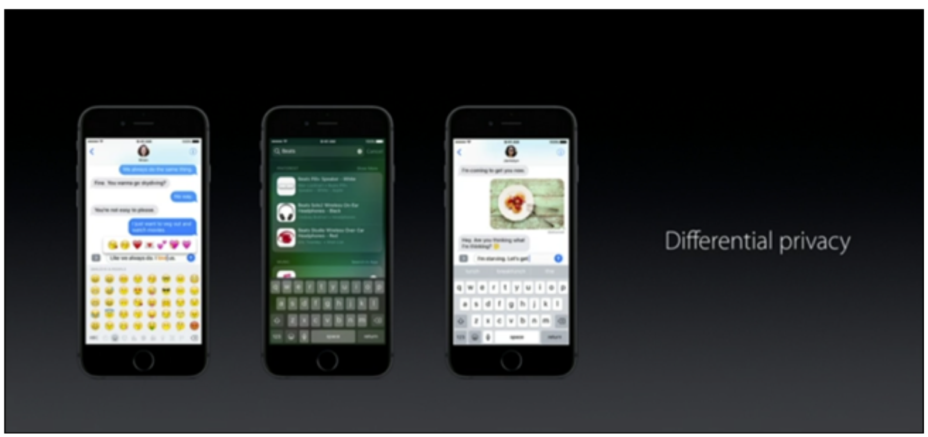
US Census' OnTheMap [2008] & 2020 Decennial



Google's RAPPOR [2014]



Apple's use of differential privacy [2016]



The Privacy Tools project [2018]

Some other efforts to bring DP to practice [partial list]

[Microsoft Research] PINQ

[UT Austin] Airavat: Security & Privacy for MapReduce

[UC Berkeley] GUPT

[CMU-Cornell-PennState] Integrating Statistical and Computational Approaches to Privacy

[US Census] OnTheMap

[Google] Rappor, TensorFlow Privacy

[UCSD] Integrating Data for Analysis, Anonymization, and Sharing (iDash)

[UPenn] Putting Differential Privacy to Work

[Stanford-Berkeley-Microsoft] Towards Practicing Privacy

[Duke-NISS] Triangle Census Research Network

[Harvard] Privacy Tools

[Georgetown-Harvard-BU] Formal Privacy Models and Title 13

[Harvard-Georgetown-Buffalo] Computing over Distributed Sensitive Data

This talk

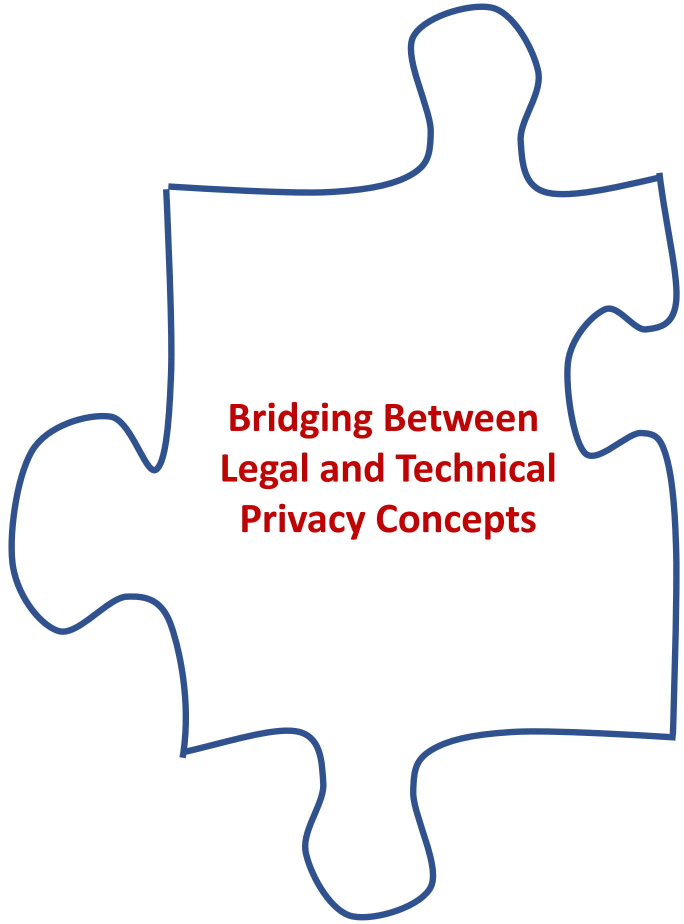
Background:

