Database Privacy - I

Outline

- Why privacy?
- Privacy Attacking Examples
- Conventional principles and limitations
  - K-anonymity
  - L-diversity
  - T-closeness

On the Internet, nobody knows you're a dog?

The New Yorker, July 5, 1993

Your personal information is kept by

- Government agencies
- Banks/Financial business
- Online shopping web sites
- Advertising companies

Publishing sensitive data about individuals.

- Medical research
  - What treatments have the best outcomes?
  - How can we recognize the onset of disease earlier?
  - Are certain drugs better for certain phenotypes?
- Web search
  - What are people really looking for when they search?
  - How can we give them the most authoritative answers?
- Public health
  - Where are our outbreaks of unpleasant diseases?
  - What behavior patterns or patient characteristics are correlated with these diseases?

Publishing sensitive data about individuals.

- Social and computer networking
  - What is the pattern of phone/data/multimedia network usage? How can we better use existing (or plan new) infrastructure to handle this traffic?
  - How do people relate to one another, e.g., as mediated by Facebook?
  - How is society evolving (Census data)?
- Industrial data (individual = company; need SMC if no TTP)
  - What were the total sales, over all companies, in a sector last year/quarter/month?
  - What were the characteristics of those sales: who were the buyers, how large were the purchases, etc.?

Today, access to these data sets is usually strictly controlled.

Only available:

- Inside the company/agency that collected the data
- Or after signing a legal contract
  - Click streams, taxi data
- Or in very coarse-grained summaries
  - Public health
- Or after a very long wait
  - US Census data details
- Or with definite privacy issues
  - US Census reports, the AOL click stream, dbGaP summary tables, Enron email
- Or with IRB (Institutional Review Board) approval
  - dbGaP summary tables

Society would benefit if we could publish some useful form of the data, without having to worry about privacy.

Why is access so strictly controlled?

No one should learn who had which disease.

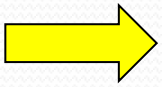
Name	Age	Sex	Zipcode	Disease
Andy	5	M	12000	gastric ulcer
Bill	9	M	14000	dyspepsia
Ken	6	M	18000	pneumonia
Nash	8	M	19000	bronchitis
Joe	12	M	22000	pneumonia
Sam	19	M	24000	pneumonia
Linda	21	F	58000	flu
Jane	26	F	36000	gastritis
Sarah	28	F	37000	pneumonia
Mary	56	F	33000	flu

"Microdata"

What if we “de-identify” the records by removing names?

Name	Age	Sex	Zipcode	Disease
Andy	5	M	12000	gastric ulcer
Bill	9	M	14000	dyspepsia
Ken	6	M	18000	pneumonia
Nash	8	M	19000	bronchitis
Joe	12	M	22000	pneumonia
Sam	19	M	24000	pneumonia
Linda	21	F	58000	flu
Jane	26	F	36000	gastritis
Sarah	28	F	37000	pneumonia
Mary	56	F	33000	flu

publish



Age	Sex	Zipcode	Disease
5	M	12000	gastric ulcer
9	M	14000	dyspepsia
6	M	18000	pneumonia
8	M	19000	bronchitis
12	M	22000	pneumonia
19	M	24000	pneumonia
21	F	58000	flu
26	F	36000	gastritis
28	F	37000	pneumonia
56	F	33000	flu

We can re-identify people, absolutely or probabilistically

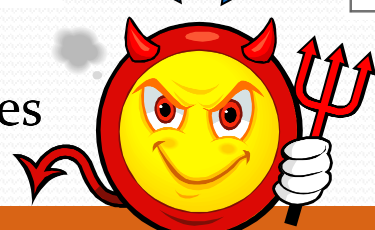
The published table

Age	Sex	Zipcode	Disease
5	M	12000	gastric ulcer
9	M	14000	dyspepsia
6	M	18000	pneumonia
8	M	19000	bronchitis
12	M	22000	pneumonia
19	M	24000	pneumonia
21	F	58000	flu
26	F	36000	gastritis
28	F	37000	pneumonia
56	F	33000	flu

Quasi-identifier (QI) attributes

A voter registration list

Name	Age	Sex	Zipcode
Andy	5	M	12000
Bill	9	M	14000
Ken	6	M	18000
Nash	8	M	19000
Mike	*7*	*M*	*17000*
Joe	12	M	22000
Sam	19	M	24000
Linda	21	F	58000
Jane	26	F	36000
Sarah	28	F	37000
Mary	56	F	33000



"Background knowledge"

87% of Americans can be uniquely identified by {zip code, gender, date of birth}.

actually 63%
[Golle 06]

Latanya Sweeney [*International Journal on Uncertainty, Fuzziness and Knowledge-based Systems*, 2002] used this approach to re-identify the medical record of an ex-governor of Massachusetts.

Outline

- Why privacy?
- Privacy Attacking Examples
- Conventional Principles and limitations
  - K-anonymity
  - L-diversity
  - T-closeness

Real query logs can be very useful to CS researchers. But click history can uniquely identify a person.

<*AnonID, Query, QueryTime, ItemRank, domain name clicked*>

What the New York Times did:

- Find all log entries for AOL user 4417749
- Multiple queries for businesses and services in Lilburn, GA (population 11K)
- Several queries for Jarrett Arnold
  - Lilburn has 14 people with the last name Arnold
- NYT contacts them, finds out AOL User 4417749 is Thelma Arnold

Just because data looks hard to re-identify, doesn't mean it *is*.

[Narayanan and Shmatikov, Oakland 08]

In 2009, the Netflix movie rental service offered a $1,000,000 prize for improving their movie recommendation service.

	High School Musical 1	High School Musical 2	High School Musical 3	Twilight
Customer #1	4	5	5	?

Training data: ~100M ratings of 18K movies from ~500K randomly selected customers, plus dates

Only 10% of their data; slightly perturbed

We can re-identify a Netflix rater if we know just a little bit about her (from life, IMDB ratings, blogs, …).

- 8 movie ratings (≤ 2 wrong, dates ±2 weeks) ➔ re-identify 99% of raters
- 2 ratings, ±3 days ➔ re-identify 68% of raters
  - Relatively few candidates for the other 32% (especially with movies outside the top 100)
- Even a handful of IMDB comments allows Netflix re-identification, in many cases
  - 50 IMDB users → re-identify 2 with very high probability, one from ratings, one from dates

Why should we care about this innocuous data set?

