Agenda for today class

- Last Class
  - Data Privacy
- Today's class
  - Data Privacy Algorithms

Example

- In March, Alice publishes an article based on the information in this database and writes that “the current freshman class at Private University is made up of 3,005 students, 202 of whom are from families earning over US$350,000 per year.”
- The following month, Bob publishes a separate article
- containing these statistics: “201 families in Private University’s freshman class of 3,004 have household incomes exceeding US$350,000 per year.”
- Neither Alice nor Bob is aware that they have both published similar information.

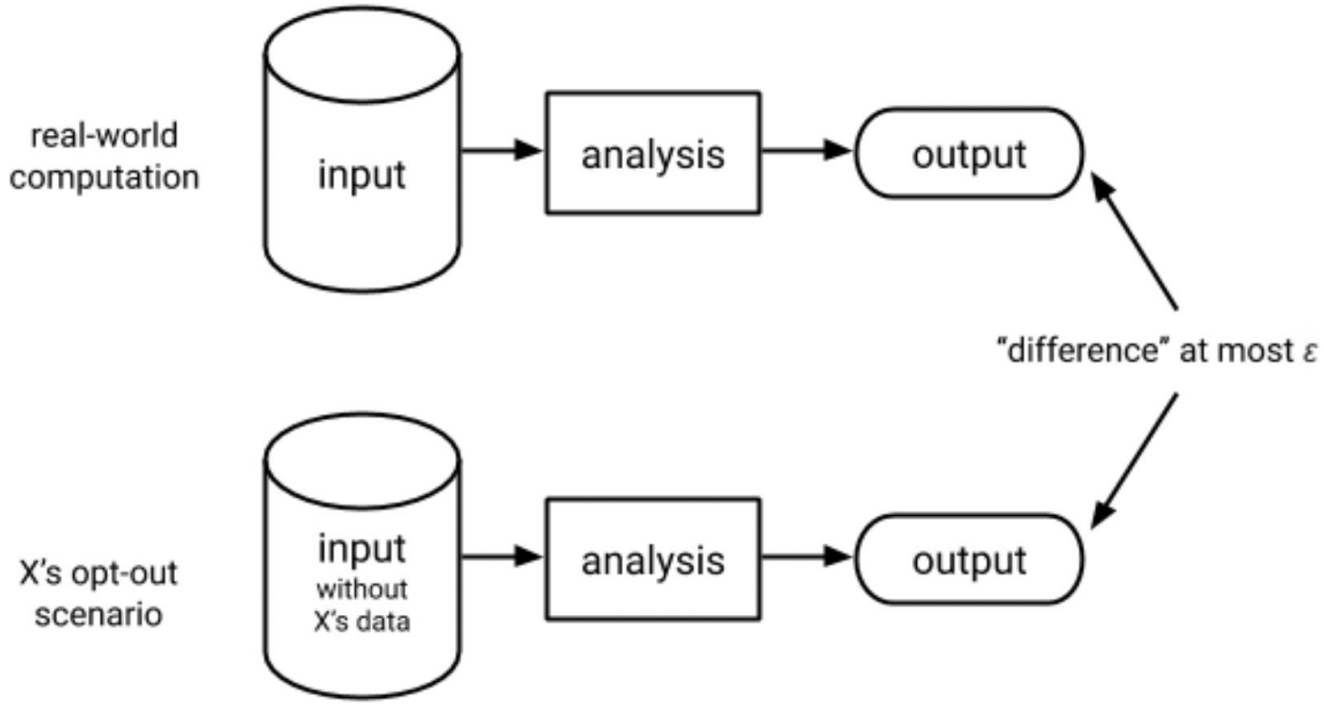
Example

- A clever student Eve reads both of these articles and makes an observation. From the published information, Eve concludes that between March and April one freshman withdrew from Private University and that the student's parents earn over US$350,000 per year.

- Eve asks around and is able to determine that a student named John dropped out around the end of March. Eve then informs her classmates that John's parents probably earn over US$350,000 per year
- John is upset – since his privacy is violated

John’s scenario

John’s scenario



Privacy

- First thought: anonymize the data
- How?
- Remove "personally identifying information" (PII)
  - Name, Social Security number, phone number, email, address… what else?
  - Anything that identifies the person directly
- Is this enough?

Re-identification by Linking

Microdata

ID	QID			SA
Name	Zipcode	Age	Sex	Disease
Alice	47677	29	F	Ovarian Cancer
Betty	47602	22	F	Ovarian Cancer
Charles	47678	27	M	Prostate Cancer
David	47905	43	M	Flu
Emily	47909	52	F	Heart Disease
Fred	47906	47	M	Heart Disease

Voter registration data

Name	Zipcode	Age	Sex
Alice	47677	29	F
Bob	47983	65	M
Carol	47677	22	F
Dan	47532	23	M
Ellen	46789	43	F

Latanya Sweeney's Attack (1997)

Massachusetts hospital discharge dataset

Medical Data Released as Anonymous

SSN	Name	...nicity	Date Of Birth	Sex	ZIP	Marital Status	Problem
			09/27/64	female	02139	divorced	hypertension
			09/30/64	female	02139	divorced	obesity
		asian	04/18/64	male	02139	married	chest pain
		asian	04/15/64	male	02139	married	obesity
		black	03/13/63	male	02138	married	hypertension
		black	03/18/63	male	02138	married	shortness of breath
		black	09/13/64	female	02141	married	shortness of breath
		black	09/07/64	female	02141	married	obesity
		white	05/14/61	male	02138	single	chest pain
		white	05/08/61	male	02138	single	obesity
		white	09/15/61	female	02142	widow	shortness of breath

Voter List

Name	Address	City	ZIP	DOB	Sex	Party	.....
.....	.....	.....	.....	.....	.....	.....	
.....	.....	.....	.....	.....	.....	.....	
Sue J. Carlson	1459 Main St.	Cambridge	*02142*	*9/15/61*	*female*	democrat	.....
.....	.....	.....	.....	.....	.....	.....	

Figure 1: ...e-identifying anonymous data by linking to external data

Public voter dataset

Quasi-Identifiers

- Key attributes
  - Name, address, phone number - uniquely identifying!
  - Always removed before release
- Quasi-identifiers half identifiers
  - (5-digit ZIP code, birth date, gender) uniquely identify 87% of the population in the U.S.
  - Can be used for linking anonymized dataset with other datasets

Classification of Attributes

- Sensitive attributes
  - Medical records, salaries, etc.
  - These attributes is what the researchers need, so they are always released directly

imp

Key Attribute	Quasi-identifier			Sensitive attribute
Name	DOB	Gender	Zipcode	Disease
Andre	1/21/76	Male	53715	Heart Disease
Beth	4/13/86	Female	53715	Hepatitis
Carol	2/28/76	Male	53703	Brochitis
Dan	1/21/76	Male	53703	Broken Arm
Ellen	4/13/86	Female	53706	Flu
Eric	2/28/76	Female	53706	Hang Nail

K-Anonymity: Intuition

write a bit

- The information for each person contained in the released table cannot be distinguished from at least k-1 individuals whose information also appears in the release
  - Example: you try to identify a man in the released table, but the only information you have is his birth date and gender. There are k men in the table with the same birth date and gender.
- Any quasi-identifier present in the released table must appear in at least k records

K-Anonymity Protection Model

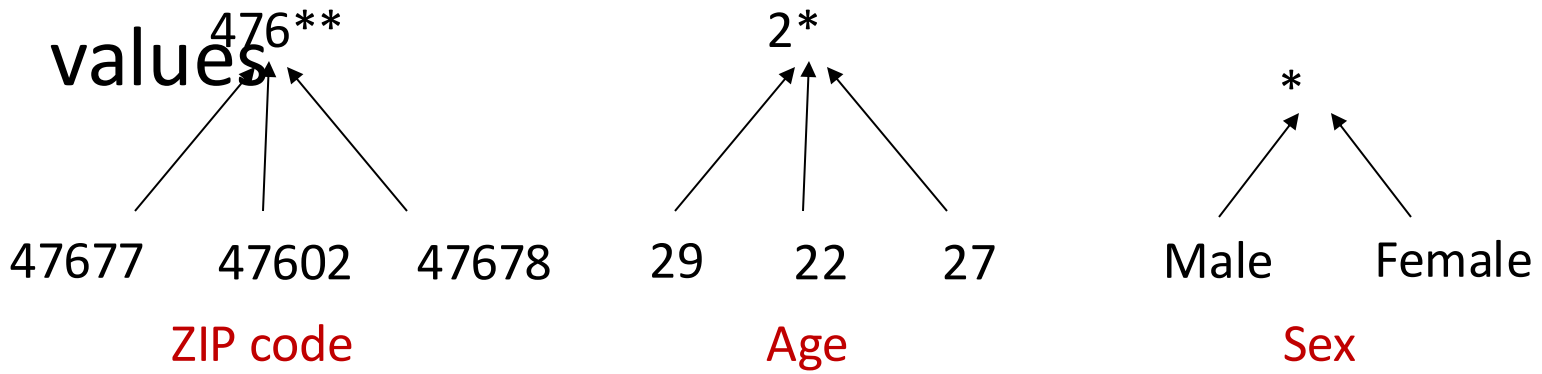
- Private table
- Released table: RT satisfies the k-anonymity
- Attributes: A_1, A_2, …, A_n
- Quasi-identifier subset: A_i, …, A_j

Let $RT(A_1,\ldots,A_n)$ be a table, $QI_{RT}=(A_i,\ldots,A_j)$ be the quasi-identifier associated with RT, $A_i,\ldots,A_j \subseteq A_1,\ldots,A_n$, and RT satisfy k-anonymity. Then, each sequence of values in $RT[A_x]$ appears with at least k occurrences in $RT[QI_{RT}]$ for $x=i,\ldots,j$.

write

Generalization

- Goal of k-Anonymity
  - Each record is indistinguishable from at least k-1 other records
  - These k records form an equivalence class
- Generalization: replace quasi-identifiers with less specific, but semantically consistent values

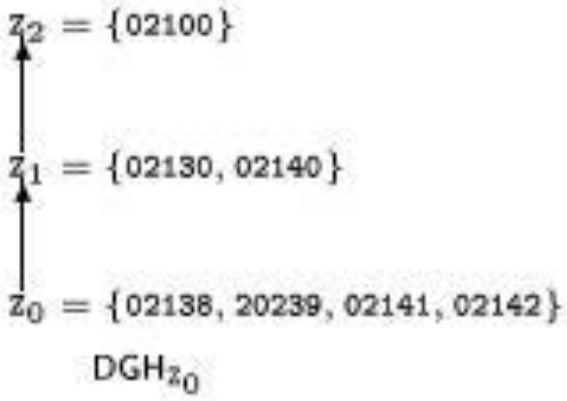


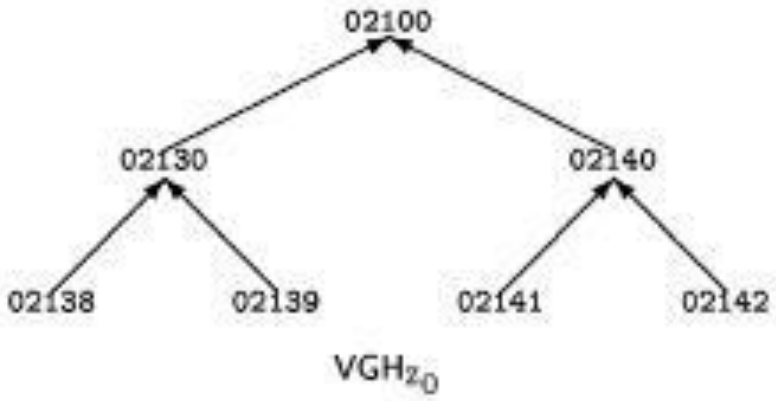
Achieving k-Anonymity

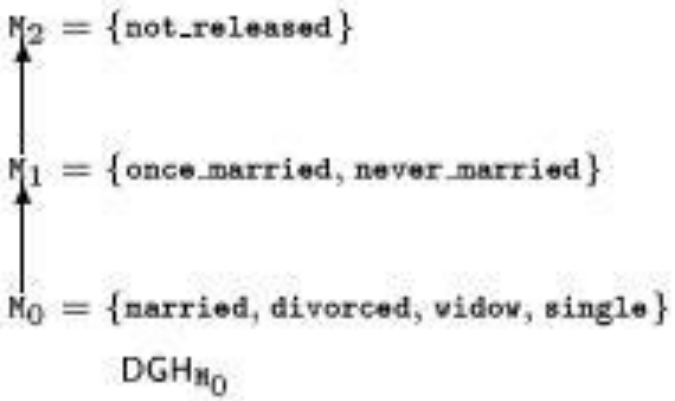
- Generalization
  - Replace specific quasi-identifiers with less specific values until get k identical values
  - Partition ordered-value domains into intervals
- Suppression
  - When generalization causes too much information loss
    - This is common with “outliers”
- Lots of algorithms in the literature
  - Aim to produce “useful” anonymizations

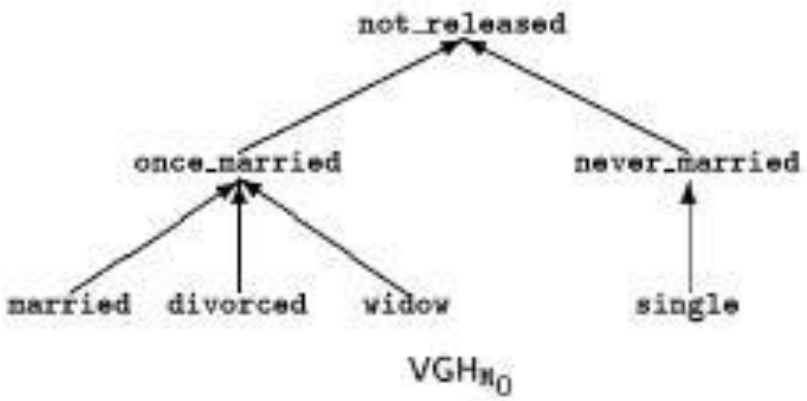
    … usually without any clear notion of utility

Generalization in Action









Example of a k-Anonymous Table

he can give to find K

each record should resemble with k-1 records

	Race	Birth	Gender	ZIP	Problem
t1	Black	1965	m	0214*	short breath
t2	Black	1965	m	0214*	chest pain
t3	Black	1965	f	0213*	hypertension
t4	Black	1965	f	0213*	hypertension
t5	Black	1964	f	0213*	obesity
t6	Black	1964	f	0213*	chest pain
t7	White	1964	m	0213*	chest pain
t8	White	1964	m	0213*	obesity
t9	White	1964	m	0213*	short breath
t10	White	1967	m	0213*	chest pain
t11	White	1967	m	0213*	chest pain

Figure 2 Example of *k*-anonymity, where *k*=2 and QI={*Race*, *Birth*, *Gender*, *ZIP*}

Example of Generalization (1)

Released table

	Race	Birth	Gender	ZIP	Problem
t1	Black	1965	m	0214*	short breath
t2	Black	1965	m	0214*	chest pain
t3	Black	1965	f	0213*	hypertension
t4	Black	1965	f	0213*	hypertension
t5	Black	1964	f	0213*	obesity
t6	Black	1964	f	0213*	chest pain
t7	White	1964	m	0213*	chest pain
t8	White	1964	m	0213*	obesity
t9	White	1964	m	0213*	short breath
t10	White	1967	m	0213*	chest pain
t11	White	1967	m	0213*	chest pain

External data

Name	Birth	Gender	ZIP	Race
Andre	1964	m	02135	White
Beth	1964	f	55410	Black
Carol	1964	f	90210	White
Dan	1967	m	02174	White
Ellen	1968	f	02237	White

By linking these 2 tables, you still don't learn Andre's problem

Example of Generalization (2)

Microdata

QID			SA
Zipcode	Age	Sex	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

Generalized table

QID			SA
Zipcode	Age	Sex	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

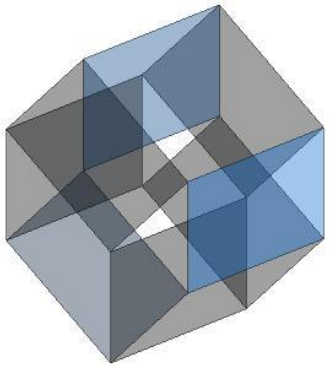
!!