- Privacy failures
- k-anonymity
- Differential privacy

CS and privacy law:

- Prior work
- Example: formalizing and reasoning about the GDPR's singling out

Summary, questions



Do k-anonymity and differential privacy meet the expectations of legal privacy standards?

It's a total waste of our time!

- "I can easily litigate use of differential privacy in court"
- An impossible task!

- Yes, but that is not the point! We need to understand how out technical concepts related with societal concepts
- Yes, but we must tackle it!
  - **With as much rigor as possible!**

Related work (1): Contextual integrity [Nissenbaum]

- Framework for reasoning about privacy as norms about information flows between contexts
  - Combines ‘technical’ and ‘normative’ notions
  - Not accurate/formal from a mathematical standpoint

- [Barth, Datta, Mitchell, Nissenbaum] Formalized aspects of CI in logic for specifying and reasoning about norms of transmission of personal info
  - Use predicates such as $\mathbf{contains}(m, q, t)$ and $t \in \mathbf{npi}$ to specify a model which restricts the transmission of a message m about an individual q if m contains an attribute t which is non-public info
  - Do not specify when it is that a message m contains an attribute t about individual q (similarly, when it is that t is non-public info)

Related work (2): Robot Lawyers [Altman, Chong, Wood]

- Robot lawyers: automatic generation of a license for researchers download files from a social-science data repository
  - Inputs: Formalizations of legislation, license template, license terms, repository specific conditions; facts about dataset (via a questionnaire), …
  - Output: Human-readable license

- Formalization uses predicates such as $\mathbf{ferpa\_datasetInScope}(\mathrm{DS})$ and $\mathbf{ferpa\_identifiable}(\mathrm{DS})$ as a basis for deciding whether a release is permitted by FERPA
  - But does not specify (mathematically) when it is that a dataset should be considered FERPA identifiable

Related work(3): “Bridging” between technical and legal approaches to privacy*

- In an earlier work we examined Family Educational Rights and Privacy Act (FERPA) which governs the disclosure of personal information contained in education records
  - Observed that FERPA + guidance documents give many clues as to who the privacy attacker is and what is his goal
  - Extracted a *conservative* mathematical definition of privacy from FERPA
  - Provided a mathematical proof that DP satisfies this definition

* [Nissim, Bembenek, Wood, Bun, Gaboardi, Gasser, O’Brien, Steinke, Vadhan] Bridging the gap between computer science and legal approaches to privacy. Harvard Journal of Law & Technology, 2018.

This talk

Background:

- Privacy failures
- k-anonymity
- Differential privacy

CS and privacy law:

- Prior work
- Example: formalizing and reasoning about the GDPR's singling out

Summary, questions

Bridging Between Legal and Technical Privacy Concepts

Towards formalizing the GDPR notion of singling out
[with Aloni Cohen]

The GDPR (General Data Protection Regulation)

- Full title: “Regulation on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (Data Protection Directive)”

- Implementation date: 25 May 2018

Singling out

GDPR, Article 1:

"This Regulation lays down rules relating to the protection of natural persons with regard to the processing of personal data . . ."

GDPR, Article 4:

"Personal data means any information relating to an identified or identifiable natural person; an identifiable natural person is one who can be identified, directly or indirectly . . ."

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 . . . to identify the natural person directly or indirectly."

Singling out

Art. 29 Working Party:

"As regards indirectly identified or identifiable persons, this category typically relates to the phenomenon of unique combinations, whether small or large in size.
. . . A name may itself not be necessary in all cases to identify an individual. This may happen when other identifiers are used to single someone out."