- *All* movie ratings → political and religious opinions, sexual orientation, …
- *Everything* bought in a store → private life details
- *Every* doctor visit → private life details

"One customer … sued Netflix, saying she thought her rental history could reveal that she was a lesbian before she was ready to tell everyone."

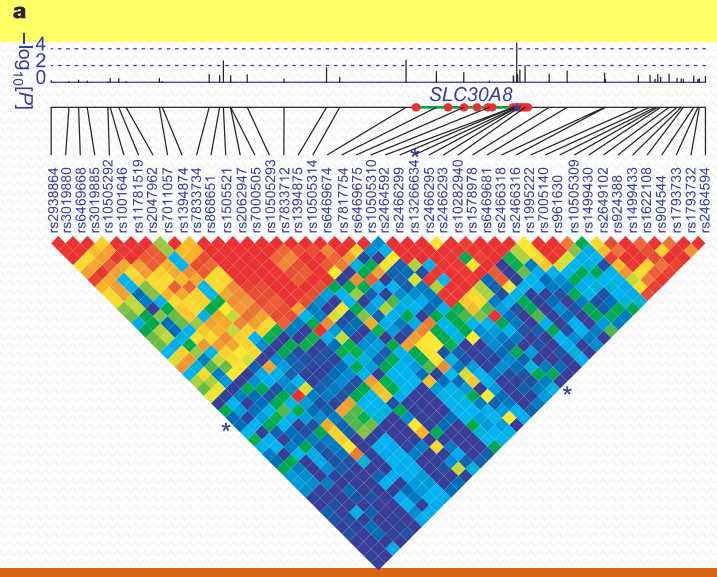
It is becoming routine for medical studies to include a genetic component.

Genome-wide association studies (GWAS) aim to identify the correlation between diseases, e.g., diabetes, and the patient's DNA, by comparing people with and without the disease.

GWAS papers usually include detailed correlation statistics.

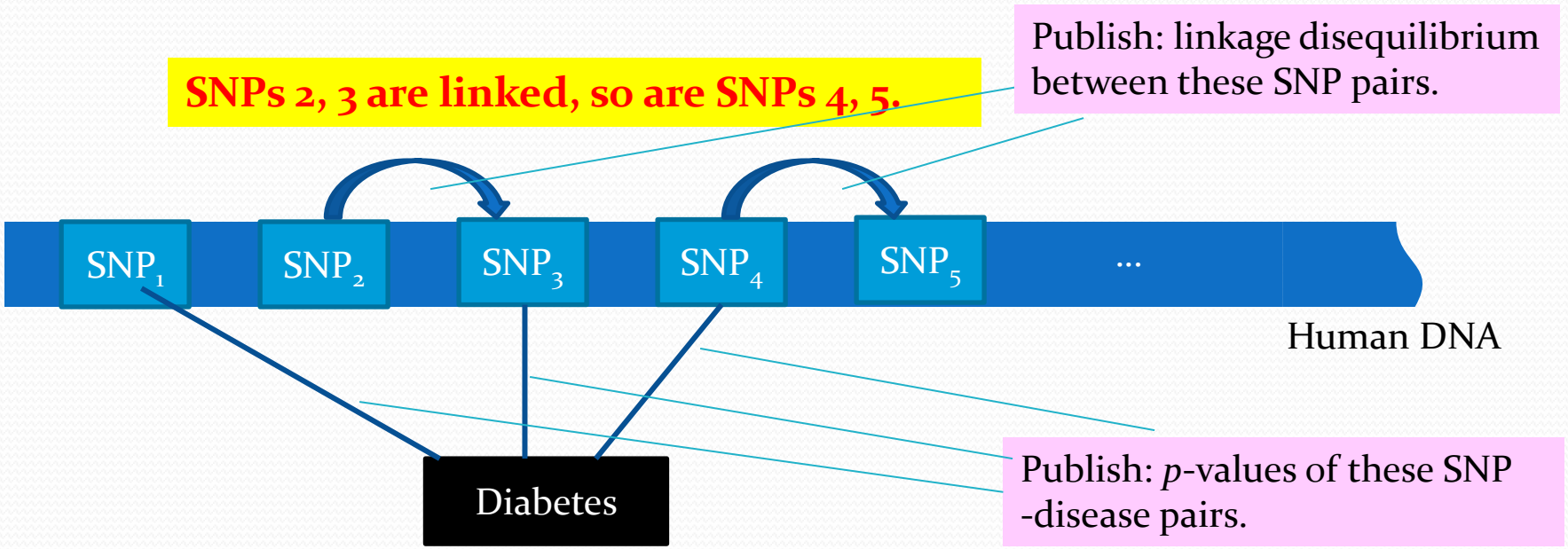
Our attack: uncover the identities of the patients in a GWAS

- For studies of up to moderate size, a significant fraction of people, determine whether a specific person has participated in a particular study within 10 seconds, with high confidence!



A genome-wide association study identifies novel risk loci for type 2 diabetes, Nature 445, 881-885 (22 February 2007)