- Released table is 3-anonymous
- If the adversary knows Alice's quasi-identifier (47677, 29, F), he still does not know which of the first 3 records corresponds to Alice's record

Curse of Dimensionality

[Aggarwal VLDB '05]

- Generalization fundamentally relies on spatial locality
  - Each record must have k close neighbors



- Real-world datasets are very sparse
  - Many attributes (dimensions)
    - Netflix Prize dataset: 17,000 dimensions
    - Amazon customer records: several million dimensions
  - "Nearest neighbor" is very far imp
- Projection to low dimensions loses all info $\Rightarrow$ k-anonymized datasets are useless

Two (and a Half) Interpretations

- Membership disclosure: Attacker cannot tell that a given person in the dataset
- Sensitive attribute disclosure: Attacker cannot tell that a given person has a certain sensitive attribute
- Identity disclosure: Attacker cannot tell which

This interpretation is correct, assuming the attacker does not know anything other than quasi-identifiers

But this does not imply any privacy!

Example: k clinical records, all HIV+

names of attacks

Unsorted Matching Attack

- Problem: records appear in the same order in the released table as in the original table
- Solution: randomize order before releasing

Race	ZIP
Asian	02138
Asian	02139
Asian	02141
Asian	02142
Black	02138
Black	02139
Black	02141
Black	02142
White	02138
White	02139
White	02141
White	02142

PT

Race	ZIP
Person	02138
Person	02139
Person	02141
Person	02142
Person	02138
Person	02139
Person	02141
Person	02142
Person	02138
Person	02139
Person	02141
Person	02142

GT1

Race	ZIP
Asian	02130
Asian	02130
Asian	02140
Asian	02140
Black	02130
Black	02130
Black	02140
Black	02140
White	02130
White	02130
White	02140
White	02140

GT2

Figure 3 Examples of *k*-anonymity tables based on PT

Complementary Release Attack

- Different releases of the same private table can be linked together to compromise k-

Race	BirthDate	Gender	ZIP	Problem
black	1965	male	02141	short of breath
black	1965	male	02141	chest pain
person	1965	female	0213*	painful eye
person	1965	female	0213*	wheezing
black	1964	female	02138	obesity
black	1964	female	02138	chest pain
white	1964	male	0213*	short of breath
person	1965	female	0213*	hypertension
white	1964	male	0213*	obesity
white	1964	male	0213*	fever
white	1967	male	02138	vomiting
white	1967	male	02138	back pain

GT1

Race	BirthDate	Gender	ZIP	Problem
black	1965	male	02141	short of breath
black	1965	male	02141	chest pain
black	1965	female	02138	painful eye
black	1965	female	02138	wheezing
black	1964	female	02138	obesity
black	1964	female	02138	chest pain
white	1960-69	male	02138	short of breath
white	1960-69	human	02139	hypertension
white	1960-69	human	02139	obesity
white	1960-69	human	02139	fever
white	1960-69	male	02138	vomiting
white	1960-69	male	02138	back pain

GT3

Linking Independent Releases

Race	BirthDate	Gender	ZIP	Problem
black	9/20/1965	male	02141	short of breath
black	2/14/1965	male	02141	chest pain
black	10/23/1965	female	02138	painful eye
black	8/24/1965	female	02138	wheezing
black	11/7/1964	female	02138	obesity
black	12/1/1964	female	02138	chest pain
white	10/23/1964	male	02138	short of breath
white	3/15/1965	female	02139	hypertension
white	8/13/1964	male	02139	obesity
white	5/5/1964	male	02139	fever
white	2/13/1967	male	02138	vomiting
white	3/21/1967	male	02138	back pain

PT

Race	BirthDate	Gender	ZIP	Problem
black	1965	male	02141	short of breath
black	1965	male	02141	chest pain
black	1965	female	02138	painful eye
black	1965	female	02138	wheezing
black	1964	female	02138	obesity
black	1964	female	02138	chest pain
white	1964	male	02138	short of breath
white	1965	female	02139	hypertension
white	1964	male	02139	obesity
white	1964	male	02139	fever
white	1967	male	02138	vomiting
white	1967	male	02138	back pain

LT

Attacks on k-Anonymity

- k-Anonymity does not provide privacy if
  - Sensitive values in an equivalence class lack diversity
  - The attacker has background knowledge

Homogeneity attack

Bob	
Zipcode	***Age***
47678	27

Background knowledge attack

Carl	
Zipcode	***Age***
47673	36

Zipcode	Age	Disease
476**	2*	Heart Disease
476**	2*	Heart Disease
476**	2*	Heart Disease
4790*	≥40	Flu
4790*	≥40	Heart Disease
4790*	≥40	Cancer
476**	3*	Heart Disease
476**	3*	Cancer
476**	3*	Cancer

k-Anonymity

write

- Each released record should be indistinguishable from at least (k-1) others on its QI attributes
- Alternatively: cardinality of any query result on released data should be at least k
- k-anonymity is (the first) one of many privacy definitions in this line of work
  - l-diversity, t-closeness, m-invariance, delta-presence...

- In this lecture, we will discuss additional privacy definitions that tries to address the limitations of k-anonymity
  - L-diversity
  - T-closeness

- Complementary Release Attack
  - Different releases can be linked together to compromise k-anonymity.
  - Solution:
    - Consider all of the released tables before release the new one, and try to avoid linking.
    - Other data holders may release some data that can be used in this kind of attack. Generally, this kind of attack is hard to be prohibited completely.

Impor: drawbacks with example

- **k-Anonymity does not provide privacy if:**
  - **Sensitive values in an equivalence class lack diversity**
  - **The attacker has background knowledge**

Homogeneity Attack

Bob	
Zipcode	***Age***
47678	27

A 3-anonymous patient table

Zipcode	Age	Disease
476**	2*	Heart Disease
476**	2*	Heart Disease
476**	2*	Heart Disease
4790*	≥40	Flu
4790*	≥40	Heart Disease
4790*	≥40	Cancer
476**	3*	Heart Disease
476**	3*	Cancer
476**	3*	Cancer

Background Knowledge Attack

Umeko (Japanese)	
Zipcode	***Age***
47673	36

- Easy to understand.
- Should prevent background knowledge attacks.
- Should be easily enforceble.

imp write

- L-diversity principle: A q-block is l-diverse if contains at least l "well represented" values for the sensitive attribute S. A table is l-diverse if every q-block is l-diverse

- Distinct *l*-diversity
  - Each equivalence class has at least *l* well-represented sensitive values
  - Limitation:
    - Doesn't prevent the probabilistic inference attacks
    - Ex.

      In one equivalent class, there are ten tuples. In the "Disease" area, one of them is "Cancer", one is "Heart Disease" and the remaining eight are "Flu". This satisfies 3-diversity, but the attacker can still affirm that the target person's disease is "Flu" with the accuracy of 80%.

write

l-diversity may be difficult and unnecessary to achieve.

- **A single sensitive attribute**
  - **Two values: HIV positive (1%) and HIV negative (99%)**
  - **Very different degrees of sensitivity**
- **l-diversity is unnecessary to achieve**

imp

  - **2-diversity is unnecessary for an equivalence class that contains only negative records**
- **l-diversity is difficult to achieve**
  - **Suppose there are 10000 records in total**
  - **To have distinct 2-diversity, there can be at most 10000*1%=100 equivalence classes**

what it means?

write with eg

l-diversity is insufficient to prevent attribute disclosure.

Similarity Attack

Bob	
Zip	***Age***
47678	27

A 3-diverse patient table

Zipcode	Age	Salary	Disease
476**	2*	20K	Gastric Ulcer
476**	2*	30K	Gastritis
476**	2*	40K	Stomach Cancer
4790*	≥40	50K	Gastritis
4790*	≥40	100K	Flu
4790*	≥40	70K	Bronchitis
476**	3*	60K	Bronchitis
476**	3*	80K	Pneumonia
476**	3*	90K	Stomach Cancer

Conclusion

1. **Bob's salary is in [20k,40k], which is relative low.**
2. **Bob has some stomach-related disease.**

l-diversity does not consider semantic meanings of sensitive values

l-Diversity

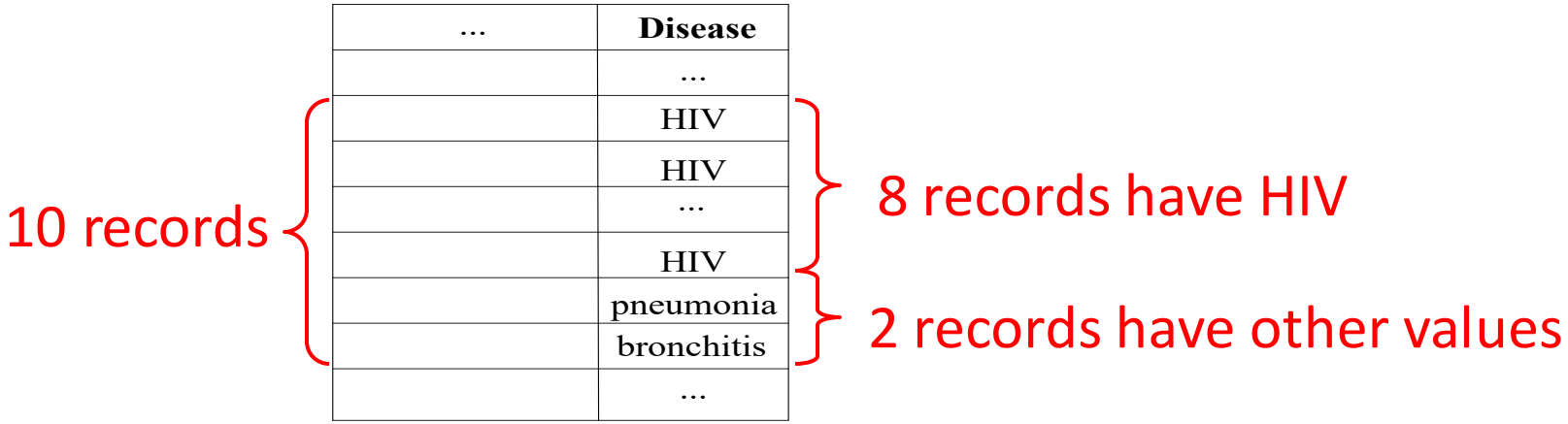
[Machanavajjhala et al. ICDE '06]

Caucas	787XX	Flu
Caucas	787XX	Shingles
Caucas	787XX	Acne
Caucas	787XX	Flu
Caucas	787XX	Acne
Caucas	787XX	Flu
Asian/AfrAm	78XXX	Flu
Asian/AfrAm	78XXX	Flu
Asian/AfrAm	78XXX	Acne
Asian/AfrAm	78XXX	Shingles
Asian/AfrAm	78XXX	Acne
Asian/AfrAm	78XXX	Flu

Sensitive attributes must be "diverse" within each quasi-identifier equivalence class

Distinct l-Diversity

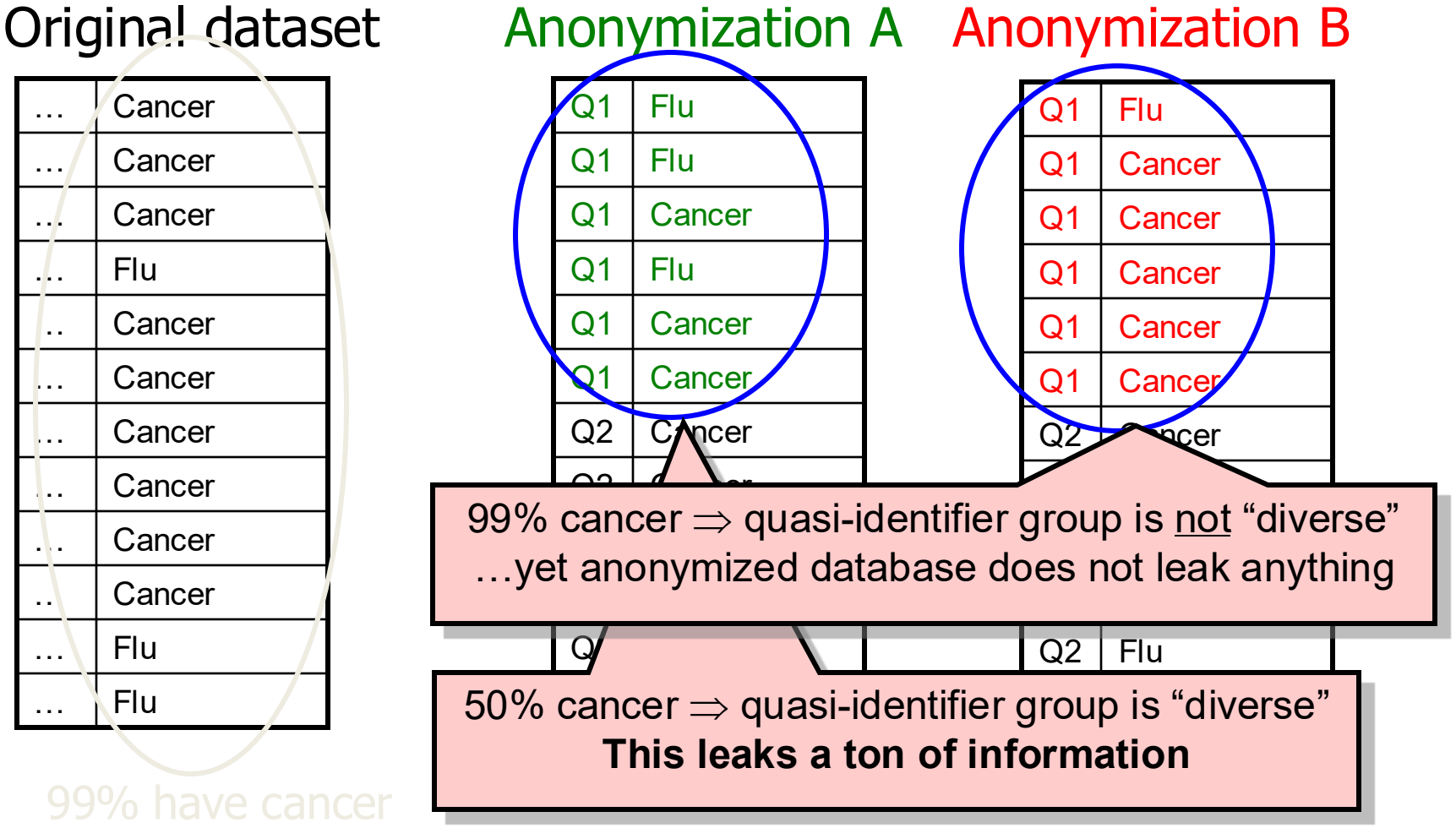
- Each equivalence class has at least l well-represented sensitive values
- Doesn't prevent probabilistic inference attacks



Other Versions of l-Diversity

- Probabilistic l-diversity
  - The frequency of the most frequent value in an equivalence class is bounded by 1/l
- Entropy l-diversity
  - The entropy of the distribution of sensitive values in each equivalence class is at least log(l)
- Recursive (c,l)-diversity
  - $r_1 < c(r_l + r_{l+1} + \ldots + r_m)$ where r_i is the frequency of the i^{th} most frequent value
  - Intuition: the most frequent value does not appear too frequently

Neither Necessary, Nor Sufficient



Skewness Attack

- Example: sensitive attribute is HIV+ (1%) or HIV- (99%)
- Consider an equivalence class that contains an equal number of HIV+ and HIV- records imp
  - Diverse, but potentially violates privacy!
- l-diversity does not differentiate:
  - Equivalence class 1: 49 HIV+ and 1 HIV-
  - Equivalence class 2: 1 HIV+ and 49 HIV-

l-diversity does not consider overall distribution of sensitive values!