	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

Singling out

Art. 29 Working Party:

> "As regards indirectly identified or identifiable persons, this category typically relates to the phenomenon of unique combinations, whether small or large in size.
> . . . A name may itself not be necessary in all cases to identify an individual. This may happen when other identifiers are used to single someone out."

Overall, by referring to singling out, the GDPR seems to higher the bar on what is considered anonymized data

Why?

- Singling out is a stepping stone towards re-identification
- Suffices for treating a person differently

Isolation

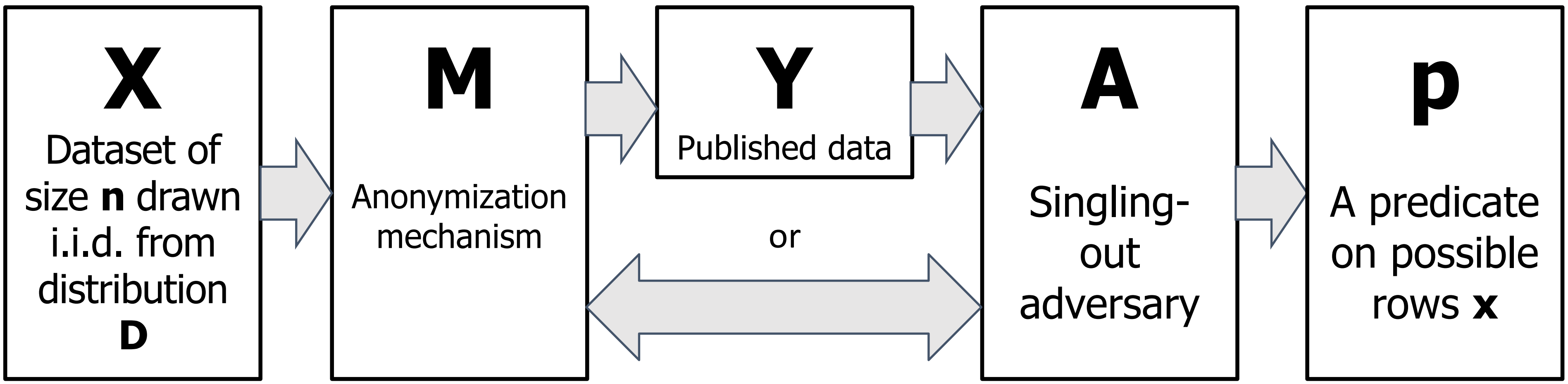
[Francis et al. 2018] Singling out as isolation: “there is exactly one person that has these attributes”

ID	Movie	Date (+/- 10)	Rating	Movie	Date	Rating	Movie	Date	Rating
1	Fargo	Jan 1	5	Mulan	Feb 2	5	Crash	Mar 3	5
2	Fargo	Jan 11	5	Mulan	Feb 29	5	Crash	Mar 13	5
3	The Sting	Jan 1	5	Mulan	Feb 2	5	Mad Max	Mar 3	5

Isolation examples: there is exactly 1 row in the underlying data that...

1. ... contains "The Sting”
2. ...... watched "Mulan" between Feb 19 and March 10
3. ... doesn't satisfy any of 1, or 2

Singling out = Isolation ?