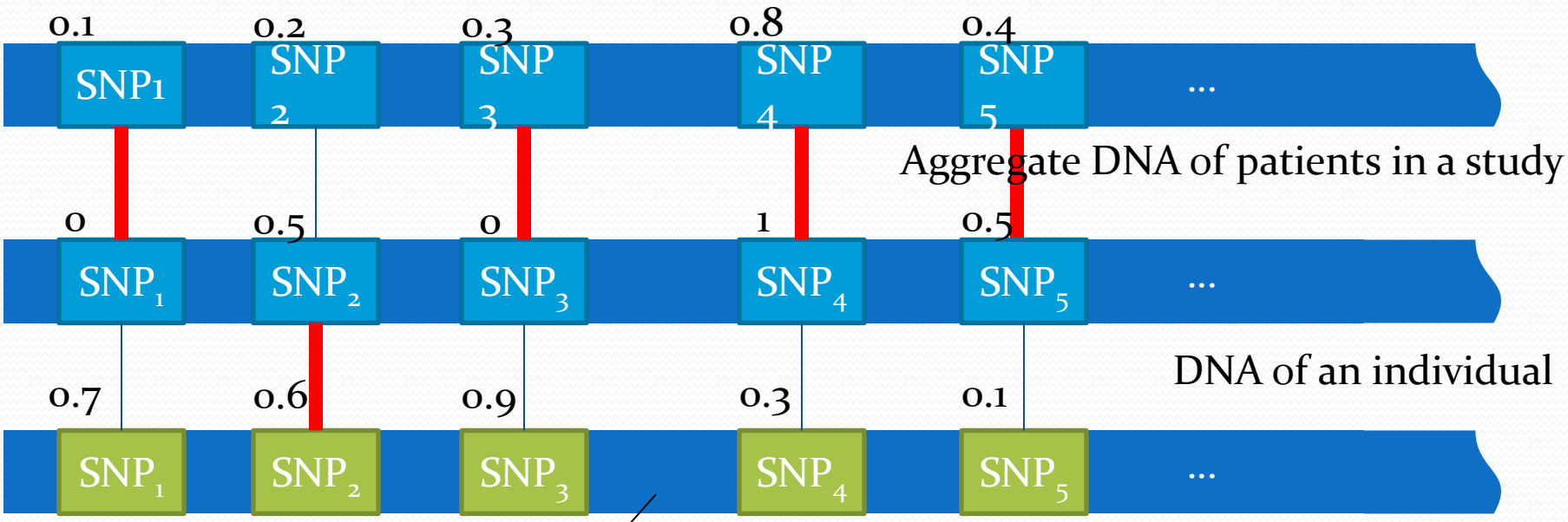
GWAS papers usually include detailed correlation statistics.



Privacy attacks can use SNP-disease association.

Idea [Homer et al. *PloS Genet.*'08, Jacobs et al. *Nature*'09]:

- Obtain aggregate SNP info from the published *p*-values (1)
- Obtain a sample DNA of the target individual (2)
- Obtain the aggregate SNP info of a ref. population (3)
- Compare (1), (2), (3)



Outline

- Why privacy?
- Privacy Attacking Examples
- Conventional Principles and limitations
  - K-anonymity
  - L-diversity
  - T-closeness

Issues

- Privacy principle
  - What is adequate privacy protection?
- Distortion approach
  - How can we achieve the privacy principle,
  - while maximizing the utility of the data?

Different applications may have different privacy protection needs.

Membership disclosure: Attacker cannot tell that a given person is/was in the data set (e.g., a set of AIDS patient records or the summary data from a data set like dbGaP).

- δ-presence [Nergiz et al., 2007].
- Differential privacy [Dwork, 2007].

Sensitive attribute disclosure: Attacker cannot tell that a given person has a certain sensitive attribute.

- *l*-diversity [Machanavajjhala et al., 2006].
- *t*-closeness [Li et al., 2007].

Identity disclosure: Attacker cannot tell which record corresponds to a given person.

- *k*-anonymity [Sweeney, 2002].

Privacy principle 1: *k*-anonymity

your quasi-identifiers are indistinguishable from ≥ *k* other people's.

[Sweeney, *Int'l J. on Uncertainty, Fuzziness and Knowledge-based Systems*, 2002]

2-anonymous generalization:

A voter registration list

Name	Age	Sex	Zipcode
Andy	5	M	12000
Bill	9	M	14000
Ken	6	M	18000
Nash	8	M	19000
Mike	*7*	*M*	*17000*
Joe	12	M	22000
Sam	19	M	24000
Linda	21	F	58000
Jane	26	F	36000
Sarah	28	F	37000
Mary	56	F	33000

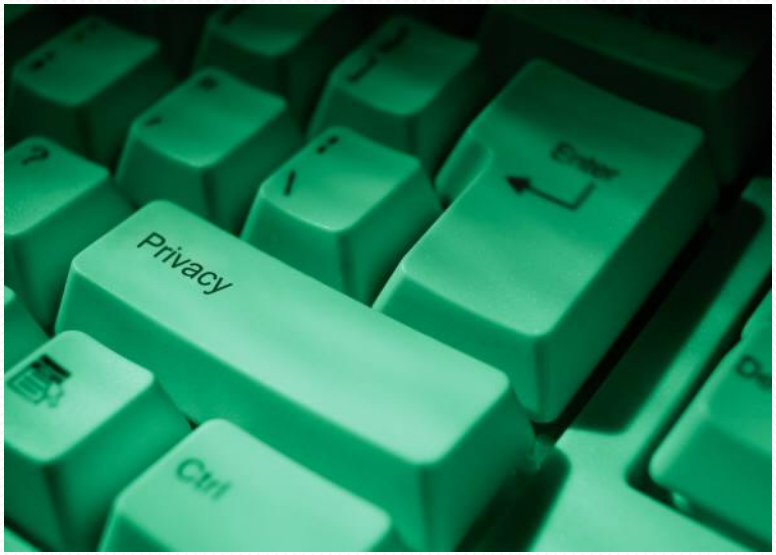
Sensitive attribute

QI attributes

4 QI groups

Age	Sex	Zipcode	Disease
[1, 10]	M	[10001, 15000]	gastric ulcer
[1, 10]	M	[10001, 15000]	dyspepsia
[1, 10]	M	[15001, 20000]	pneumonia
[1, 10]	M	[15001, 20000]	bronchitis
[11, 20]	M	[20001, 25000]	pneumonia
[11, 20]	M	[20001, 25000]	pneumonia
[21, 60]	F	[30000, 60000]	flu
[21, 60]	F	[30000, 60000]	gastritis
[21, 60]	F	[30000, 60000]	pneumonia
[21, 60]	F	[30000, 60000]	flu

The biggest advantage of k-anonymity is that people can understand it.



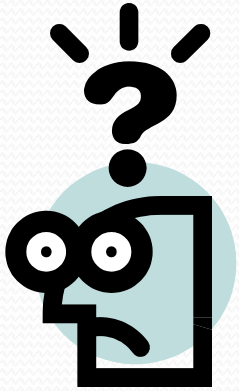
And often it can be computed fast.