Sensitive Attribute Disclosure

Similarity attack

Bob	
Zip	***Age***
47678	27

A 3-diverse patient table

Zipcode	Age	Salary	Disease
476**	2*	20K	Gastric Ulcer
476**	2*	30K	Gastritis
476**	2*	40K	Stomach Cancer
4790*	≥40	50K	Gastritis
4790*	≥40	100K	Flu
4790*	≥40	70K	Bronchitis
476**	3*	60K	Bronchitis
476**	3*	80K	Pneumonia
476**	3*	90K	Stomach Cancer

Conclusion

1. Bob's salary is in [20k,40k], which is relatively low
2. Bob has some stomach-related disease

l-diversity does not consider semantics of sensitive values!

t-Closeness

[Li et al. ICDE '07]

Caucas	787XX	Flu
Caucas	787XX	Shingles
Caucas	787XX	Acne
Caucas	787XX	Flu
Caucas	787XX	Acne
Caucas	787XX	Flu
Asian/AfrAm	78XXX	Flu
Asian/AfrAm	78XXX	Flu
Asian/AfrAm	78XXX	Acne
Asian/AfrAm	78XXX	Shingles
Asian/AfrAm	78XXX	Acne
Asian/AfrAm	78XXX	Flu

imp

Distribution of sensitive attributes within each quasi-identifier group should be "close" to their distribution in the entire original database

Trick question: Why publish quasi-identifiers at all??

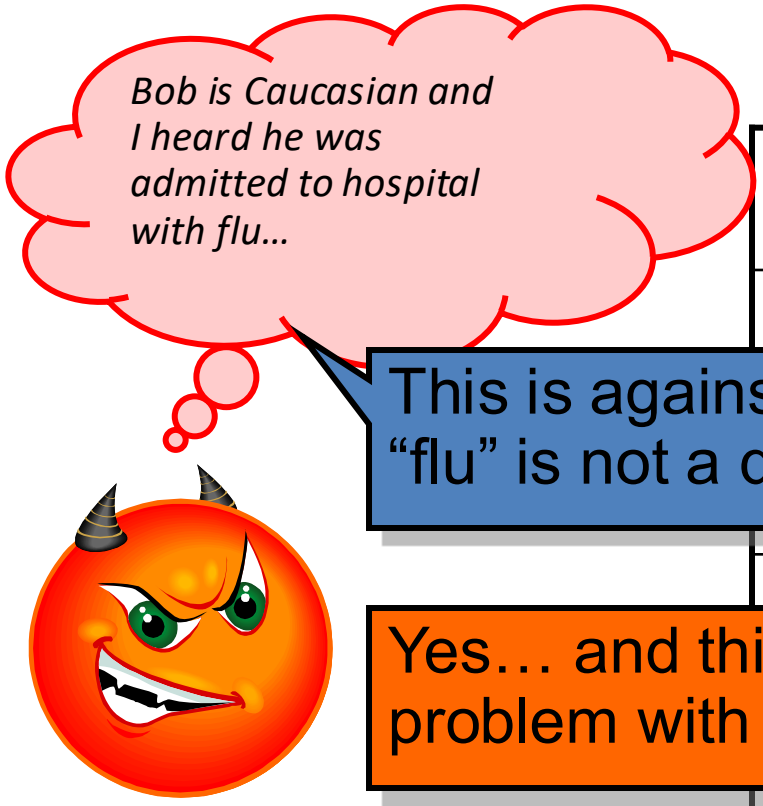
Anonymous, “t-Close” Dataset

Caucas	787XX	HIV+	Flu
Asian/AfrAm	787XX	HIV-	Flu
Asian/AfrAm	787XX	HIV+	Shingles
Caucas	787XX	HIV-	Acne
Caucas	787XX	HIV-	Shingles
Caucas	787XX	HIV-	Acne

This is k-anonymous, l-diverse and t-close…

…so secure, right?

What Does Attacker Know?



Caucas	787XX	HIV+	Flu
Asian/AfrAm	787XX	HIV-	Flu
		IV+	Shingles
		HIV-	Acne
		HIV-	Shingles
Caucas	787XX	HIV-	Acne

This is against the rules!
"flu" is not a quasi-identifier

Yes… and this is yet another problem with k-anonymity

AOL Privacy Debacle

- In August 2006, AOL released anonymized search query logs
  - 657K users, 20M queries over 3 months (March-May)
- Opposing goals
  - Analyze data for research purposes, provide better services for users and advertisers
  - Protect privacy of AOL users
    - Government laws and regulations
    - Search queries may reveal income, evaluations, intentions to acquire goods and services, etc.

AOL User 4417749

- AOL query logs have the form

  <AnonID, Query, QueryTime, ItemRank, ClickURL>

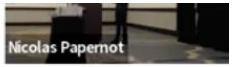
  - ClickURL is the truncated URL
- NY Times re-identified AnonID 4417749
  - Sample queries: “numb fingers”, “60 single men”, “dog that urinates on everything”, “landscapers in Lilburn, GA”, several people with the last name Arnold
    - Lilburn area has only 14 citizens with the last name Arnold
  - NYT contacts the 14 citizens, finds out AOL User

k-Anonymity Considered Harmful

- Syntactic
  - Focuses on data transformation, not on what can be learned from the anonymized dataset
  - "k-anonymous" dataset can leak sensitive information
- "Quasi-identifier" fallacy
  - Assumes a priori that attacker will not know certain information about his target
- Relies on locality
  - Destroys utility of many real-world datasets

Definition

- Differential Privacy



- Setup: n datapoints $X = X_1, \dots, X_n$, given to a "trusted curator"
- The curator has an algorithm $M : \mathcal{X}^n \to \mathcal{Y}$, outputs $M(X)$

imp

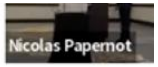
Definition: An algorithm M is (ε, δ)-differentially private (DP) if for all datasets X and X' which differ in one entry ("neighbouring"), and for all events $S \subseteq \mathcal{Y}$,

$$\Pr[M(X) \in S] \leq e^{\varepsilon} \Pr[M(X') \in S] + \delta.$$

"An algorithm is differentially private if its distribution over outputs doesn't change much after adding/removing one point."

DP Definition

Comments on Differential Privacy



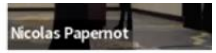
Definition: An algorithm M is (ε, δ)-differentially private (DP) if for all datasets X and X' which differ in one entry ("neighbouring"), and for all events $S \subseteq \mathcal{Y}$,

$$\Pr[M(X) \in S] \leq e^{\varepsilon} \Pr[M(X') \in S] + \delta.$$

- Bounds the multiplicative increase in probability of any event
  - With small additive change
- Quantitative in ε, δ, smaller = more private
- $e^{\varepsilon} \approx 1 + \varepsilon$ (for small ε): multiplicative increase in any probability
  - $\varepsilon \approx 1$ is reasonable, double-digit ε should make you suspicious
- δ: probability of (potential) "total privacy failure"

DP Definition

What *doesn't* DP do?



- Important: does **not** prevent inferences (statistics/machine learning)
  - (Public) smoker participates in (differentially private) study investigating whether smoking causes cancer
  - Reveals that smoking causes cancer! Smoker's insurance premiums increase!
  - Was their (differential) privacy violated?
  - No: smoking → cancer could be inferred whether or not they participated
  - Differential privacy: outcome of algorithm is similar, whether or not someone participates
- Not appropriate when individual identities are important
  - "Private" contact tracing

Data Privacy

- "Imagine you are in a class with 10 students total. You've just taken an exam, and are eagerly awaiting the results. One day, the professor walks into the room and writes a number on the board: 85%. They announce that this was the average grade on the exam, and that they will be passing back papers shortly.

- Suddenly, you get a text from your good friend Ari: "Big news! Just got a job with the circus and had to drop the class. See you this summer!" You mention to the professor that your friend is no longer enrolled in the class. With this information, they open their computer, type for a moment, then walk over to the board, erase the old average, and write 87%.

Differential Data Privacy

He begins to explain that you know how to calculate an average in general. Suppose each student's score is denoted by x_i, the set of all such scores is X, and the size of this set is n. Then

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

And you know there were 10 people in the class before your friend dropped it:

$$\bar{x}_{\text{before}} = \frac{1}{10}\sum_{i=1}^{10} x_i = 85$$

Differential Data Privacy

where each x_i is an exam score. Further, you know there were 9 people after he left, and that the average score was 84:

$$\bar{x}_{\text{after}} = \frac{1}{9}\sum_{i=1}^{9} x_i = 87$$

Subtract the two equations from each other:

$$\bar{x}_{\text{before}} - \bar{x}_{\text{after}} = \frac{1}{10}\sum_{i=1}^{10} x_i - \frac{1}{9}\sum_{i=1}^{9} x_i = -2$$

Differential Data Privacy

Simplify the subtraction term and using the definition of $\bar{x}_{\text{after}}$:

$$-\frac{1}{90} \cdot 9 \cdot \bar{x}_{\text{after}} + \frac{1}{10} x_{10} = -2$$

and simplify the fraction to its lowest common denominator:

$$-\frac{1}{10} \cdot \bar{x}_{\text{after}} + \frac{1}{10} x_{10} = -2$$

Now you just need to isolate x_{10}, your friend's exam score:

$$x_{10} = 10 \cdot -2 + \frac{1}{10} \bar{x}_{\text{after}} = -20 + \bar{x}_{\text{after}}$$

You already know $\bar{x}_{\text{after}}$, it is simply the average written on the board:

$$x_{10} = -20 + 87 = 67$$

How could this be prevented

- **Randomized response**

- Randomized response is a method that modifies the value of each item in a data set according to certain probabilistic rules. For each item, if a “coin flip” returns tails (false), then the value recorded may not be the true value. This algorithm originated in the social sciences, where the goal was to prevent embarrassment to survey participants who were answering sensitive questions about their health or behavior.

How could this be prevented

- **Adding noise**

- Alternatively, what if the professor samples a value from a distribution and adds it to the mean? For example, if the professor tells you that the mean has noise added to it, then reconstructing your friend's score becomes impossible, since the equation has two unknowns: the exam score, and the amount of noise added. In this case, the means from the previous section become:

$$\bar{x}_{\text{before}} = \frac{1}{10}\sum_{i=1}^{10} x_i + N_1 = 85$$

$$\bar{x}_{\text{after}} = \frac{1}{9}\sum_{i=1} x_i + N_2 = 87$$

How could this be prevented

- where $N1$ and $N2$ are values drawn from some distribution and not disclosed to the students. With this noise added, attempting to calculate your friend's grade yields the following equation:

$$| \bar{x}_{\text{before}} - \bar{x}_{\text{after}} | = \frac{1}{10} \sum_{i=1}^{10} x_i + N_1 - \frac{1}{9} \sum_{i=1}^{9} x_i + N_2 = |1 + N_1 - N_2|$$

How could this be prevented

- You now have three unknowns: $x10$,$N1$, and $N2$, so the equation can no longer be solved analytically. At best, if you know the distribution that $N1$ and $N2$ were drawn from, then you can estimate the probabilities of different scores, without knowing with perfect information the true score. The question becomes: how do you add noise in such a way that you strike a balance between protecting privacy and keeping the statistic as useful as possible? If you know that your friend is the only person who is leaving, you can choose the noise from a known distribution in such a way that you sufficiently protect his score.

Agenda for today class

- Last Class
  - Data Privacy
- Today's class
  - Data Privacy Algorithms

K-Anonymity

- K Anonymity works on the principle that if you combine data with similar attributes, you can obscure identifying information about any individual contributing to that data. It's basically the ability to disappear in a crowd – since a sensitive data attribute masked using K Anonymity could actually correspond to any single individual in the pooled dataset.

- A dataset is k-anonymous if, for each group of records with the same quasi-identifiers (attributes that can potentially identify an individual), there are at least k records in that group.

L-diversity

- L-diversity is a privacy enhancement for datasets that goes beyond k-anonymity by ensuring sufficient variation in sensitive attributes within equivalence classes.

- It means that for each group of records with the same quasi-identifiers, there should be at least l distinct values for the sensitive attribute. T

- This helps prevent inferences about an individual's sensitive information even if there's k-anonymity.

The L-Diversity Data Anonymization Model: Extending K Anonymity

imp

L-Diversity reduces the risk of re-identification of sensitive data by ensuring that individual records in a dataset are not too similar to each another.

What is L-Diversity?

An enhancement to the K Anonymity data masking model, the L-Diversity extension was developed to reduce the granularity of data representation in a dataset.

The L Diversity Data Anonymization Model: Extending K Anonymity

How?

K-Anonymity leverages generalization, suppression ת‌and other techniques that enable mapping of each specific record onto a minimum of “K minus 1” other records in the dataset. Yet protecting identities down the K individual level cannot always protect the sensitive values which were masked, especially when these values are homogenic within the dataset. To solve this, L Diversity promotes intra-group diversity of sensitive values as one of the key data masking best practices.

imp

L-diversity

- **l-diversity:**

Beyond k-anonymity, l-diversity ensures that within each group of records with the same quasi-identifiers, there are at least l distinct values for the sensitive attribute. This adds another layer of protection by ensuring there's variation in the sensitive data even within these groups.

- **Why it matters:**

If all records within a group have the same sensitive value, even with k-anonymity, an attacker might still be able to infer the sensitive information by identifying the group. L-diversity helps to prevent this by ensuring enough variation within the groups.

L-diversity

- l-diversity can be applied in addition to k-anonymity if there is a risk that too much homogeneity in a sensitive attribute's values, in combination with other quasi-identifying attributes, might lead to loss of privacy.