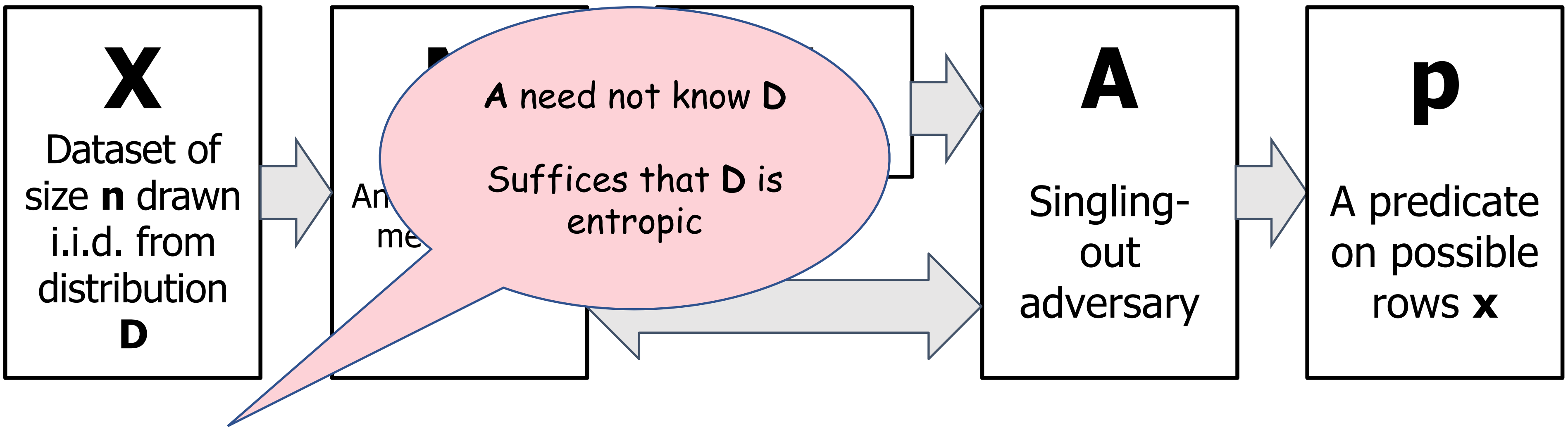


The **adversary's goal**: Given Y, output predicate p matching **exactly 1 row** in X.

Definition attempt: M is **secure against singling out** if no adversary can isolate a row except with **negligible probability** (over coins of X, M, A)

Impossible

Isolation with a trivial adversary



Choose p^* that matches a **random** ~1/n **fraction** of the universe.

$$\Pr[p^* \text{ isolates a row}] = n\left(\frac{1}{n}\right)\left(1-\frac{1}{n}\right)^{n-1} \approx \frac{1}{e} \approx 0.37$$

Can isolate (hence, single out) **without seeing Y**, succeed with probability 37%

Baseline: How well would a trivial adversary do?

- Definition: **weight(p)** = $\Pr_{x \leftarrow D}[p(x) = 1]$
- Def: **baseline(w)** to be the probability that a weight w predicate singles out.

$$\text{baseline}(w) = nw(1-w)^{n-1} \approx \mathbf{nwe^{-nw}}$$

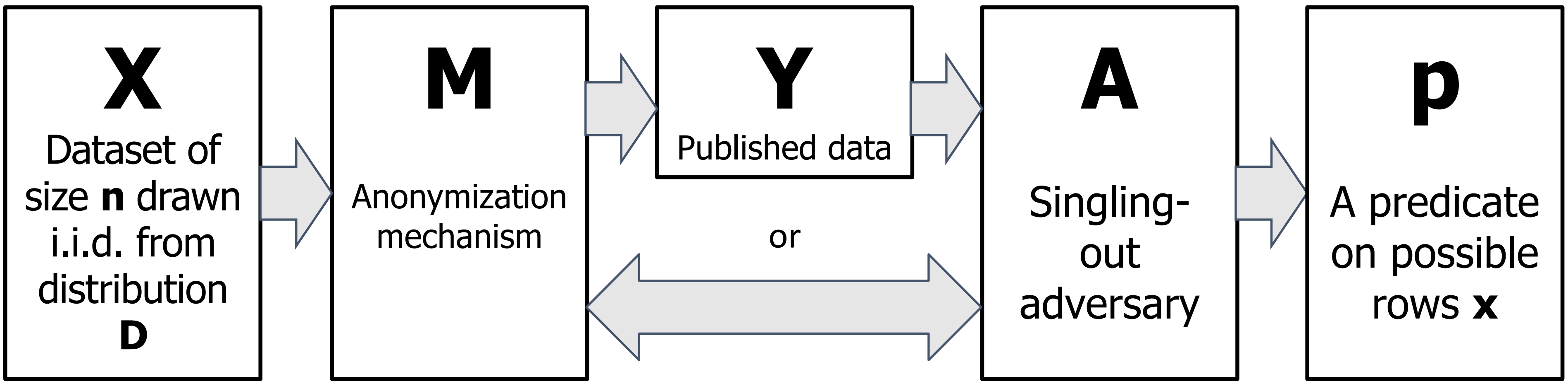
w	**baseline(w)**
negl(n)	negl(n)
$1/n^c$, for $c > 1$	$\approx 1/n^{c-1}$
c/n, for $c > 0$	$\approx ce^{-c}$
$\log(n^c)/n$, for $c > 0$	$\approx \log(n^c)/n^c$
$\omega(\log(n)/n)$	negl(n)

