

Differentially Private: Meaning and Caveats

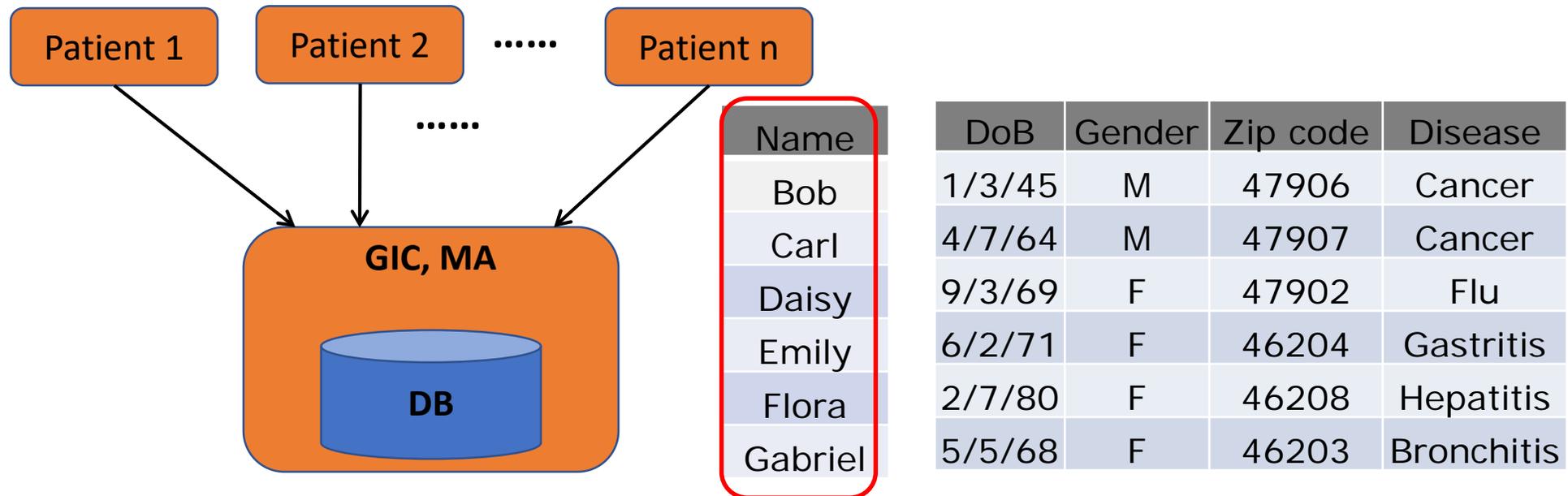
Presenter: Ninghui Li

Purdue University

2023.08.07 @ AI for Open Society 2023

GIC Incidence [Sweeny 2002]

- Group Insurance Commissions (GIC, Massachusetts)
 - Collected patient data for ~135,000 state employees.
 - Gave to researchers and sold to industry.
 - Medical record of the former state governor is identified.



Re-identification occurs!

k-Anonymity [Sweeney, Samarati]

The Microdata

QID			SA
Zipcode	Age	Gen	Disease
47677	29	F	Ovarian Cancer
47602	22	F	Ovarian Cancer
47678	27	M	Prostate Cancer
47905	43	M	Flu
47909	52	F	Heart Disease
47906	47	M	Heart Disease

A 3-Anonymous Table

QID			SA
Zipcode	Age	Gen	Disease
476**	2*	*	Ovarian Cancer
476**	2*	*	Ovarian Cancer
476**	2*	*	Prostate Cancer
4790*	[43,52]	*	Flu
4790*	[43,52]	*	Heart Disease
4790*	[43,52]	*	Heart Disease

□ k-Anonymity

- Each record is indistinguishable from $\geq k-1$ other records when only “quasi-identifiers” are considered
- These k records form an equivalence class

What is Privacy?

It is complicated!

Some concepts from the book “Understanding Privacy” by Daniel J. Solove:

1. the right to be let alone
2. limited access to the self
3. **secrecy—the concealment of certain matters from others;**
4. **control over others' use of information about oneself**
5. personhood—the protection of one’s personality, individuality, and dignity;
6. intimacy—control over, or limited access to, one’s intimate relationships or aspects of life.

Impossibility of “Privacy as Secrecy”

- Dalenius [in 1977] proposes this as privacy notion: *“Access to a statistical database should not enable one to learn anything about an individual that could not be learned without access.”*
 - Similar to the notion of semantic security for encryption
 - Not possible in the context if one wants utility.
- The Smoker example:
 - *Adversary knows “Terry is a smoker”*
 - *Published data reveal that “smokers are much more likely to get lung cancer”*
 - *Seeing published info enable learning that Terry’s risk for lung cancer is high*

Hiding personal information is hard because of correlation.

Differential Privacy [Dwork et al. 2006]

- Definition: Mechanism A satisfies ϵ -Differential Privacy if and only if
 - for any **neighboring** datasets D and D'
 - and any possible transcript $t \in \text{Range}(A)$,
$$\Pr[A(D) = t] \leq e^\epsilon \Pr[A(D') = t]$$
 - For relational datasets, typically, datasets are said to be **neighboring** if they differ by a single record.

Genius of Idea Behind DP

- Privacy is hard because which information to hide is difficult to enumerate and information may correlate
- By identifying a world without one individual's data as an ideal world for the individual, and providing real-world-ideal-world bound, DP does not need to provide prior-to-posterior bound, and does not need to deal with data correlation
- DP simulates privacy is “control over others' use of information about oneself”

The Personal Data Principle

- Data privacy means giving an individual control over his or her personal data. An individual's privacy is not violated if no personal data about the individual is used.
- Privacy does not mean that no information about the individual is learned, or no harm is done to an individual; enforcing the latter is infeasible and unreasonable.