- For example, suppose that all women in the age group 40-45 and living in a particular district fall within the same income bracket. If you live in that district and you have a female neighbor who is 44, then you can deduce what she earns. The sensitive information has been leaked.

imp

- *l-diversity is considered as an addition to k-anonymity. Conversely, k-anonymity can be seen as a special case of l-diversity where l=1.*

L-diversity

- Using the l-diversity parameter, you can reduce the risk of identification by specifying that a sensitive attribute must have a minimum number of distinct values within each equivalence class. An equivalence class is a set of identical quasi-identifying attributes resulting from k-anonymity.

imp

The following data set contains identifying, quasi-identifying, and sensitive information:

L-diversity

The following data set contains identifying, quasi-identifying, and sensitive information:

The following data set contains identifying, quasi-identifying, and sensitive information:

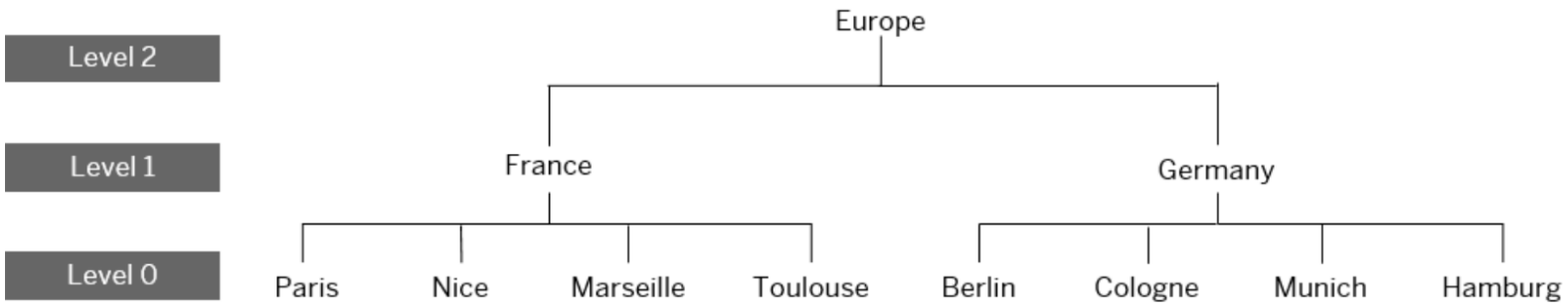
Row ID	Name	Gender	Location	Age
1	Peter	m	Berlin	30
2	Sigrid	f	Cologne	31
3	François	m	Paris	24
4	Bernhard	m	Munich	31
5	Pierre	m	Nice	25
6	Andrea	f	Hamburg	32
7	Juliette	f	Marseille	28
8	Fabienne	f	Toulouse	28

Identifier | Quasi-Identifiers | Sensitive data

L-diversity

The following data set contains identifying, quasi-identifying, and sensitive information:

A hierarchy is defined for the location attribute as follows:



L-diversity

If *k*=2, the result of data anonymization for the different equivalence classes looks like this:

imp

Equivalence classes

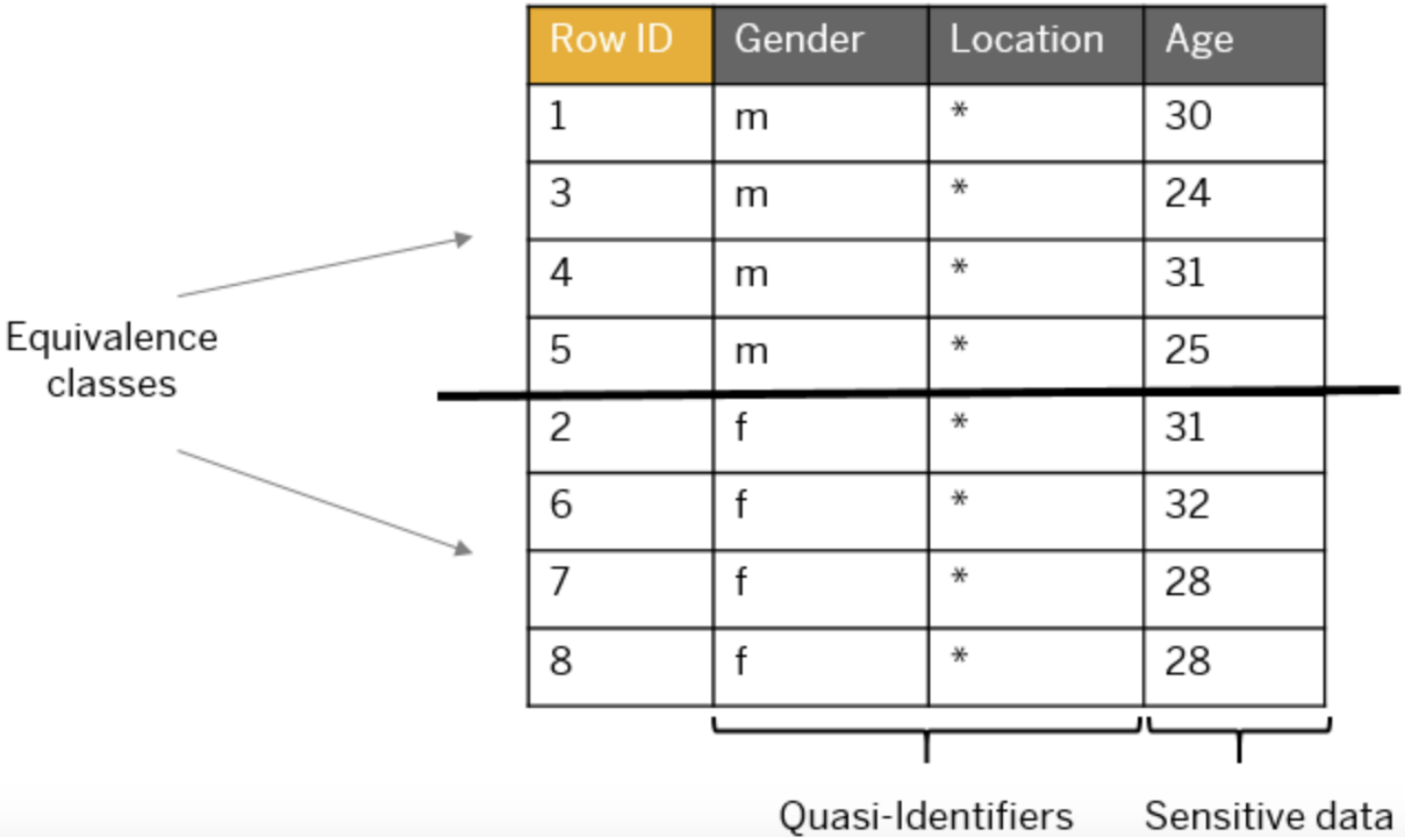
Row ID	Gender	Location	Age
1	m	Germany	30
4	m	Germany	31
2	f	Germany	31
6	f	Germany	32
3	m	France	24
5	m	France	25
7	f	France	28
8	f	France	28

Quasi-Identifiers

Sensitive data

L-diversity

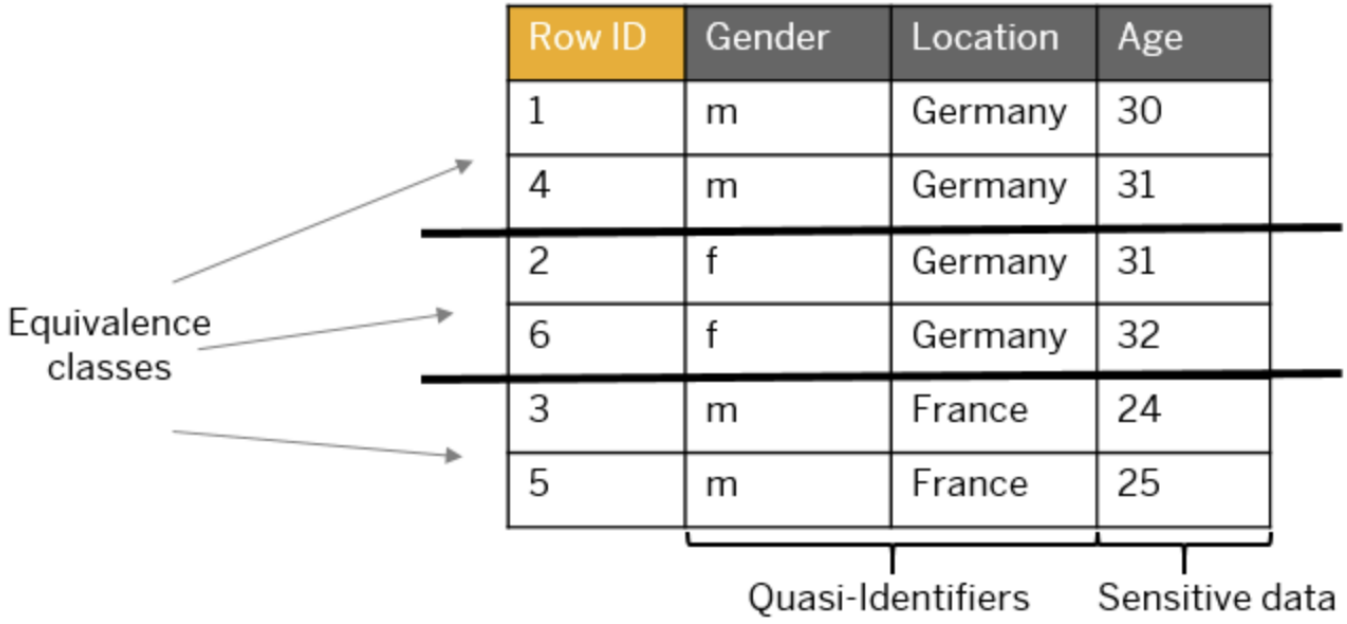
The following data set contains identifying, quasi-identifying, and sensitive information:



L-diversity

The following data set contains identifying, quasi-identifying, and sensitive information:

If a `loss` parameter is added to the definition, that is, k=2, l=2, `loss`=0.5, the following happens:



Row ID	Gender	Location	Age
1	m	Germany	30
4	m	Germany	31
2	f	Germany	31
6	f	Germany	32
3	m	France	24
5	m	France	25

T-Closeness

- **T-Closeness**

imp

T-closeness is a privacy technique that extends k-anonymity and l-diversity by ensuring that the distribution of a sensitive attribute within a group of records is close to the overall distribution of that attribute in the entire dataset. This closeness is measured using a threshold "t", which defines the maximum allowable distance between the two distributions

Relationship to K-anonymity and L-diversity:

- K-anonymity ensures that each record is indistinguishable from at least k - 1 other records within the dataset.

- L-diversity extends k-anonymity by requiring that each equivalence class has at least l distinct values for each sensitive attribute.

- T-closeness goes further than l-diversity by considering the distribution of sensitive attributes, not just the number of distinct values.

Benefits - T-Closeness

- T-closeness provides stronger privacy guarantees than k-anonymity and l-diversity by preventing the leakage of information about the distribution of sensitive attributes within groups of records.

Benefits - T-Closeness

- Imagine a dataset with age, sex, and income as attributes. If "t" is set to 0.1, then for each group of people with the same age and sex, the distribution of their incomes must be within 10% of the overall income distribution in the entire dataset. This prevents someone from knowing the income of a specific person in the group just by knowing their age and sex.

imp

Differential Privacy

- Differential privacy anonymizes data by randomizing sensitive information but in a way that regardless of whether an individual record is included in the data set or not, the outcome of statistical queries remains approximately the same. Differential privacy provides formal statistical privacy guarantees.

- The differentially private approach to anonymizing data is typically applied to numerical data in statistical databases. It works by adding noise to the sensitive values to protect privacy, while maximizing the accuracy of queries.

Differential Privacy

- What questions do you want the data to answer?
  - Knowing which queries will be executed on the data determines which columns to include in the data set. For example, to average salaries grouped by gender, region, start year and level, the data table to be anonymized could look like the table below. Direct identifiers and any other unrelated columns are removed.

How could this be prevented

- where $N1$ and $N2$ are values drawn from some distribution and not disclosed to the students. With this noise added, attempting to calculate your friend's grade yields the following equation:

$$\left| \bar{x}_{\text{before}} - \bar{x}_{\text{after}} \right| = \frac{1}{10} \sum_{i=1}^{10} x_i + N_1 - \frac{1}{9} \sum_{i=1}^{9} x_i + N_2 = \left| 1 + N_1 - N_2 \right|$$

Differential Privacy

Gender	Region	Hire Year	Level	Salary
m	APJ	1998	L2	20000
f	EMEA	1990	L5	50000
m	NA	2016	L2	20000
f	NA	2005	L3	30000

Sensitive data

Agenda for today class

- Today's class
  - Project Ideas
  - PyDeequ

Differential Data Privacy

- "Differential privacy makes it possible for tech companies to collect and share aggregate information about user habits, while maintaining the privacy of individual users."

- Differential privacy mathematically guarantees that anyone seeing the result of a differentially private analysis will essentially make the same inference about any individual's private information, whether or not that individual's private information is included in the input to the analysis.

imp

- **Applications**
  - We were confronting many big data breaches that necessitate governments, organizations, and companies to give a reconsideration of privacy.