A predicate of weight negl(n) results in negl(n) success

A predicate of weight 1/n results in ~37% success probability

A predicate of weight ω(log(n)/n) results in negl(n) success

Security against predicate singling out (PSO)



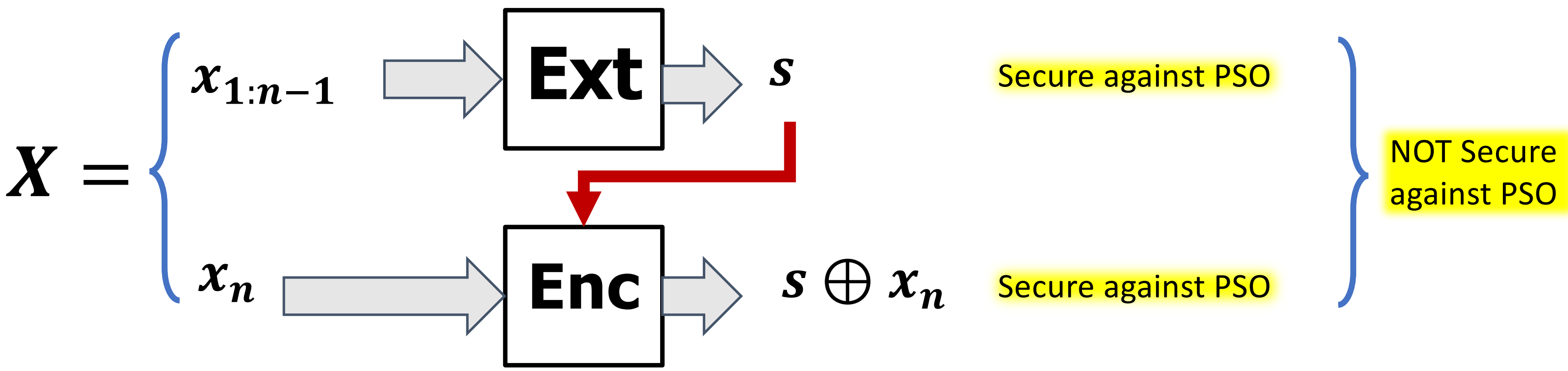
Definition*: M is **secure against predicate singling out** if no adversary can with non-negligible probability output a predicate p s.t.:

1) p matches **exactly 1 row** in X
2) p has **weight bounded away from 1/n**

* Some parameters omitted

Properties of security against PSO

- Given a definition, we can analyze its properties
- Claim: security against PSO does not self-compose



- A more natural example: there exists $\omega(\log n)$ count query mechanisms
  - Each secure against PSO; Their composition is not

Towards legal theorems

- Do k-anonymity and differential privacy protect against predicate singling out?
- Theorem: DP protects against predicate predicate singling out
- Proof via a Connection to generalization properties of differential privacy [Dwork, Feldman, Hardt, Pitassi, Reingold, Roth '15] [Bassily, Nissim, Smith, Steinke, Stemmer, Ullman '16]