But in general, it is easy to attack.

k-anonymity... or how not to define privacy.

[Shmatikov]

- Does not say anything about the computations to be done on the data (utility).
- Assumes that attacker will be able to join only on quasi-identifiers.



Intuitive reasoning:

- *k*-anonymity prevents attacker from telling which record corresponds to which person.
- Therefore, attacker cannot tell that a certain person has a particular value of a sensitive attribute.

This reasoning is fallacious!

k-anonymity does not provide privacy if the sensitive values in an equivalence class lack diversity, or the attacker has certain background knowledge.

From a voter registration list

Homogeneity Attack

Bob	
Zipcode	***Age***
47678	27

Background Knowledge Attack

Carl	
Zipcode	***Age***
47673	36

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

Updates can also destroy k-anonymity.

What is Joe's disease? Wait for his birthday.

A voter registration list
plus dates of birth (not shown)

Name	Age	Sex	Zipcode
Andy	5	M	12000
Bill	9	M	14000
Ken	6	M	18000
Nash	8	M	19000
Mike	*7*	*M*	*17000*
Joe	10	M	17000
Sam	19	M	24000
Linda	21	F	58000
Jane	26	F	36000
Sarah	28	F	37000
Mary	56	F	33000

No "diversity" in this QI group.

Age	Sex	Zipcode	Disease
[1, 10]	M	[10001, 15000]	gastric ulcer
[1, 10]	M	[10001, 15000]	dyspepsia
[1, 10]	M	[15001, 20000]	pneumonia
[1, 10]	M	[15001, 20000]	bronchitis
[11, 20]	M	[20001, 25000]	pneumonia
[11, 20]	M	[20001, 25000]	pneumonia
[21, 60]	F	[30000, 60000]	flu
[21, 60]	F	[30000, 60000]	gastritis
[21, 60]	F	[30000, 60000]	pneumonia
[21, 60]	F	[30000, 60000]	flu

Principle 2: *l*-diversity

[Machanavajjhala et al., *ICDE*, 2006]

Each QI group should have at least *l* "well-represented" sensitive values.

Maybe each QI-group must have *l different* sensitive values?

A 2-diverse table

Age	Sex	Zipcode	Disease
[1, 5]	M	[10001, 15000]	gastric ulcer
[1, 5]	M	[10001, 15000]	dyspepsia
[6, 10]	M	[15001, 20000]	pneumonia
[6, 10]	M	[15001, 20000]	bronchitis
[11, 20]	F	[20001, 25000]	flu
[11, 20]	F	[20001, 25000]	pneumonia
[21, 60]	F	[30001, 60000]	gastritis
[21, 60]	F	[30001, 60000]	gastritis
[21, 60]	F	[30001, 60000]	flu
[21, 60]	F	[30001, 60000]	flu

We can attack this probabilistically.

If we know Joe's QI group, what is the probability he has HIV?

A QI group with 100 tuples

...	Disease
	...
	HIV
	HIV
	...
	HIV
	pneumonia
	bronchitis
	...

98 tuples

The conclusion researchers drew: The most frequent sensitive value in a QI group cannot be too frequent.

Even then, we can still attack using background knowledge.

Joe has HIV.

Sally knows Joe does not have pneumonia.

Sally can guess that Joe has HIV.

A QI group with 100 tuples

...	**Disease**
	...
	HIV
	...
	HIV
	pneumonia
	...
	pneumonia
	bronchitis
	...