DP Example

- The students are the individuals in the data set.
- Suppose the professor has the student scores in a data set:

- exam_scores ---------- 85 84 85 89 92 95 100 83 70 67

- The professor then queries the student database using a mean function to get the mean test score for the class:

- exam_scores.mean() >>> 85.0

DP Example

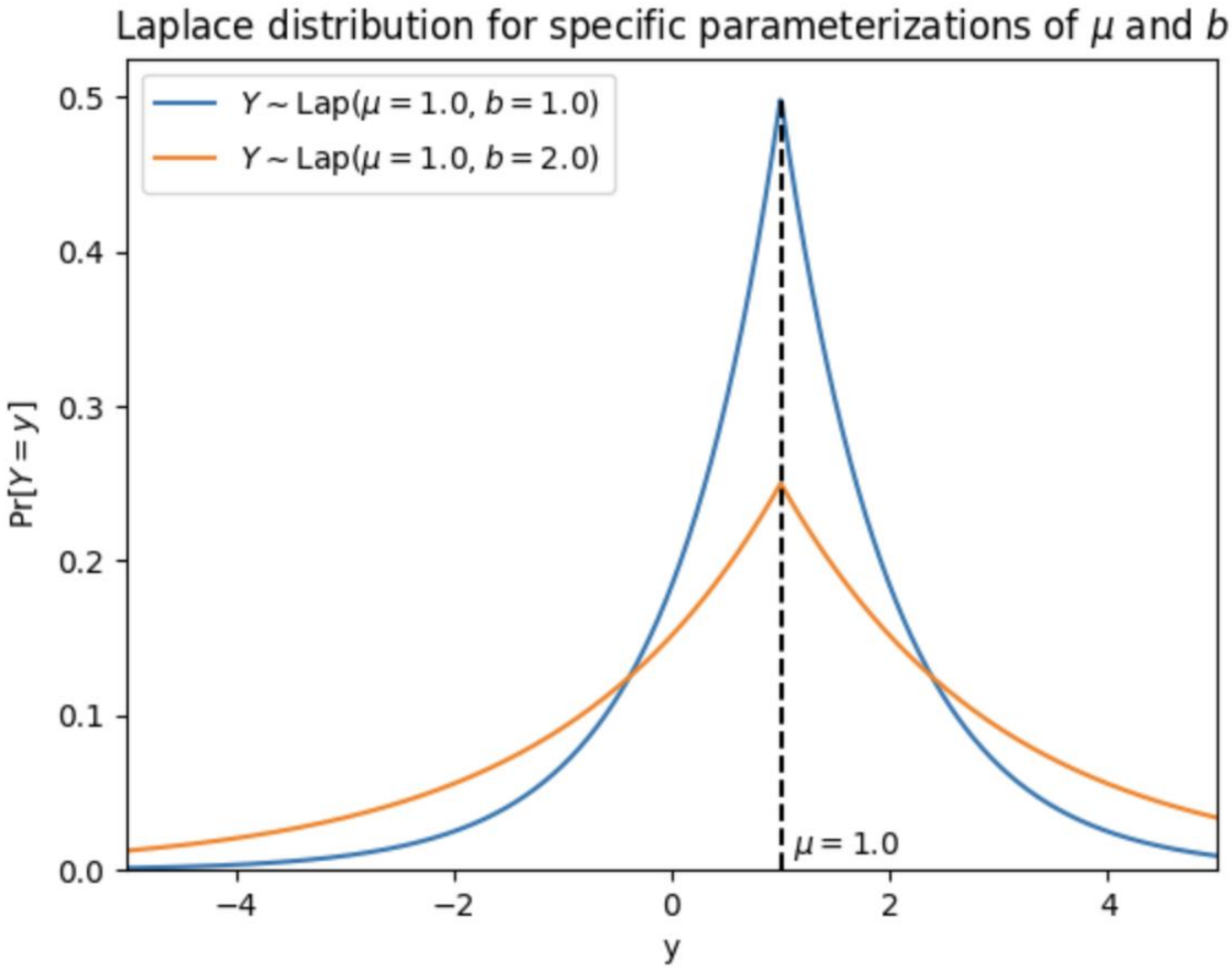
- As before, student *0* joins the circus and drops the class. The professor computes a new mean and updates the score on the board:
- exam_scores[:-1].mean() >>> 87.0

- You now know that anyone who knows how many students the school has could easily calculate the student's grade, which is a 67.

DP Example

- To protect the student's privacy, the professor needs a function, let's call it dp_mean. Under the hood, this function is calculating the mean and adding noise sampled from a Laplace distribution. With this function, the scenario looks like the following:
- dp_mean(exam_scores)
- >>> 83.84393665240519
- dp_mean(exam_scores[:-1])
- >>> 84.49003889353587
- Now when the professor releases these statistics, the students cannot calculate their friend's grade, even if they know the type of noise and scale parameters used!

Laplace Distribution



Differential Data Privacy

- "Differential privacy makes it possible for tech companies to collect and share aggregate information about user habits, while maintaining the privacy of individual users."

- What does it not guarantee?

- DP does not guarantee that one believes to be one's secrets will remain secret. It's important to identify which is general information and which is private information to get benefits from DP umbrella and reduce harm. DP guarantees to protect only private information (mentioned above). So, if one's secret is general information, it will be not protected!

Differential Data Privacy

- To understand this, let's consider a scenario when you, a smoker, decided to be included in a survey. Then, analysis on the survey data reveals that smoking causes cancer. Will you, as a smoker, be harmed by the analysis?

- Perhaps, Based on the fact that you're a smoker, one may guess at your health status. It is certainly the case that he knows more about you after the study than was known before (this is also the reason behind saying it is "general information", not "public information"), but was your information leaked? Differential privacy will take the view that it was not, with the rationale that the impact on the smoker is the same independent of whether or not he was in the study. It is the conclusions reached in the study that affect the smoker, not his presence or absence in the data set.

Example

- Let consider a canonical example to see how a DP algorithm, what satisfies DP criterion, works: Image that you are a social data scientist, who wants to perform an analysis on a survey data about a very taboo behavior. Each entry in the data is an answer (the truth) of individuals, “yes” or “no”, in the surveyed population. Because of a privacy policy, the data holder, or curator, never permits you for direct access to the data.

Example

- You, a DP expert, suggested to the curator a DP algorithm for removing private information in the data, whereby you can perform the analysis on the data. Thus, for each entry, the curator will apply this algorithm:

- Flip a coin (the coin's bias is the probability that its outcome is head and it will be denoted as p_head.).
- If heads, return the answer in the entry.
- If tails, then flip a second coin and return "yes" if heads and "no" if tails.

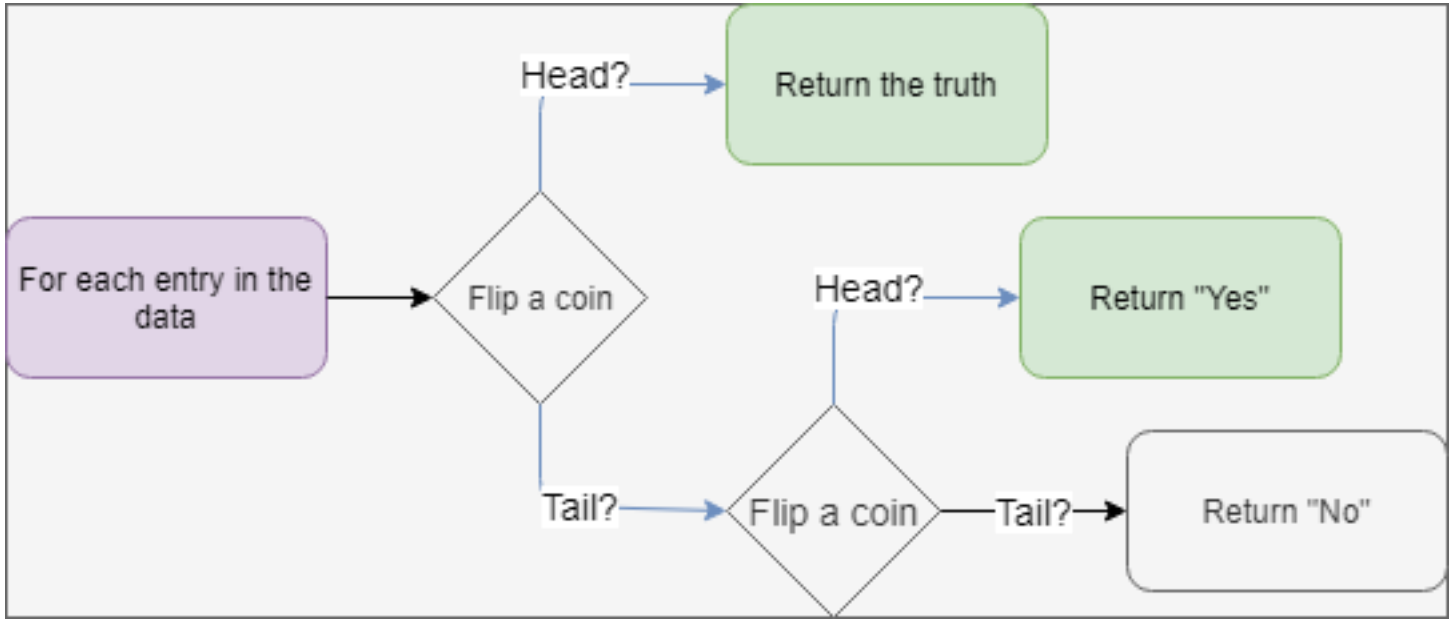
Example

- Now, each person is protected with “plausible deniability”, because a person is plausible to deny the answer by the randomness of flipping a coin. Assume you want to infer the percentage of innocents in the population (p_innocent) from that noisy data. It can be done by these steps:

Example

Learn....Imp

- Compute probability of returning “yes” given that individual isn’t an innocent: P(“yes”|not innocent) = p_head+(1-p_head)*p_head.



Example

- Now, each person is protected with “plausible deniability”, because a person is plausible to deny the answer by the randomness of flipping a coin. Assume you want to infer the percentage of innocents in the population (p_innocent) from that noisy data. It can be done by these steps:

- Compute probability of returning “yes” given that individual isn’t an innocent: P(“yes”|not innocent) = p_head+(1-p_head)*p_head.

Example

later

- What does DP tell us?
- As you can see in Figure the variance of p_innocent increases dramatically and approaches infinity when p_head approaches 0, this leads to a rapid decrease in privacy loss. DP also gives us the same conclusion.

- Thus, when p_head is 0, the distribution of returned result is identical, no matter an individual is innocent or not (the distance of 2 distributions is P(“yes”|not innocent)-P(“yes”|innocent)=p_head, the bias). If the number of innocents participating in the data changed, it does not lead to any changes in information in the noisy returned data. It means that there is no private information in the noisy returned data.

Part 1: Without DP

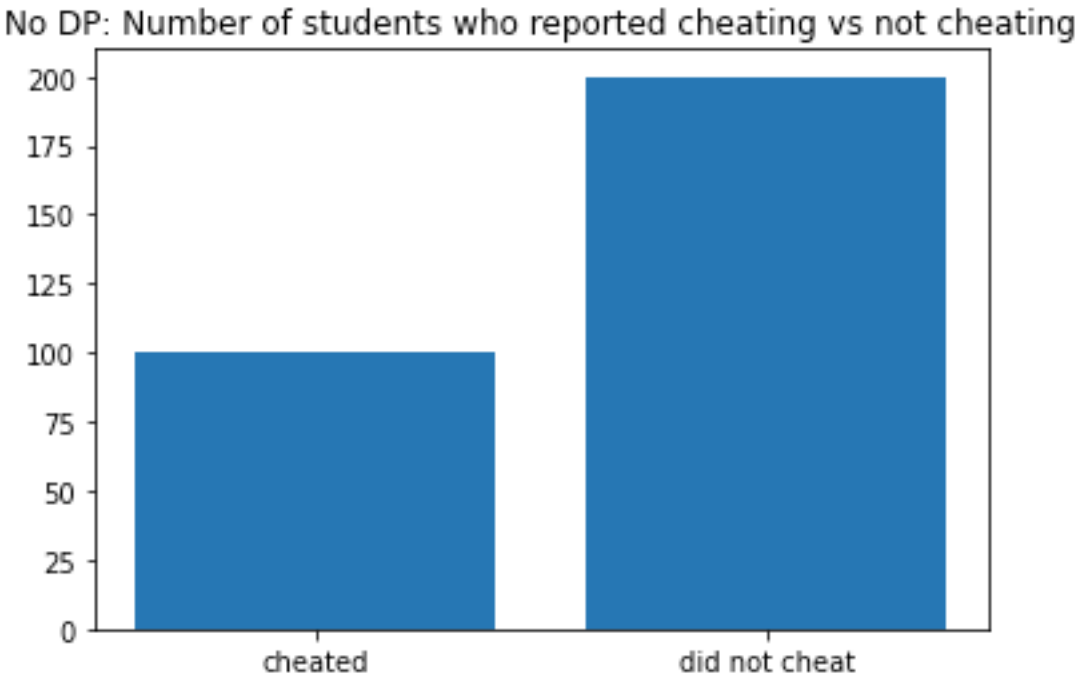
- First, we will see what the mock data looks like without differential privacy. Our “raw data” is represented with 0s and 1s, where 0 means “did not cheat,” and 1 means “cheated.” Each binary number accounts for a different student, and their true “cheating status” is the only metric measured here.

Part 1: Without DP

- rawData = [0, 1, 1, 1, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 0, 0, 1, 1, \ 1, 0, 0, 1, 0, 0, 0, 0, 1, 1, 0, 1, 0, 1, 0, 1, 0, 0, 0, 0, \ 0, 1, 1, 0, 0, 0, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 1, \ 1, 0, 1, 1, 0, 0, 0, 0, 1, 0, 1, 0, 0, 1, 1, 0, 0, 0, 0, 0, \ 0, 0, 0, 0, 1, 1, 0, 1, 0, 0, 1, 0, 1, 0, 0, 0, 0, 0, 0, 0, \ 0, 1, 1, 1, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 0, 0, 1, 1, \ 1, 0, 0, 1, 0, 0, 0, 0, 1, 1, 0, 1, 0, 1, 0, 1, 0, 0, 0, 0, \ 0, 1, 1, 0, 0, 0, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 1, \ 1, 0, 1, 1, 0, 0, 0, 0, 1, 0, 1, 0, 0, 1, 1, 0, 0, 0, 0, 0, \ 0, 0, 0, 0, 1, 1, 0, 1, 0, 0, 1, 0, 1, 0, 0, 0, 0, 0, 0, 0, \ 0, 1, 1, 1, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 0, 0, 1, 1, \ 1, 0, 0, 1, 0, 0, 0, 0, 1, 1, 0, 1, 0, 1, 0, 1, 0, 0, 0, 0, \ 0, 1, 1, 0, 0, 0, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 1, \ 1, 0, 1, 1, 0, 0, 0, 0, 1, 0, 1, 0, 0, 1, 1, 0, 0, 0, 0, 0, \ 0, 0, 0, 0, 1, 1, 0, 1, 0, 0, 1, 0, 1, 0, 0, 0, 0, 0, 0, 1]

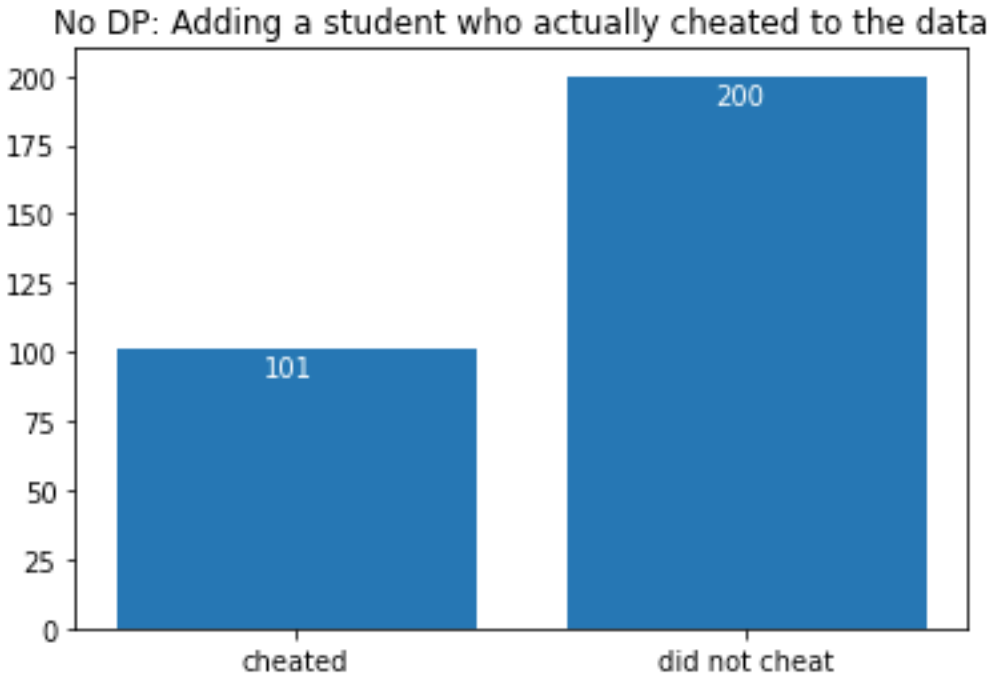
- plotData = {'cheated': rawData.count(1), 'did not cheat': rawData.count(0)}
- labels = list(plotData.keys())
- values = list(plotData.values())

Part 1: Without DP



As you can see, 100 students report cheating, and 200 report not cheating.