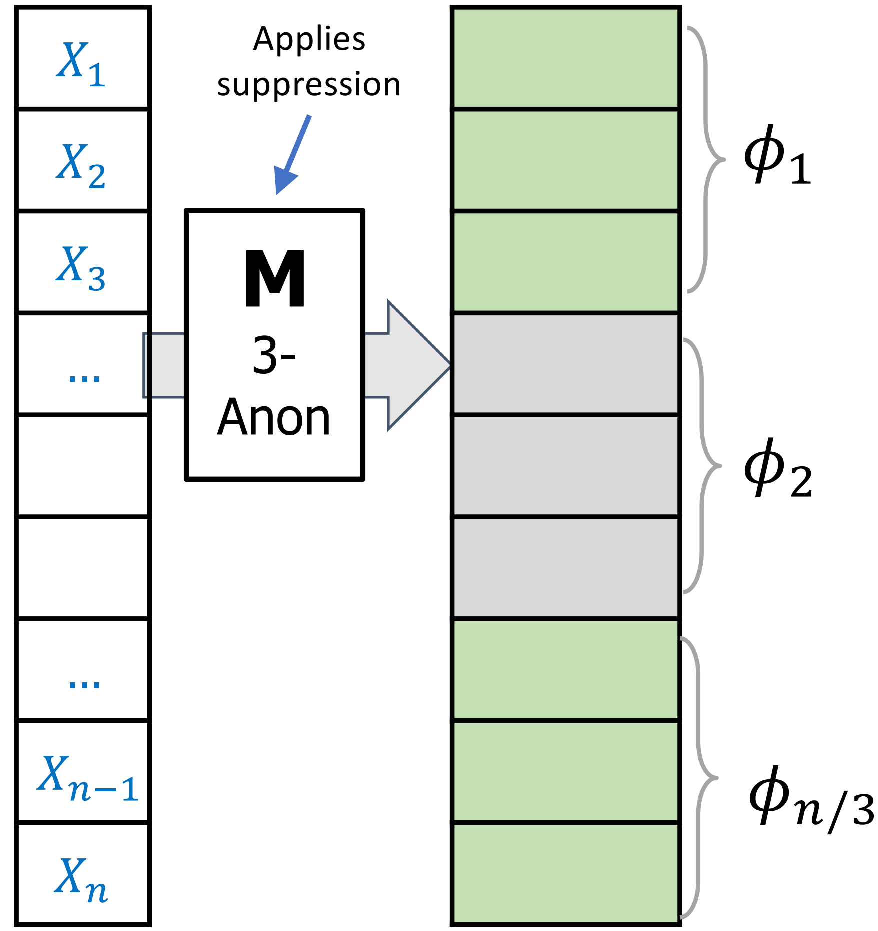
k-Anonymity & Predicate singling out

- Observation: k-anonymizer outputs predicates ϕ s.t. $\phi(X) \geq \frac{k}{n}$
  - "Typically" $\text{weight}(\phi)$ is tiny
    - Anonymizers try to suppress as little as possible

Singling out adversary:
- Choose p_k^* with weight $1/k$
- Output $p = p_k^* \wedge \phi$
  - $\text{weight}(p) \leq \text{weight}(\phi)$
- p isolates if p_k^* isolates in a chunk

$$\Pr[\text{iso}(p, X)] \approx \text{base}(k, k^{-1}) \approx e^{-1}$$

- Proof idea: Each row in a chunk has min-entropy, conditioned the rest of the chunk



Implications for GDPR compliance

- Positive results have restricted implications:
  - PSO security may be too weak (X drawn i.i.d. from D, no auxiliary knowledge)
  - Preventing predicate singling out attacks is necessary, but possibly not sufficient
  - Hence, determining whether the use of differential privacy satisfies GDPR requires more research
- Negative results most legally meaningful:
  - Restricted scope (X drawn i.i.d. from D, no auxiliary knowledge) strengthens negative results
  - Show that k-anonymity likely does not provide sufficient protection against singling out; Probably does most of the work for a singling out attacker

Back to the Art. 29 Working Party assesment

We respectfully disagree…

	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

Is predicate singling out a good privacy concept?