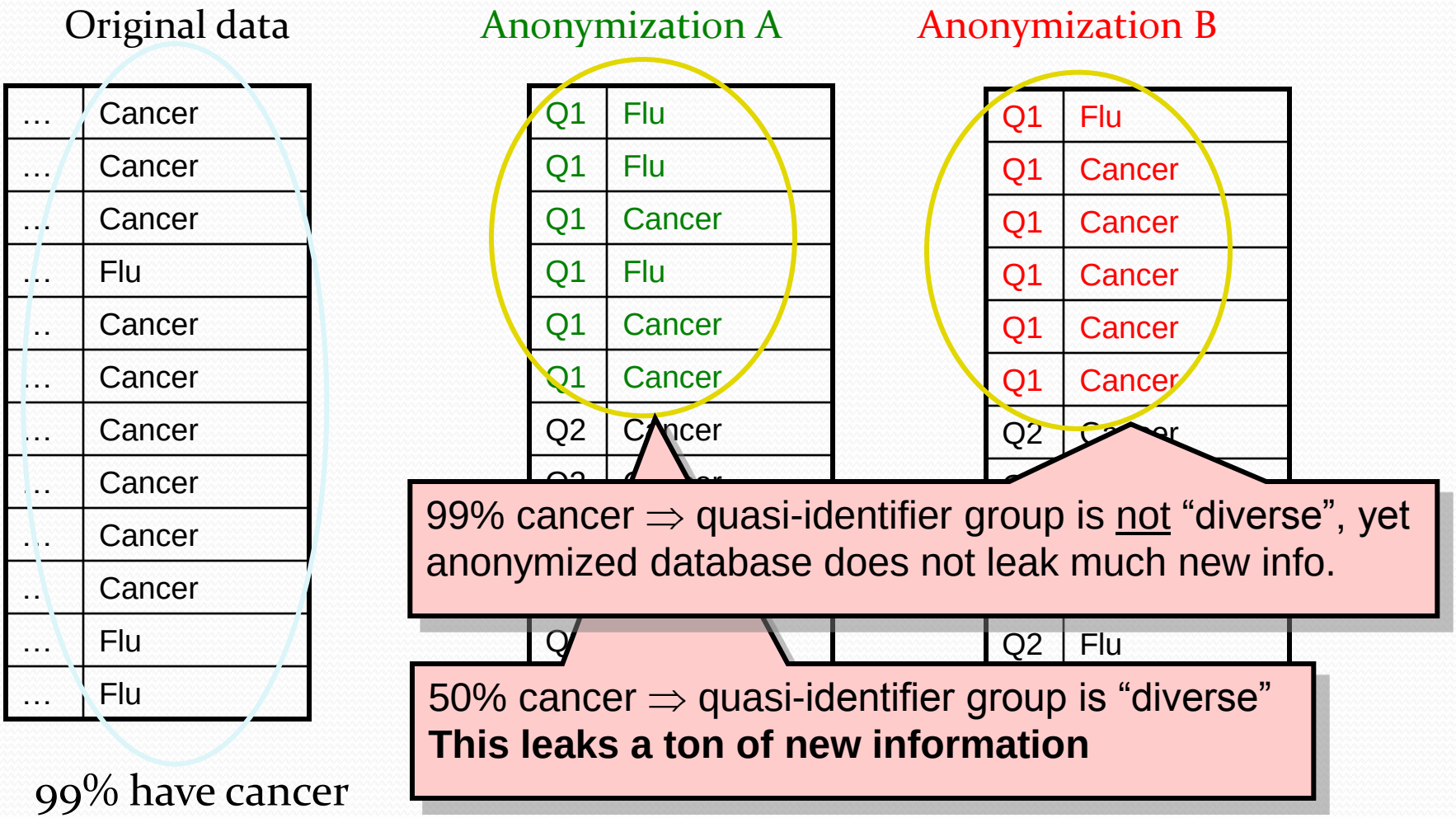
50 tuples

49 tuples

l-diversity variants have been proposed to address these weaknesses.

- 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
  - The most frequent value does not appear too frequently
  - $r_1 < c(r_l + r_{l+1} + \ldots + r_m)$, where r_i is the frequency of the *i*-th most frequent value.

l-diversity can be overkill or underkill.



Diversity does not *inherently* benefit privacy.

Principle 3: 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

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

Then we can bound the knowledge that the attacker gains by seeing a particular anonymization.

Adversarial belief

Belief	Knowledge
B_0	External Knowledge
B_1	Overall distribution of sensitive values
B_2	Distribution of sensitive values in a particular group

Released table

Age	Zip code	.....	Gender	Disease
2*	479**	.....	Male	Flu
2*	479**	.....	Male	Heart Disease
2*	479**	.....	Male	Cancer
.	.	.....	.	.
.	.	.....	.	.
.	.	.....	.	.
≥50	4766*	.....	*	Gastritis

Only applicable when we can define the distance between values, e.g., using a hierarchy of diagnoses.

How anonymous is this 4-anonymous, 3-diverse, and perfectly-*t*-close data?

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

That depends on the attacker's background knowledge.

My coworker Bob's shingles got so bad that he is in the hospital. He looks Asian to me…

This is against the rules, because flu is not a quasi-identifier.

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

In the real world, almost *anything* could be personally identifying (as we saw with Netflix).

There are probably 100 other related proposed privacy principles…

- *k*-gather, (*a*, *k*)-anonymity, personalized anonymity, positive disclosure-recursive (*c*, *l*)-diversity, non-positive-disclosure (*c1*, *c2*, *l*)-diversity, *m*-invariance, (*c*, *t*)-isolation, …

And for other data models, e.g., graphs:

- *k*-degree anonymity, *k*-neighborhood anonymity, *k*-sized grouping, (*k*, *l*) grouping, …

… and they suffer from related problems. [Shmatikov]

~~Trying to achieve "privacy" by syntactic transformation of the data~~

- ~~Scrubbing of PII, k-anonymity, l-diversity…~~

Fatally flawed!

- Insecure against attackers with arbitrary background info
- Do not compose (anonymize twice ⇒ reveal data)
- No meaningful notion of privacy
- No meaningful notion of utility

Does he go too far?

And there is an impossibility result that applies to all of them.

[Dwork, Naor 2006]

For any reasonable definition of "privacy breach" and "sanitization", with high probability some adversary can breach some sanitized DB.

Example:

- Private fact: my exact height
- Background knowledge: I'm 5 inches taller than the average American woman
- San(DB) allows computing average height of US women
- This breaks my privacy … even if my record is not in the database!

DATABASE PRIVACY - II

Formulation of Privacy

- What information can be published?
  - Average height of US people 
  - Height of an individual 
- Intuition:
  - If something is insensitive to the change of any individual tuple, then it should not be considered private
- Example:
  - Assume that we arbitrarily change the height of an individual in the US
  - The average height of US people would remain roughly the same
  - i.e., The average height reveals little information about the exact height of any particular individual

ε-Differential Privacy

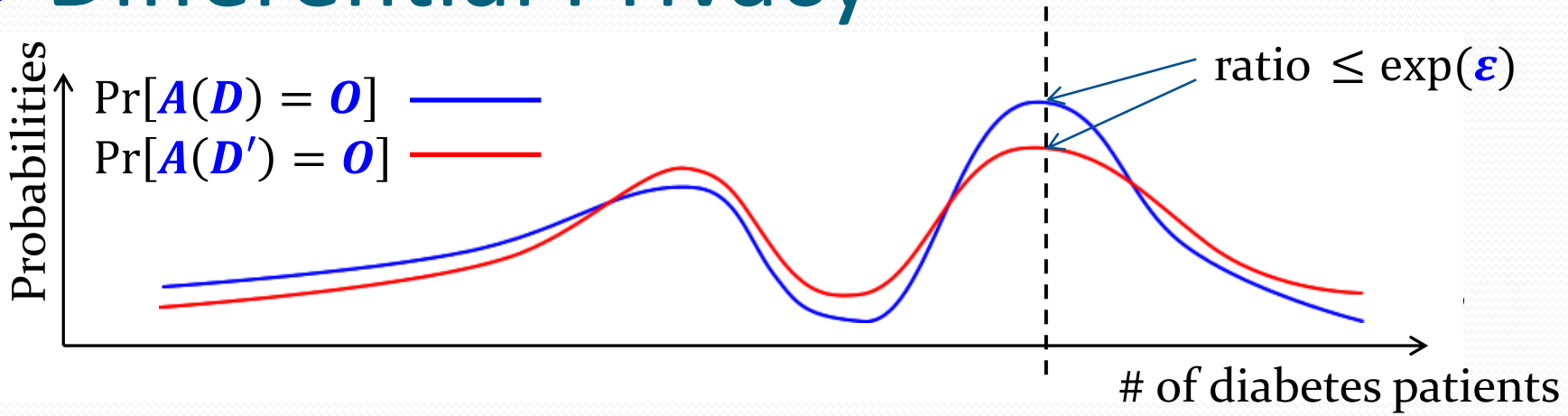
- Definition:
  - Neighboring datasets: Two datasets D and D', such that D' can be obtained by changing one single tuple in D
  - A randomized algorithm A satisfies ε-differential privacy, iff for any two neighboring datasets D and D' and for any output O of A,