Part 1: Without DP



I

n this traditional survey, when we add a student, it's very easy to tell the difference between the two results. If the "cheated" count goes up, the student cheated. If the "did not cheat" count goes up, then they didn't cheat. because the cheated column is now 101 instead of 100, that means student #301 cheated.

nice

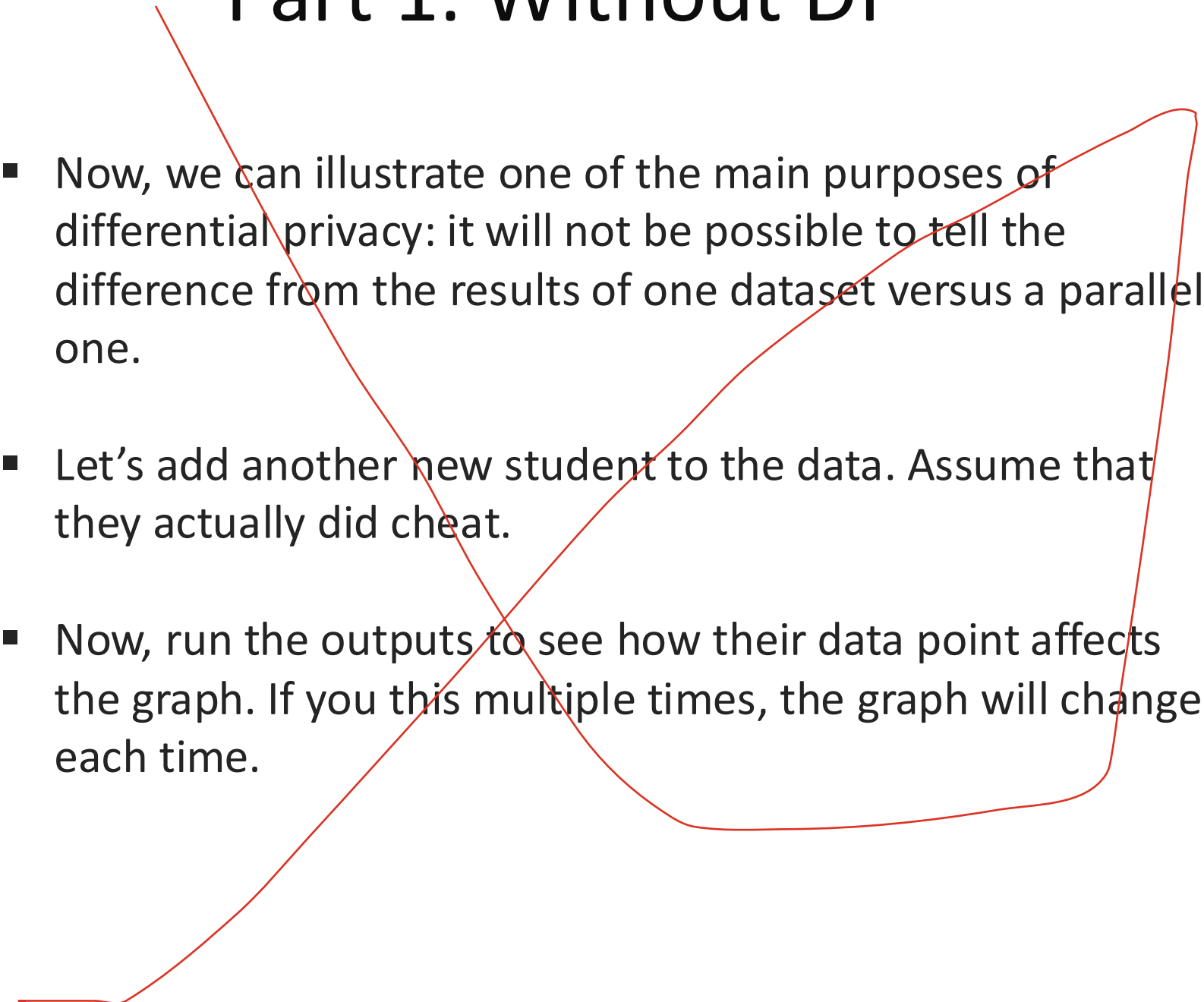
Part 1: Without DP

- Now, we can illustrate one of the main purposes of differential privacy: it will not be possible to tell the difference from the results of one dataset versus a parallel one.

- Let's add another new student to the data. Assume that they actually did cheat.

- Now, run the outputs to see how their data point affects the graph. If you this multiple times, the graph will change each time.

Part 1: Without DP



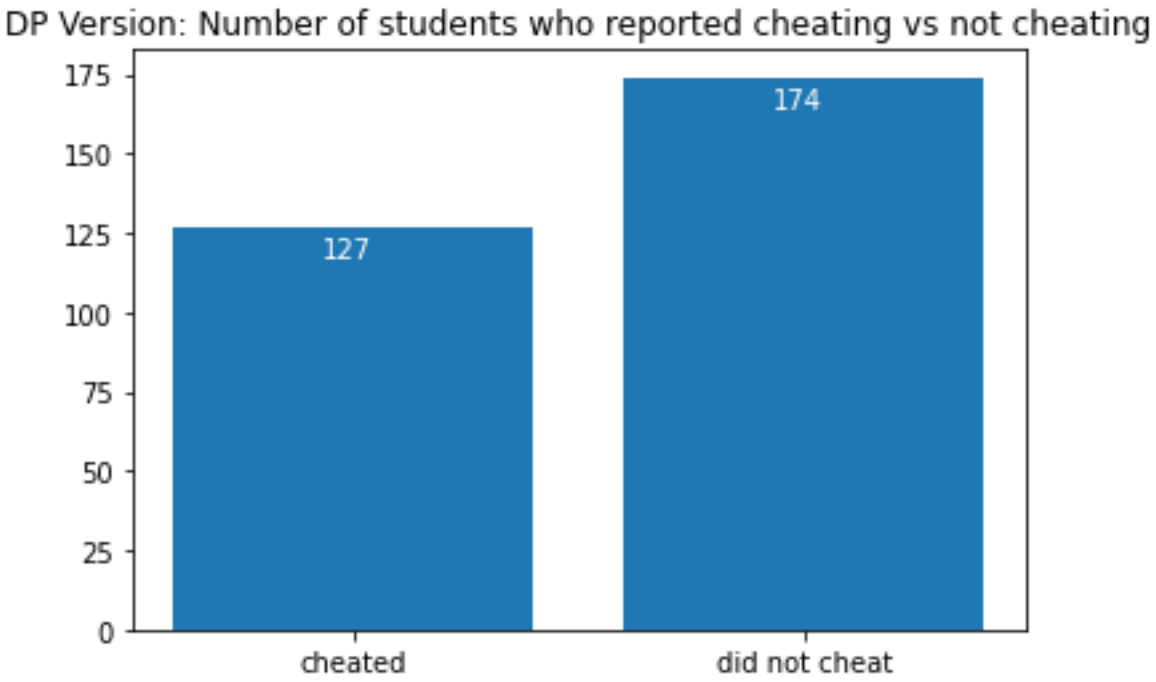
- Now, we can illustrate one of the main purposes of differential privacy: it will not be possible to tell the difference from the results of one dataset versus a parallel one.
- Let's add another new student to the data. Assume that they actually did cheat.
- Now, run the outputs to see how their data point affects the graph. If you this multiple times, the graph will change each time.

Part 2: With DP

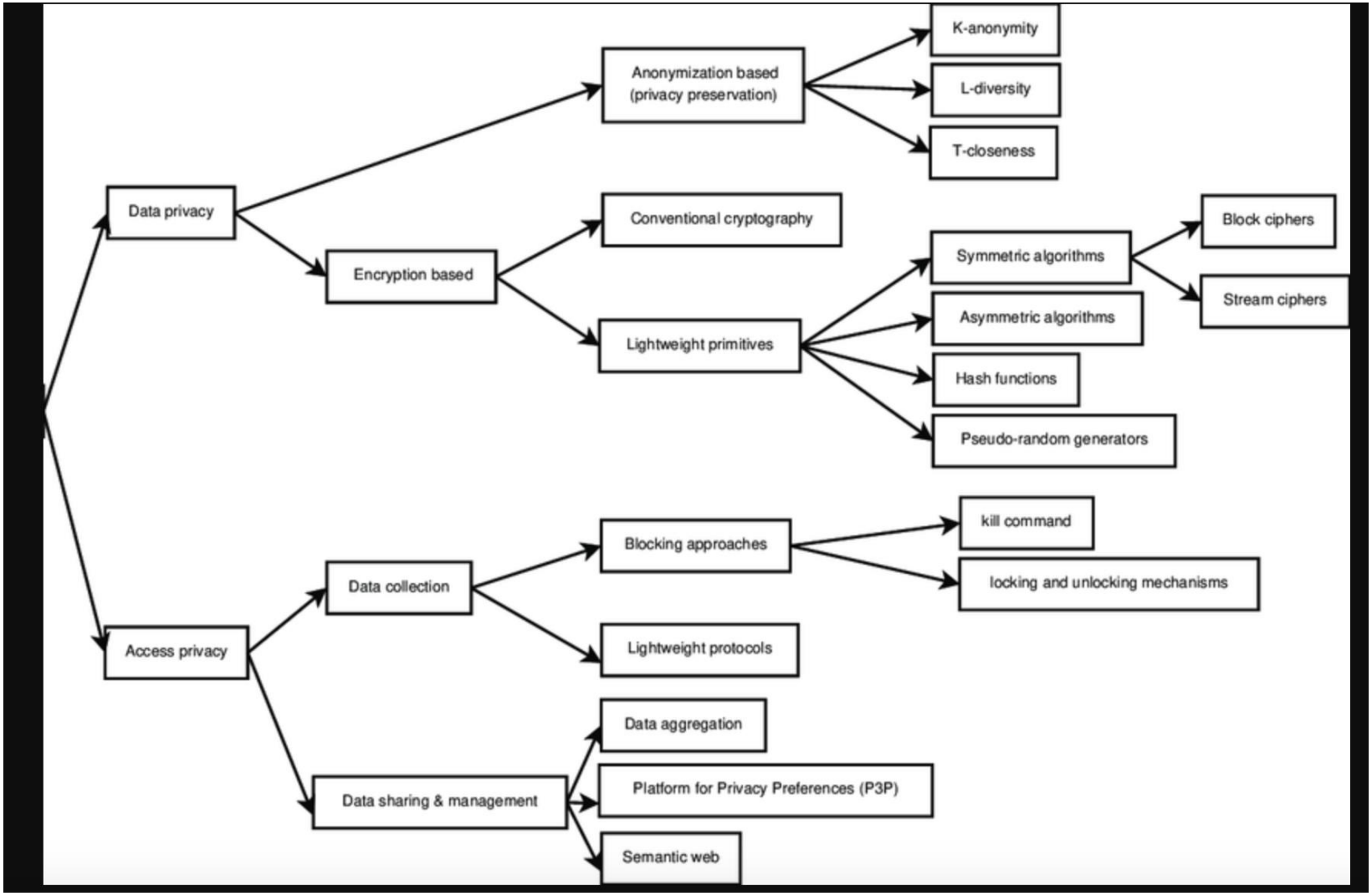
- To implement DP, each student flips a coin. If it lands on heads, they will truthfully answer. If it lands on tails, they flip another coin. If it lands on heads, they respond that they haven't cheated, and if it lands on tails, they respond that they have cheated.

- **If you run the code multiple times, you'll notice how the graph changes every time. That's because DP algorithms inject randomness, such as we did with a coin flip.**

Part 1: With DP



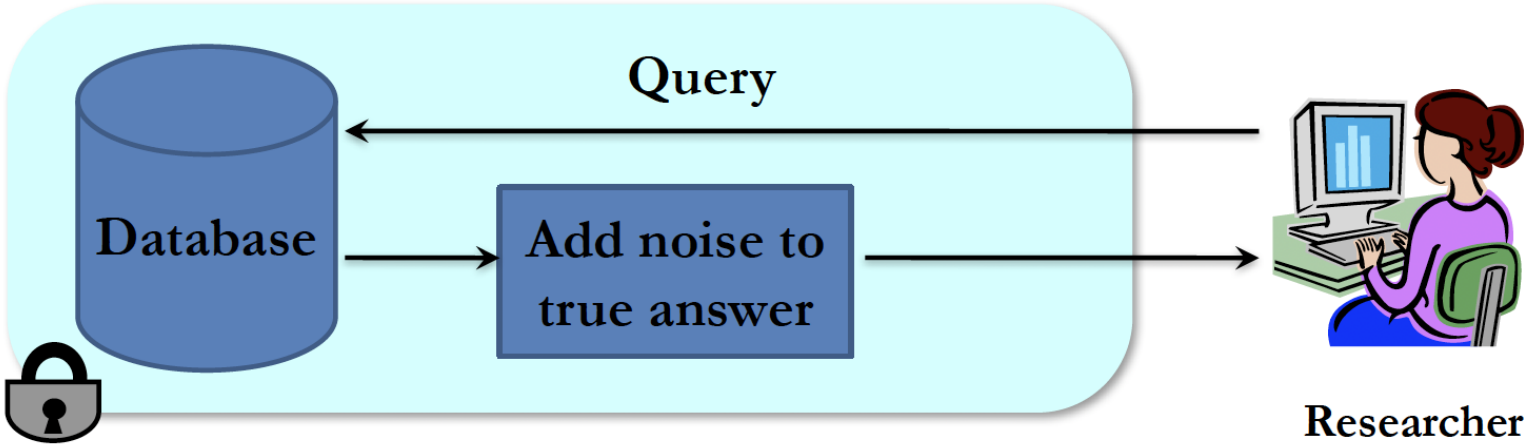
As you can see, 100 students report cheating, and 200 report not cheating.


Data privacy
Anonymization based (privacy preservation)
K-anonymity
L-diversity
T-closeness
Encryption based
Conventional cryptography
Lightweight primitives
Symmetric algorithms
Block ciphers
Stream ciphers
Asymmetric algorithms
Hash functions
Pseudo-random generators
Access privacy
Data collection
Blocking approaches
kill command
locking and unlocking mechanisms
Lightweight protocols
Data sharing & management
Data aggregation
Platform for Privacy Preferences (P3P)
Semantic web

Agenda for today class

- Last Class
  - Data Privacy
- Today's class
  - Differential Privacy

Output Randomization



- Add noise to answers such that:
  - Each answer does not leak too much information about the database.
  - Noisy answers are close to the original answers.