- It is useful for examining disclosure limitation concepts such as differential privacy and k-anonymity w.r.t. legal requirements such as in the GDPR 👍

- Does not self compose! 👎

This talk

Background:

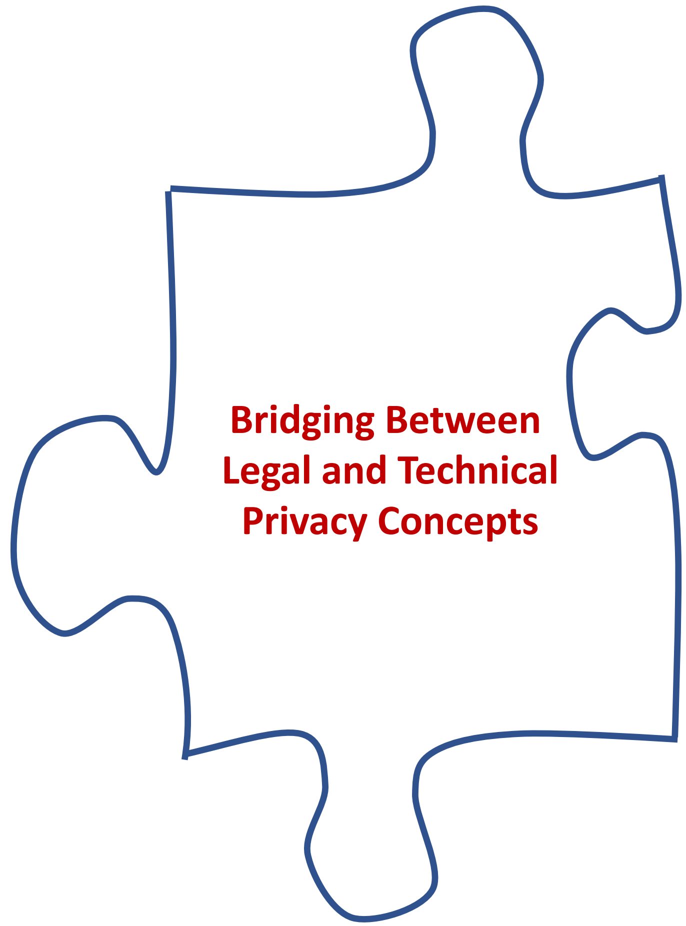
- Privacy failures
- k-anonymity
- Differential privacy

CS and privacy law:

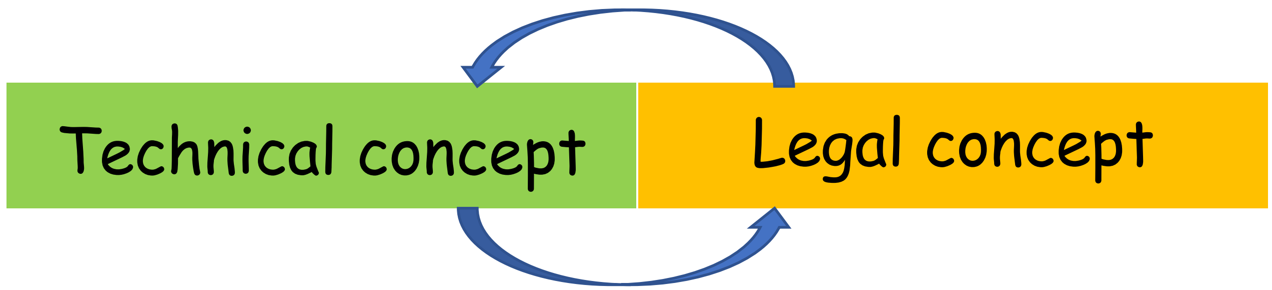
- Prior work
- Example: formalizing and reasoning about the GDPR's singling out

Summary, questions

- (answers not guaranteed)
- (Then coffee)