Some Caveats of Applying DP

- How neighboring datasets is defined.
- What constitutes an individual's data: One individual's data or personal data under one individual's control
- Group privacy
- Moral challenge
- Choosing epsilon value
- Learning models and applying to individuals

An Example Adapted from [Kifer and Machanavajjhala, 2011]

- Bob or one of his 9 immediate family members may have contracted a highly contagious disease, in which case the entire family would have been infected. An adversary asks the query “how many people at Bob's family address have this disease?”
- What can be learned from an answer produced while satisfying ϵ -DP?
 - Answer: Adversary's belief change on Bob's disease status may change by something close to $e^{10\epsilon}$.
- Anything wrong here?

In A Sense, No

1. An adversary's belief about Bob's disease status may change by a factor of $e^{10\epsilon}$ due to data correlation. This is an example that DP cannot bound prior-to-posterior belief change against arbitrary external knowledge.
2. DP's guarantee about real-to-ideal bound remains valid.
3. Applying PDP, ϵ -DP is doing what it is supposed to do, but stay tuned

My Personal Data or Personal Data Under My Control?

- Consider the following variants of the Bob example.
- Case (a). Bob lives in a dorm building with 9 other unrelated individuals. Either they all have the disease or none. One can query how many individuals at this address have the disease.
- Case (b). The original example: Bob and 9 family members.
- Case (c). Bob and 9 minors for which Bob is the legal guardian.

What Constitutes An Individual's Personal Data?

- Is the genome of my parents, children, sibling, cousins “my personal information”?
- Example: DeCode Genetics, based in Reykjavík, says it has collected full DNA sequences on 10,000 individuals. And because people on the island are closely related, DeCode says it can now also extrapolate to accurately guess the DNA makeup of nearly all other 320,000 citizens of that country, including those who never participated in its studies.

Such legal and ethical questions still need to be resolved

- Evidences suggest that such privacy concerns will be recognized.
- In 2003, the supreme court of Iceland ruled that a daughter has the right to prohibit the transfer of her deceased father's health information to a Health Sector Database, not because her right acting as a substitute of her deceased father, but in the recognition that she might, on the basis of her right to protection of privacy, have an interest in preventing the transfer of health data concerning her father into the database, as information could be inferred from such data relating to the hereditary characteristics of her father which might also apply to herself.

https://epic.org/privacy/genetic/iceland_decision.pdf

Lesson

- When dealing with genomic and health data, one cannot simply say correlation doesn't matter because of Personal Data Principle, and may have to quantify and deal with such correlation.

Group Privacy as a Potential Challenge to Personal Data Principle

- Can a group of individuals, none of whom has specifically authorized usage of their personal information, together sue on privacy grounds that aggregate information about them is leaked?
 - If so, satisfying DP is not sufficient.
 - Would size of group matter?

A Moral Challenge to DP

Say I steal 2 cents from every bank account in America. I am proven guilty, but everyone I stole from says they're fine with it. What happens?

 Answer

 Follow · 115

 Request



 5



- If one makes profit from applying DP to a dataset of many individuals, isn't this morally the same as the above?

How to Choose ϵ

- From the inventors of DP: *“The choice of ϵ is essentially a social question. We tend to think of ϵ as, say, 0.01, 0.1, or in some cases, $\ln 2$ or $\ln 3$ ”.*
- Our position.
 - ϵ of between 0.1 and 1 is often acceptable
 - ϵ close to 5 might be applicable in rare cases, but needs careful analysis
 - ϵ above 10 means very little
- Why?

Consult This Table of Change in Belief: p is prior; numbers in table are posterior

ϵ	0.01	0.1	1	5	10
$\gamma = e^\epsilon$	1.01	1.11	2.72	148	22026
$p = 0.001$	0.0010	0.0011	0.0027	0.1484	1.0000
$p = 0.01$	0.0101	0.0111	0.0272	0.9933	1.0000
$p = 0.1$	0.1010	0.1105	0.2718	0.9939	1.0000
$p = 0.5$	0.5050	0.5476	0.8161	0.9966	1.0000
$p = 0.75$	0.7525	0.7738	0.9080	0.9983	1.0000
$p = 0.99$	0.9901	0.9910	0.9963	0.9999	1.0000

Apply a Model Learned with DP

- There are two steps in Big Data
 - Learning a model from data from individuals in A
 - Apply the model to individuals in B, using some (typically less sensitive) personal info of each individual, one can learn (typically more sensitive) personal info.
 - The sets A and B may overlap
- The notion of DP deals with only the first step.
- Even if a model is learned while satisfying DP, applying it may still result in privacy concern, because it uses each individual's personal info.

The Target Pregnancy Prediction Example

- Target assigns every customer a Guest ID number and stores a history of everything they've bought and any demographic information Target has collected from them or bought from other sources.
- Looking at historical buying data for all the ladies who had signed up for Target baby registries in the past, Target's algorithm was able to identify about 25 products that, when analyzed together, allowed Target to assign each shopper a "pregnancy prediction" score.
- Target could also estimate her due date to within a small window, so Target could send coupons timed to very specific stages of her pregnancy.

Using an ϵ -DP Definition is Good Enough When

1. For each dataset D , and each individual X whose data is in D , there exists an dataset D' such that
 - D and D' are neighboring
 - All data that belong to X in D have been removed or overwritten in D'
 - That is, we can say that D' is “an ideal world of privacy” for individual X
2. The privacy parameter ϵ is suitable for the setting

There are many applications of ϵ -DP where one cannot automatically assume that using ϵ -DP provides strong privacy protection.

Chapter 3 of Li et al.: Differential Privacy: From Theory to Practice. Morgan & Claypool Publishers 2016.

- Thank you!

- Questions?