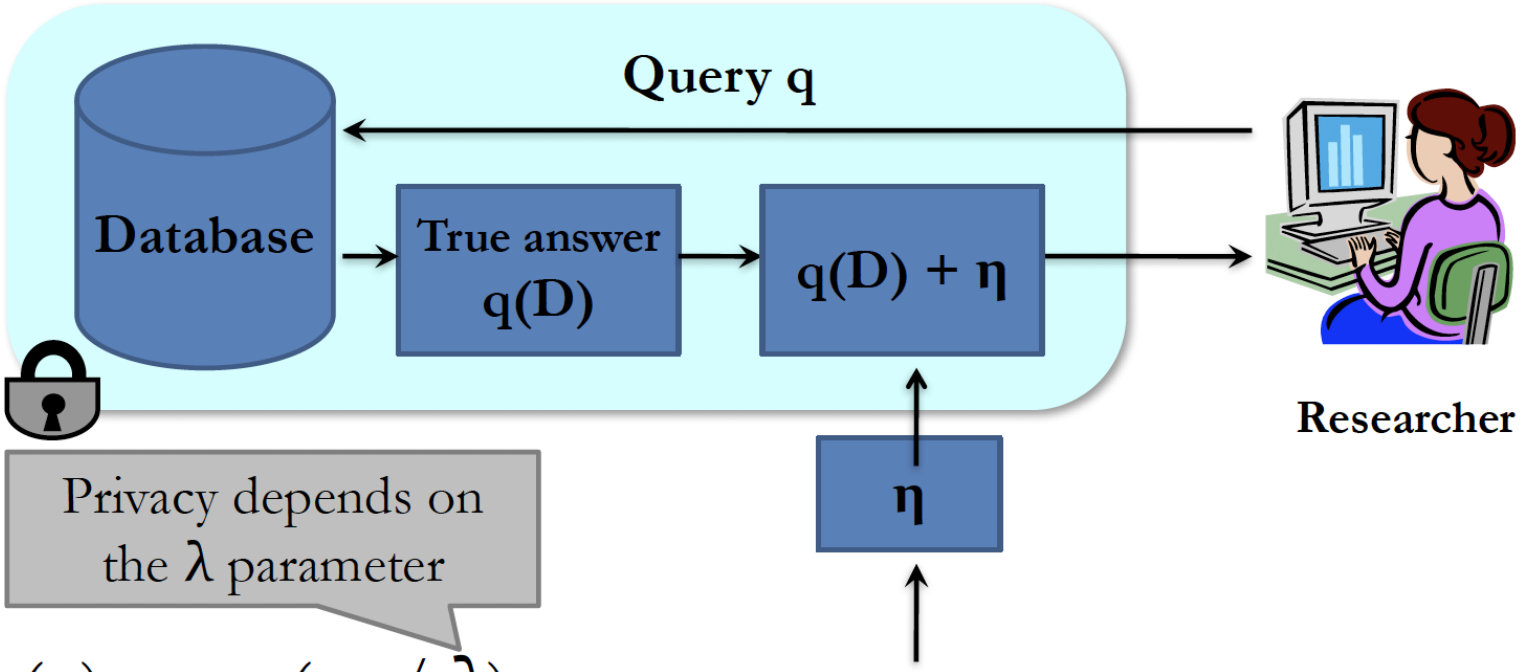
Differential Privacy

Basic setup:

- There is a database $\mathcal{D}$ which potentially contains sensitive information about individuals.
- The database curator has access to the full database. We assume the curator is trusted.
- The data analyst wants to analyze the data. She asks a series of queries to the curator, and the curator provides a response to each query.
- The way in which the curator responds to queries is called the mechanism. We'd like a mechnism that gives helpful responses but avoids leaking sensitive information about individuals.

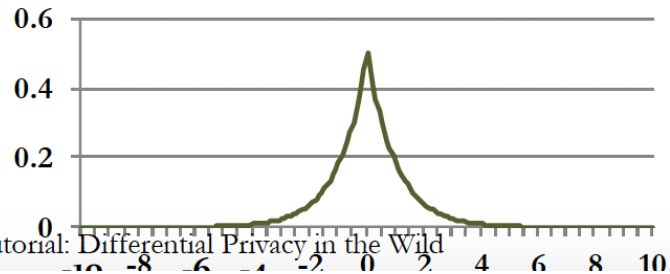
imp

Laplace Mechanism



$h(\eta) \propto \exp(-\eta / \lambda)$

Laplace Distribution – Lap(λ)



Mean: 0,
Variance: $2\lambda^2$

You are still most likely to say "google.com," but there's a chance (~42%) you will report facebook.com or wikipedia.org to preserve privacy!

Example

- Data: x_i = website visited by student i today
- • Range: Y = {website names}
- For each name y, let q(y; X) = #{i : x_i = y}

imp write
no of times the website visted given X

Goal: output **the most frequently visited site**

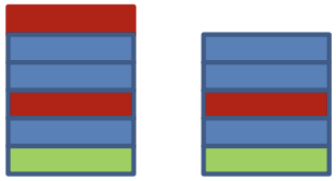
- **Procedure: Given X,** Output website y ***with probability prop to*** $e^{\varepsilon\, q(y,X)}$
-
- Popular sites exponentially more likely than rare ones
- Website scores don't change too quickly

Differential Privacy

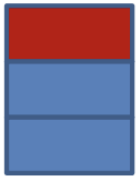
[Dwork ICALP 2006]

For every pair of inputs that differ in one row

For every output …



D_1 D_2



O

Adversary should not be able to distinguish between any D_1 and D_2 based on any O

$$\ln\left(\frac{\Pr[A(D_1) = o]}{\Pr[A(D_2) = o]}\right) \leq \varepsilon, \qquad \varepsilon > 0$$

Differential Privacy

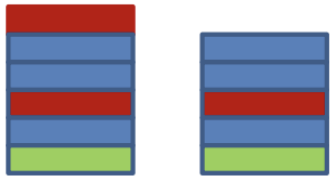
imp

- Epsilon represents the privacy budget, which is the amount of privacy an algorithm "spends" when processing data. Each time a function is applied to a dataset, it consumes a portion of the privacy budget (ε), and the total epsilon across all operations determines the overall privacy guarantee.

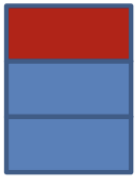
Privacy Parameter ε

For every pair of inputs that differ in one row

For every output …



D_1 D_2



O

Imp

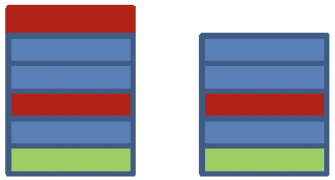
$$\Pr[A(D_1) = o] \leq e^{\varepsilon} \Pr[A(D_2) = o]$$

Controls the degree to which D_1 and D_2 can be distinguished.
Smaller the ε more the privacy (and worse the utility)

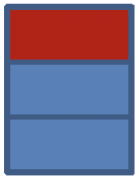
Differential Privacy

For every pair of inputs that differ in one row

For every output …



D_1 D_2



O

imp

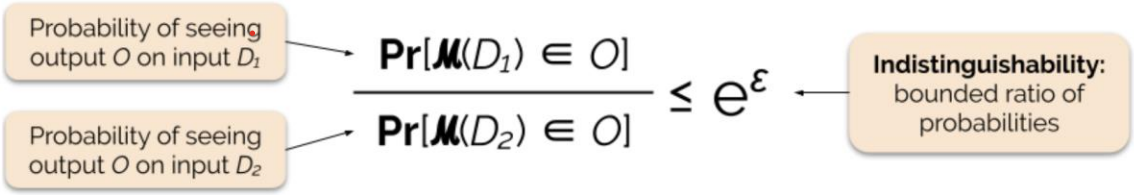
If algorithm A satisfies differential privacy then

$$\frac{\Pr[A(D_1) = O]}{\Pr[A(D_2) = O]} < \exp(\varepsilon) \quad (\varepsilon > 0)$$

Intuition: adversary should not be able to use output O to distinguish between any D_1 and D_2

Important def of diff privacy

What is Differential Privacy



- The strength of this privacy guarantee is a policy decision and is often expressed as Epsilon (lower epsilon is more privacy)
- The higher the privacy guarantee the more the data need to be 'fudged' and the more risk that usability suffers

Differential Privacy

- Two databases $\mathcal{D}_1$ and $\mathcal{D}_2$ are neighbouring if they agree except for a single entry.
- **Idea:** if the mechanism behaves nearly identically for $\mathcal{D}_1$ and $\mathcal{D}_2$, then an attacker can't tell whether $\mathcal{D}_1$ or $\mathcal{D}_2$ was used (and hence can't learn much about the individual).
- **Definition:**
  - A mechanism $\mathcal{M}$ is ε-differentially private if for any two neighbouring databases $\mathcal{D}_1$ and $\mathcal{D}_2$, and any set $\mathcal{R}$ of possible responses

$$\Pr(\mathcal{M}(\mathcal{D}_1) \in \mathcal{R}) \leq \exp(\varepsilon) \Pr(\mathcal{M}(\mathcal{D}_2) \in \mathcal{R}).$$

- **Note:** for small ε, $\exp(\varepsilon) \approx 1 + \varepsilon$.
- **A consequence:** for any possible response y,

$$\exp(-\varepsilon) \leq \frac{\Pr(\mathcal{M}(\mathcal{D}_1) = y)}{\Pr(\mathcal{M}(\mathcal{D}_2) = y)} \leq \exp(\varepsilon)$$

The Role of Laplace Distribution

- The Laplace distribution, also known as the double exponential distribution, plays a crucial role in differential privacy mechanisms. It is characterized by its heavy tails, which make it well-suited for modeling data with outliers or extreme values. The Laplace distribution is symmetric around its mean, making it ideal for adding noise to data without biasing the results.

write

The Laplace Distribution:

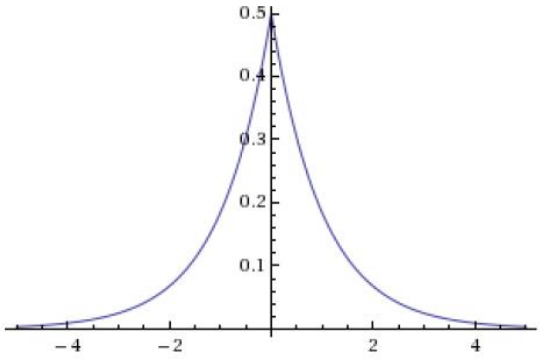
Lap(b) is the probability distribution with p.d.f.:

$$p(x\,|b) = \frac{1}{2b}\exp\left(-\frac{|x|}{b}\right)$$

i.e. a symmetric exponential distribution

$Y \sim \mathrm{Lap}(b), \qquad E[|Y|] = b$

$\Pr[|Y| \geq t \cdot b] = e^{-t}$



The Role of Laplace Distribution

- In the context of differential privacy, the Laplace distribution is often used to generate the noise that is added to the data. The amount of noise added is determined by a parameter called the privacy budget, which quantifies the level of privacy protection desired. By adjusting the scale parameter of the Laplace distribution based on the privacy budget, it's possible to achieve the desired level of privacy while still allowing for accurate analysis of the data.

The Laplace Mechanism

Laplace$(D, Q: \mathbb{N}^{|X|} \rightarrow \mathbb{R}^k, \epsilon)$:

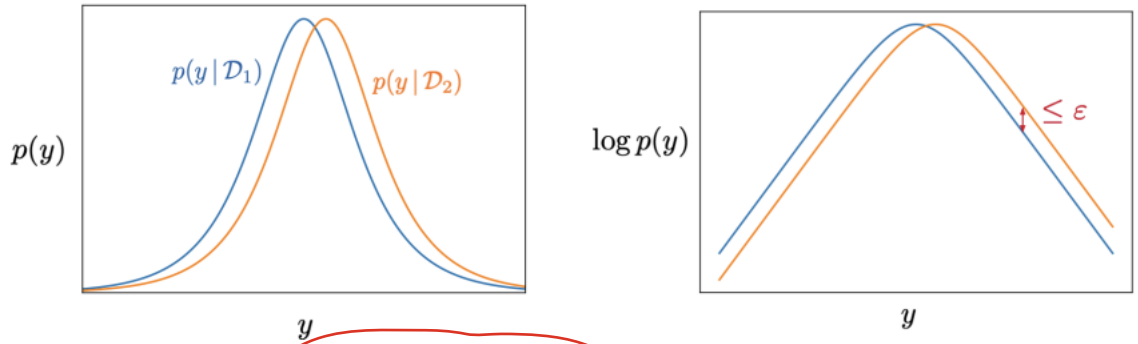
1. Let $\Delta = GS(Q)$.
2. For $i = 1$ to k: Let $Y_i \sim \text{Lap}(\frac{\Delta}{\epsilon})$.
3. Output $Q(D) + (Y_1, \ldots, Y_k)$

Independently perturb each coordinate of the output with Laplace noise scaled to the sensitivity of the function.

Idea: This should be enough noise to hide the contribution of any single individual, no matter what the database was.

Differential Privacy

Visually:



Notice that the tail behavior is important.

Differential Privacy

- The Laplace mechanism adds noise from a Laplace distribution, where the scale parameter is determined by the sensitivity of the function and the epsilon value. The noise scales linearly with the sensitivity divided by epsilon ($\Delta f/\varepsilon$). Therefore, a smaller epsilon leads to larger noise values, and vice versa.

Differential Privacy

- The choice of epsilon often depends on the specific application and the desired balance between privacy and utility. Common values for epsilon might range from 10^-5 to 0.1 or higher, depending on the context. In medical research, for example, an epsilon of 0.1 might be used per emoji to provide a relatively strong privacy guarantee, while a delta of 1e-5 might be used for a probabilistic privacy guarantee.

- Anna is an attacker who wants to figure out if Patrick (x) is in the cancer database $\mathcal{D}$. Her prior probability for him being in the database is 0.4. $\mathcal{D}$ is ε-differentially private. She makes a query and gets back $y = \mathcal{M}(\mathcal{D})$.
- She's narrowed it down to two possible databases $\mathcal{D}_1$ and $\mathcal{D}_2$, which are identical except that $x \in \mathcal{D}_1$ and $x \notin \mathcal{D}_2$.
- After observing y, she computes her posterior probability using Bayes' Rule:

Bayes Theorem

$$\begin{aligned}
\Pr(x \in \mathcal{D} \mid y) &= \frac{\Pr(x \in \mathcal{D}) \Pr(y \mid x \in \mathcal{D})}{\Pr(x \in \mathcal{D}) \Pr(y \mid x \in \mathcal{D}) + \Pr(x \notin \mathcal{D}) \Pr(y \mid x \notin \mathcal{D})} \\
&\geq \frac{\Pr(x \in \mathcal{D}) \Pr(y \mid x \in \mathcal{D})}{\Pr(x \in \mathcal{D}) \Pr(y \mid x \in \mathcal{D}) + \exp(\varepsilon) \Pr(x \notin \mathcal{D}) \Pr(y \mid x \in \mathcal{D})} \\
&= \frac{\Pr(x \in \mathcal{D})}{\Pr(x \in \mathcal{D}) + \exp(\varepsilon) \Pr(x \notin \mathcal{D})} \\
&\geq 0.4 \exp(-\varepsilon)
\end{aligned}$$

- Similarly, $\Pr(x \in \mathcal{D} \mid y) \leq 0.4 \exp(\varepsilon)$. So Anna hasn't learned much about Patrick.