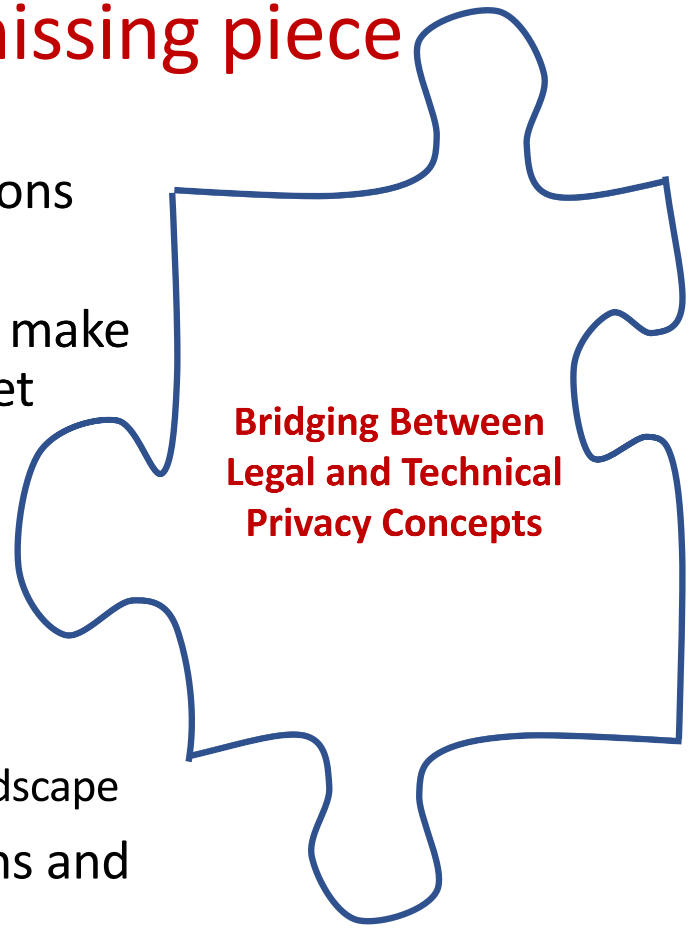
Summary: what have we seen?



• Definition of PSO security	• GDPR notion of singling out
• PSO security does not compose	• GDPR singling out security likely doesn't compose
• k-anonymization is not PSO secure	• k-anonymization likely does not prevent GDPR singling out
• DP is PSO secure	• Evidence that DP prevents GDPR singling out

Summary: An important missing piece

- More and more technologists need to make decisions with normative ethical and legal implications
- More and more lawyers and policymakers need to make decisions on the sufficiency of technologies to meet ethical and legal expectations
- A llitany of bad/uninformed decisions on privacy
- Missing in the current discussion:
  - Common vocabulary (we use the same words, but with different and incompatible meanings)
  - Ways to argue, rigorously, about the legal-technical landscape
- The CSF community has interests in these questions and tools to address them



References

- Bridging the Gap between Computer Science and Legal Approaches to Privacy. K. Nissim, A. Bembenek, A. Wood, M Bun, M Gaboardi, U. Gasser, D. O'Brien, T. Steinke, & S. Vadhan. Harvard Journal of Law and Technology, Spring 2018.
- Is Privacy *Privacy?* K. Nissim & A. Wood. Philosophical Transaction of the Royal Society, August 2018.
- Differential Privacy: A Primer for a Non-Technical Audience. A. Wood, M. Altman, A. Bembenek, M. Bun, M. Gaboardi, J. Honaker, K. Nissim, D. O'Brien, T. Steinke, S. Vadhan. Vanderbilt Journal of Entertainment and Technology Law, 2018.
- Towards Formalizing the GDPR's Notion of Singling Out. A. Cohen & K. Nissim. 2019. (available on arXiv).
- Hybrid Legal-Technical Concepts of Privacy. K. Nissim, A. Wood, M. Altman, & A. Cohen. (very preliminary version presented in PLSC 2018, available from authors).

Thank you!

*

Learning More About Differential Privacy

- [Page et al, 2018] Differential Privacy: An Introduction For Statistical Agencies, UK ONS.
- [Wood et al, 2019] Differential Privacy: A Primer for a Non-technical Audience, Vanderbilt JETLaw.
- [Nissim et al, 2018] Bridging the gap between computer science and legal approaches to privacy, Harvard JOLT.
- [Dwork 2011] A Firm Foundation for Private Data Analysis, CACM January 2011.
- [Heffetz & Ligett, 2014] Privacy and Data-Based Research, Journal of Economic Perspectives.
- [Dwork & Roth, 2014] The Algorithmic Foundations of Differential Privacy, Now publishers.
- [Vadhan, 2017] The Complexity of Differential Privacy

less technical

technical