$$\Pr[A(D) = O] \leq \exp(\varepsilon) \cdot \Pr[A(D') = O]$$

Name	Gender	Age	Diabetes
Alice	F	28	Y
Bob	M	19	Y
Chris	M	25	N
Doug	M	30	N

Name	Gender	Age	Diabetes
Alice	F	28	Y
Bob	M	19	Y
Chris	M	23	Y
Doug	M	30	N

ε-Differential Privacy



- Definition:
  - Neighboring datasets: Two datasets D and D', such that D' can be obtained by changing one single tuple in D
  - A randomized algorithm A satisfies ε-differential privacy, iff for any two neighboring datasets D and D' and for any output O of A,

$$\Pr[A(D) = O] \leq \exp(\varepsilon) \cdot \Pr[A(D') = O]$$

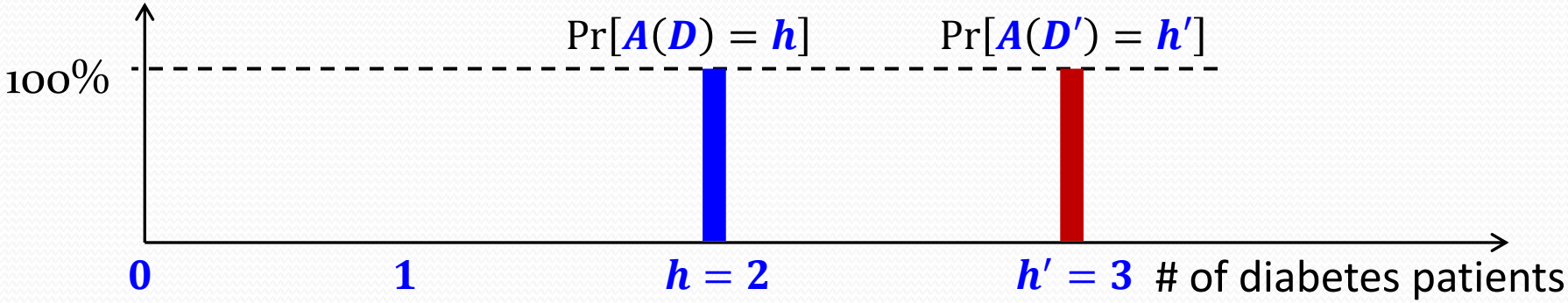
  - The value of ε decides the degree of privacy protection

Achieving ε-Differential Privacy

Name	Gender	Age	Diabetes
Alice	F	28	Y
Bob	M	19	Y
Chris	M	25	N
Doug	M	30	N

Name	Gender	Age	Diabetes
Alice	F	28	Y
Bob	M	19	Y
Chris	M	23	Y
Doug	M	30	N

- It won't work if we release the number directly:
  - $\boldsymbol{D}$: the original dataset
  - $\boldsymbol{D'}$: modify an arbitrary patient in $\boldsymbol{D}$
  - $\Pr[\boldsymbol{A}(\boldsymbol{D}) = \boldsymbol{O}] \leq \exp(\boldsymbol{\varepsilon}) \cdot \Pr[\boldsymbol{A}(\boldsymbol{D'}) = \boldsymbol{O}]$ does not hold for any $\boldsymbol{\varepsilon}$

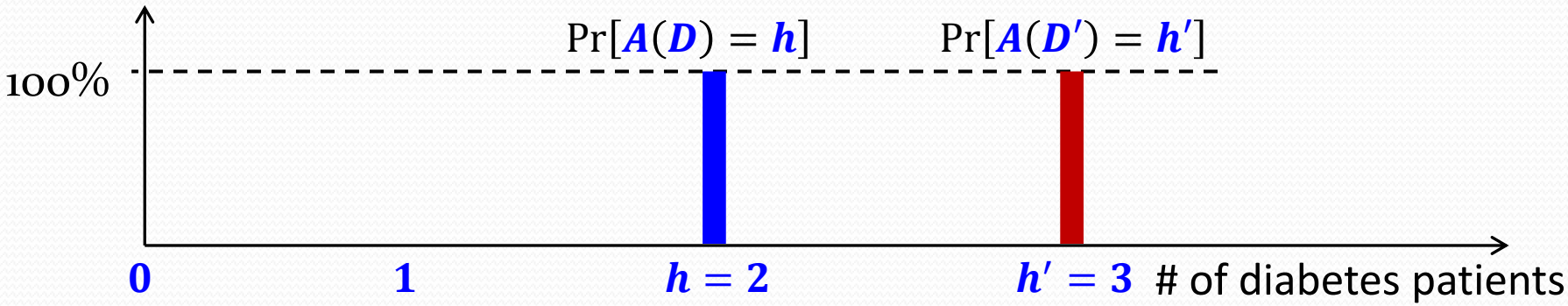


Achieving ε-Differential Privacy

Name	Gender	Age	Diabetes
Alice	F	28	Y
Bob	M	19	Y
Chris	M	25	N
Doug	M	30	N

Name	Gender	Age	Diabetes
Alice	F	28	Y
Bob	M	19	Y
Chris	M	23	Y
Doug	M	30	N

- Idea:
  - Perturb the number of diabetes patients to obtain a smooth distribution