- In what sense does this definition guarantee privacy?
- Suppose a data analyst takes the result $y = \mathcal{M}(\mathcal{D})$ and further processes it with some algorithm f (without peeking at the data itself). Is it still private?
- Let $\mathcal{R}$ be a set of possible outputs, and $\mathcal{R}'$ be the pre-image under f, i.e. $\mathcal{R}' = \{y : f(y) \in \mathcal{R}\}$.

can give, re study

$$\begin{aligned}\Pr(f(\mathcal{M}(\mathcal{D}_1)) \in \mathcal{R}) &= \Pr(\mathcal{M}(\mathcal{D}_1) \in \mathcal{R}') \\ &\leq \exp(\varepsilon) \Pr(\mathcal{M}(\mathcal{D}_2) \in \mathcal{R}') \\ &= \exp(\varepsilon) \Pr(f(\mathcal{M}(\mathcal{D}_2)) \in \mathcal{R})\end{aligned}$$

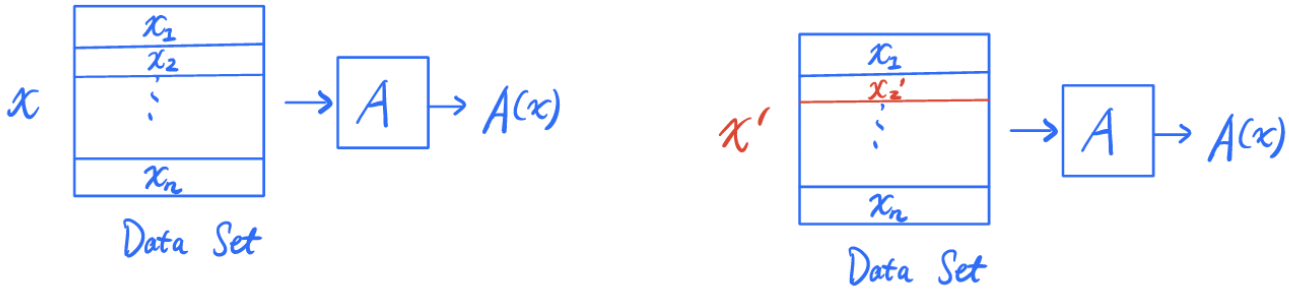
- Hence, the composition $f \circ \mathcal{M}$ is also ε-differentially private. No matter how clever the analyst is, or the resources she throws at it, she can't learn more than ε about an individual entry!

Differential Privacy Composition

- Last Class
  - Data Privacy
- Today's class
  - Differential Privacy

Differential Privacy

Thought Experiment.

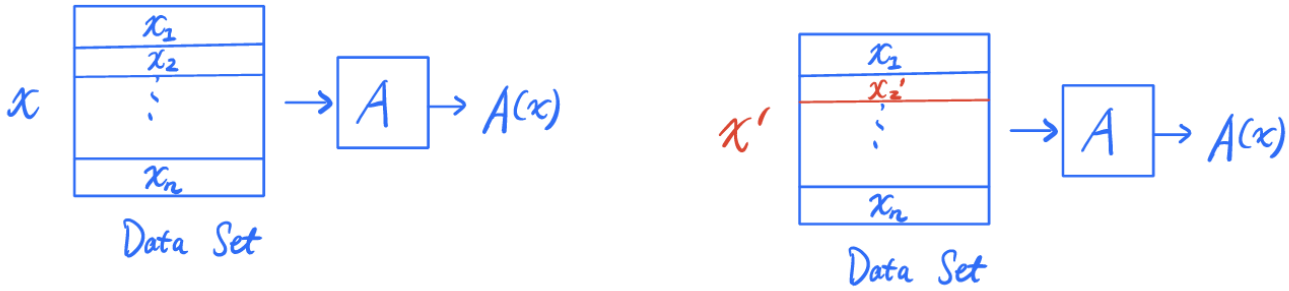


x' is a neighbor of x if they differ in one data point.

Idea of DP: Neighboring data sets induce close output distributions

Differential Privacy

Thought Experiment.



x' is a neighbor of x if they differ in one data point.

Idea of DP: Neighboring data sets induce close output distributions

Differential Privacy

write

Definition. (Differential Privacy)

A is ε-differentially private if

for all neighbors x and x'

for all subsets E of outputs

$$\mathbb{P}[A(x) \in E] \leq e^{\varepsilon} \mathbb{P}[A(x') \in E]$$

↑

This is an algorithmic property.

Differential Privacy

Definition. (Differential Privacy)

A is ε-differentially private if
for all neighbors x and x'
for all subsets E of outputs

$$\mathbb{P}[A(x) \in E] \leq e^{\varepsilon} \mathbb{P}[A(x') \in E]$$

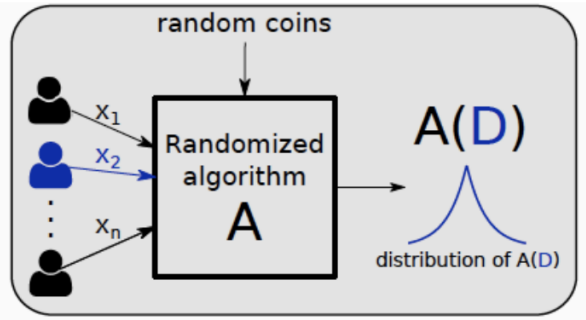
What is ε?

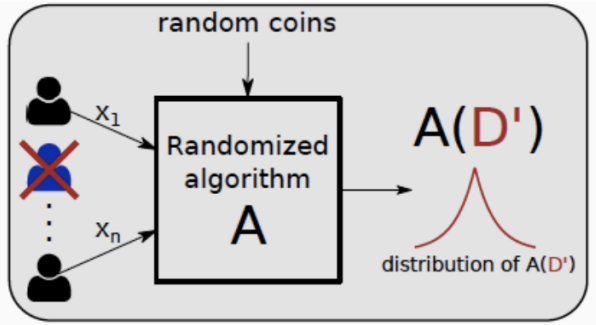
- Measure of info leakage (called max divergence)
- Small constant $\approx \frac{1}{10}, 1$, but not $\frac{1}{2^{80}}$ or 100

Differential Privacy

Differential Privacy (DP)

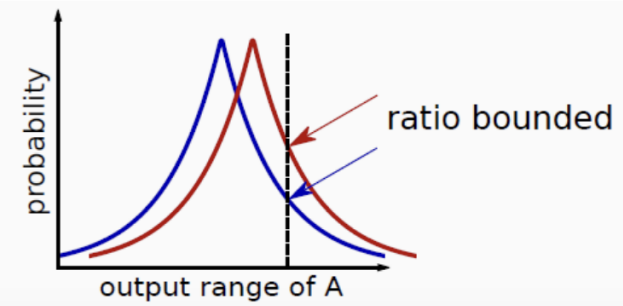
Dwork, McSherry, Nissim and Smith [2006]





A thought experiment:

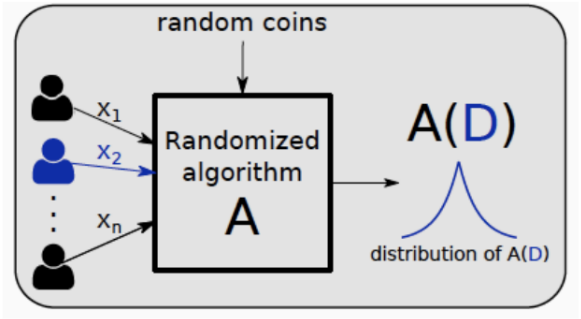
- Change, add or remove one person's data
- Will the probabilities of the outcomes change?

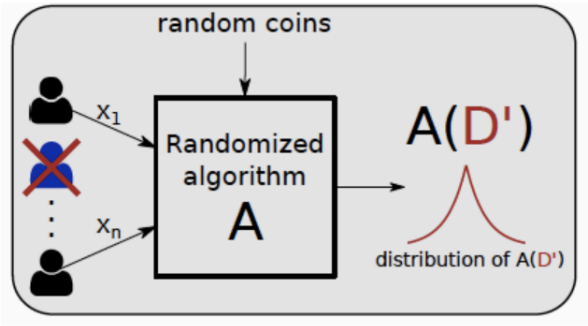


Differential Privacy

Differential Privacy (DP)

Dwork, McSherry, Nissim and Smith [2006]





The randomized algorithm A is ϵ-differentially private if for all neighboring datasets D, D' and for all outputs S:

(a) $P[A(D) \in S] \leq e^{\epsilon} \cdot P[A(D') \in S]$

(b) $P[A(D') \in S] \leq e^{\epsilon} \cdot P[A(D) \in S]$

A thought experiment:

- Change, add or remove one person's data
- Will the probabilities of the outcomes change?

Neighboring datasets

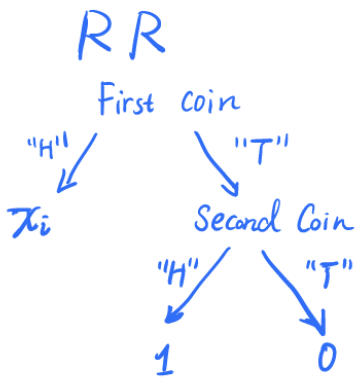
Requirement of DP: Both distributions should be close $(\epsilon \approx 0)$

Differential Privacy

Example: Randomized Response (In lecture 1)

Each person has a secret bit $x_i = 0$ or $x_i = 1$
(Have you ever done XYZ?)

Input: $x_1, \ldots, x_n$
Output: $y_1, \ldots, y_n$



Differential Privacy

RR is $\ln(3)$- diffentially private

Proof.

- Fix two neighboring data sets

$$x = (x_1, \ldots, x_i, \ldots, x_n), \quad x' = (x_1, \ldots, x_i', \ldots, x_n)$$

- To start, fix some output $y = (y_1, \ldots, y_n) \in \{0,1\}^n$

$$\frac{\mathbb{P}[RR(x) = y]}{\mathbb{P}[RR(x') = y]} = \frac{\mathbb{P}[Y_i = y_i \mid x_i]}{\mathbb{P}[Y_i = y_i \mid x_i']} \qquad 3 \text{ or } \frac{1}{3}$$

$$\Rightarrow \mathbb{P}[RR(x) = y] \leq e^{\ln(3)} \mathbb{P}[RR(x') = y]$$

- To Complete, For any $E \subseteq \{0,1\}^n$

$$\mathbb{P}[RR(x) \in E] = \sum_{y \in E} \mathbb{P}[RR(x) = y]$$

$$\leq e^{\varepsilon} \sum_{y \in E} \mathbb{P}[RR(x') = y] = \mathbb{P}[RR(x') \in E]$$

Differential Privacy Composition

Composition.

Cross referencing :

- 28 years old
- Zipcode 13012
- In both data sets

Overlap datasets

	Non-Sensitive			Sensitive
	Zip code	Age	Nationality	Condition
1	130**	<30	*	AIDS
2	130**	<30	*	Heart Disease
3	130**	<30	*	Viral Infection
4	130**	<30	*	Viral Infection
5	130**	≥40	*	Cancer
6	130**	≥40	*	Heart Disease
7	130**	≥40	*	Viral Infection
8	130**	≥40	*	Viral Infection
9	130**	3*	*	Cancer
10	130**	3*	*	Cancer
11	130**	3*	*	Cancer
12	130**	3*	*	Cancer

	Non-Sensitive			Sensitive
	Zip code	Age	Nationality	Condition
1	130**	<35	*	AIDS
2	130**	<35	*	Tuberculosis
3	130**	<35	*	Flu
4	130**	<35	*	Tuberculosis
5	130**	<35	*	Cancer
6	130**	<35	*	Cancer
7	130**	≥35	*	Cancer
8	130**	≥35	*	Cancer
9	130**	≥35	*	Cancer
10	130**	≥35	*	Tuberculosis
11	130**	≥35	*	Viral Infection
12	130**	≥35	*	Viral Infection

Differential Privacy Composition

Pitfalls of K-Anonymization: Composition

	Non-Sensitive			Sensitive
	Zip code	Age	Nationality	Condition
1	130**	<30	*	AIDS
2	130**	<30	*	Heart Disease
3	130**	<30	*	Viral Infection
4	130**	<30	*	Viral Infection
5	130**	≥40	*	Cancer
6	130**	≥40	*	Heart Disease
7	130**	≥40	*	Viral Infection
8	130**	≥40	*	Viral Infection
9	130**	3*	*	Cancer
10	130**	3*	*	Cancer
11	130**	3*	*	Cancer
12	130**	3*	*	Cancer

	Non-Sensitive			Sensitive
	Zip code	Age	Nationality	Condition
1	130**	<35	*	AIDS
2	130**	<35	*	Tuberculosis
3	130**	<35	*	Flu
4	130**	<35	*	Tuberculosis
5	130**	<35	*	Cancer
6	130**	<35	*	Cancer
7	130**	≥35	*	Cancer
8	130**	≥35	*	Cancer
9	130**	≥35	*	Cancer
10	130**	≥35	*	Tuberculosis
11	130**	≥35	*	Viral Infection
12	130**	≥35	*	Viral Infection

2 hospital release K anonymous tables for patients' medical history

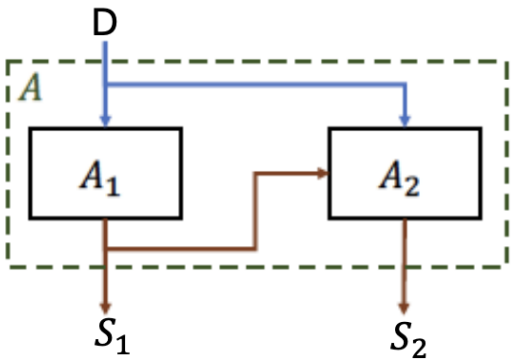
A 28 year old person visited both hospitals

The person has AIDS

Ganta, Kashivishwanathan, Smith 2008

Differential Privacy

Properties of DP: Basic Composition



Theorem: Let $A: D \rightarrow S_1 \times S_2$ be an composed algorithm that outputs (s_1, s_2) where $s_1 = A_1(D)$ and $s_2 = A_2(s_1, D)$. Then A is $(\epsilon_1 + \epsilon_2)$-DP

Allows to control cumulative privacy for multiple queries on the same dataset

$A_1: D \rightarrow S_1$ is ϵ_1-DP

$A_2: S_1 \times D \rightarrow S_2$ is ϵ_2-DP ⟶ $A_2(s_1, \cdot)$ is ϵ_2-DP for all $s_1 \in S_1$

Extends to k such DP algorithms (one for each query): cumulative privacy scales linearly with number of queries

Can be improved using **Advanced Composition**: cumulative privacy scales sub-linearly with number of queries

Differential Privacy

Setting ϵ: Group Privacy

Theorem: Let D_1, D_2 be two datasets of n records that differ in $1 \leq k \leq n$ positions. If an algorithm A is ϵ-DP, then for all outputs S, we have

$$P[A(D_1) \in S] \leq e^{k\epsilon} \cdot P[A(D_2) \in S]$$

Different than composition

Need to set $\epsilon \geq \frac{1}{n}$ for reasonable utility

Hide participation of
1. An individual who contribute several records
2. Groups of people whose data are strongly correlated

Why?

DP algorithms can't give useful output for small datasets

If $\epsilon \ll \frac{1}{n}$ then regardless of number of differing positions k, the distributions of $A(D_1)$ and $A(D_2)$ are almost same

→ To ensure high privacy, the algorithm ignores its input