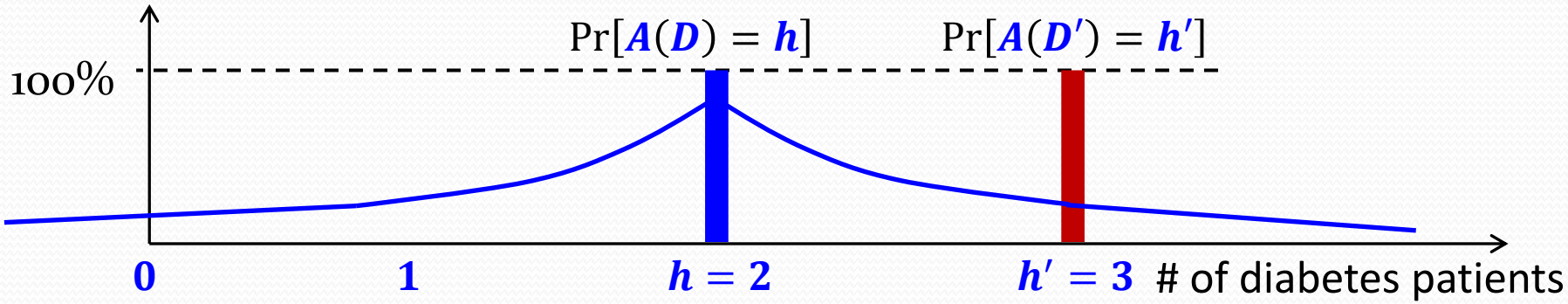


Achieving ε-Differential Privacy

Name	Gender	Age	Diabetes
Alice	F	28	Y
Bob	M	19	Y
Chris	M	25	N
Doug	M	30	N

Name	Gender	Age	Diabetes
Alice	F	28	Y
Bob	M	19	Y
Chris	M	23	Y
Doug	M	30	N

- Idea:
  - Perturb the number of diabetes patients to obtain a smooth distribution

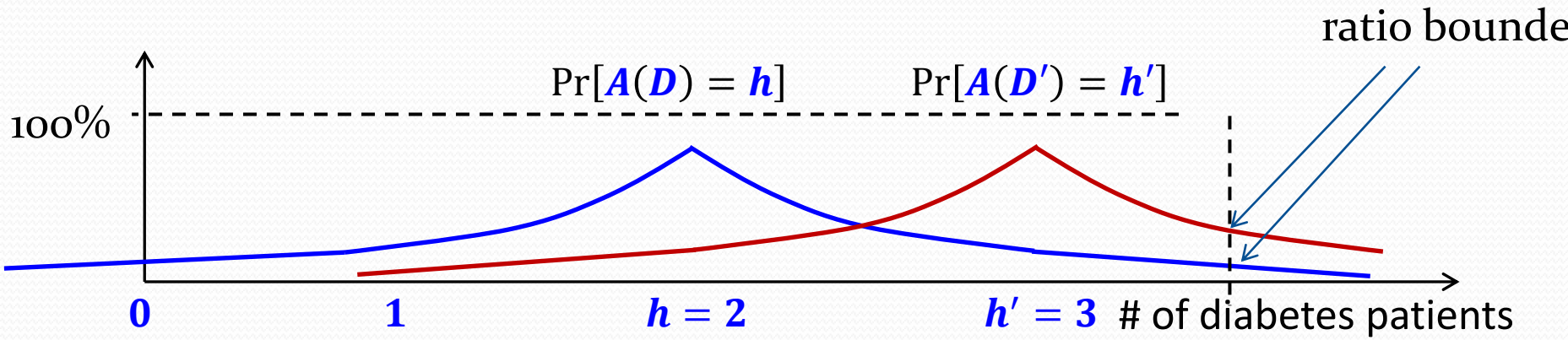


Achieving ε-Differential Privacy

Name	Gender	Age	Diabetes
Alice	F	28	Y
Bob	M	19	Y
Chris	M	25	N
Doug	M	30	N

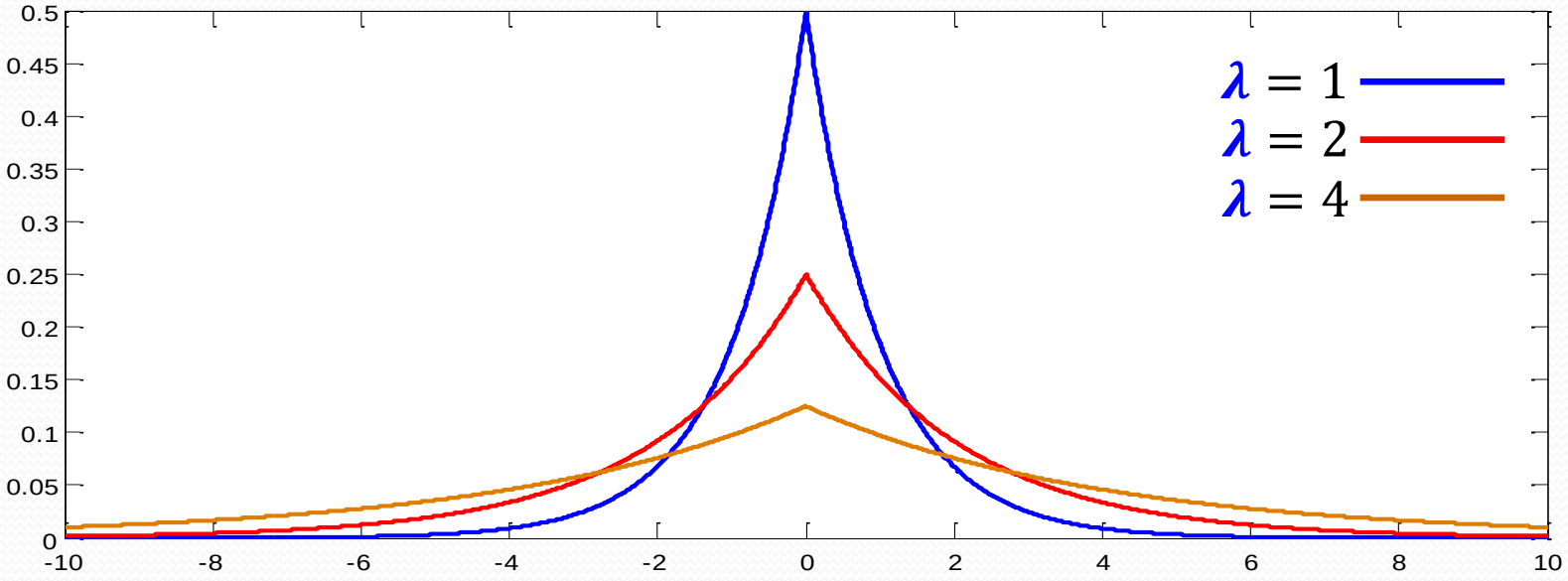
Name	Gender	Age	Diabetes
Alice	F	28	Y
Bob	M	19	Y
Chris	M	23	Y
Doug	M	30	N

- Idea:
  - Perturb the number of diabetes patients to obtain a smooth distribution



Laplace Distribution

- $pdf(\boldsymbol{x}) = \exp\left(-\frac{|\boldsymbol{x}|}{\boldsymbol{\lambda}}\right)/2\boldsymbol{\lambda}$;
- increase/decrease $\boldsymbol{x}$ by 1
- → $pdf(\boldsymbol{x})$ changes by a factor of $\exp\left(-\frac{1}{\boldsymbol{\lambda}}\right)$
- $\boldsymbol{\lambda}$ is referred as the *scale*



Differential Privacy via Laplace Noise

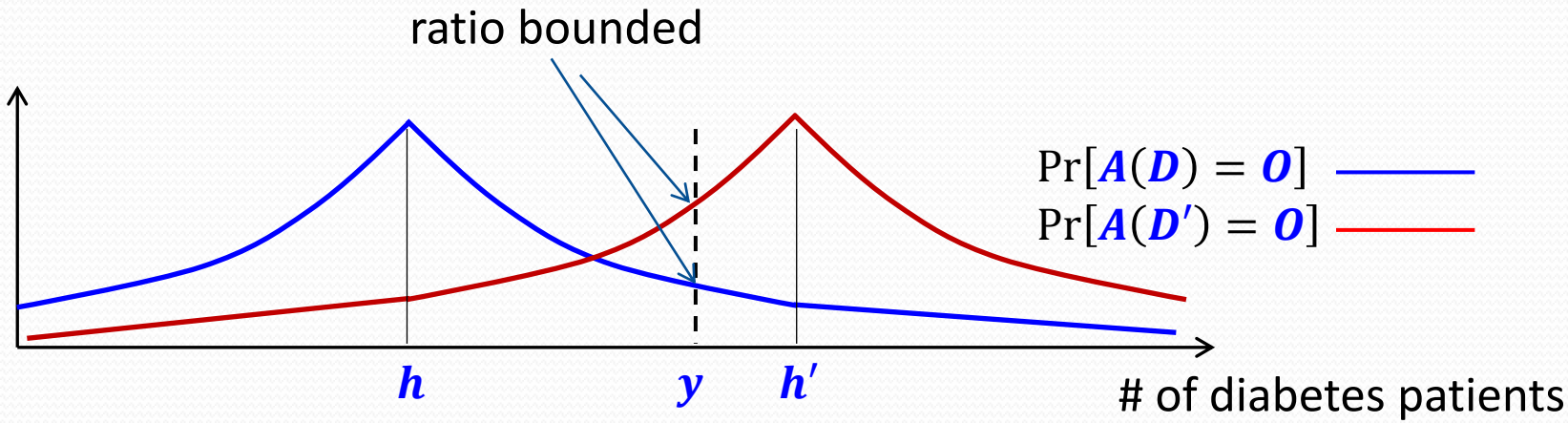
- Dataset: A set of patients
- Objective: Release # of diabetes patients with $\boldsymbol{\varepsilon}$-differential privacy

$$\Pr[\boldsymbol{A}(\boldsymbol{D}) = \boldsymbol{O}] \leq \exp(\boldsymbol{\varepsilon}) \cdot \Pr[\boldsymbol{A}(\boldsymbol{D}') = \boldsymbol{O}]$$

- Method: Release the number + Laplace noise

$$pdf(\boldsymbol{x}) = \exp\left(-\frac{|\boldsymbol{x}|}{\boldsymbol{\lambda}}\right)/2\boldsymbol{\lambda}$$

- Rationale:
  - $\boldsymbol{D}$: the original dataset; # of diabetes patients = $\boldsymbol{h}$
  - $\boldsymbol{D}'$: modify a patient in $\boldsymbol{D}$; # of diabetes patients = $\boldsymbol{h}'$



Differential Privacy via Laplace Noise

- Dataset: A set of patients
- Objective: Release # of diabetes patients with $\boldsymbol{\varepsilon}$-differential privacy

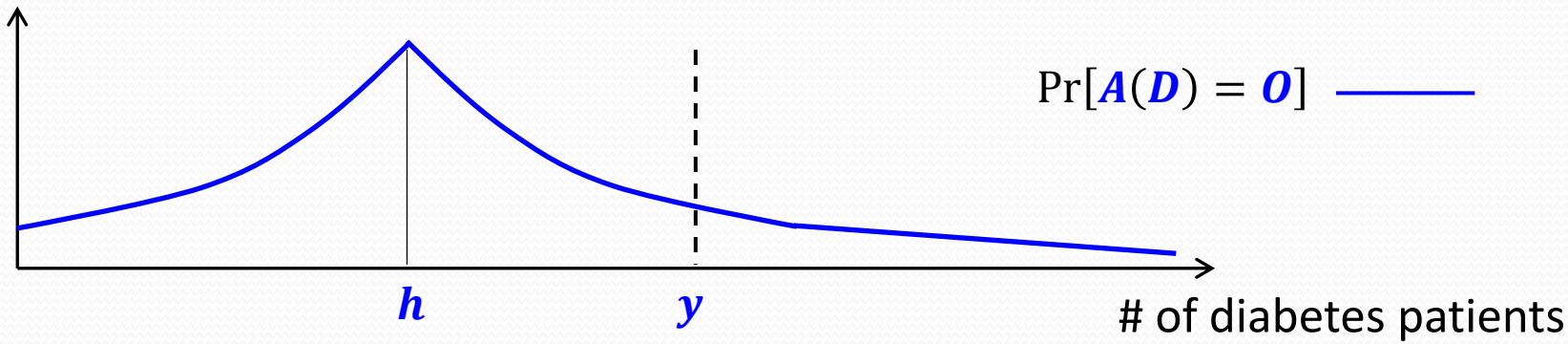
$$\Pr[\boldsymbol{A}(\boldsymbol{D}) = \boldsymbol{O}] \leq \exp(\boldsymbol{\varepsilon}) \cdot \Pr[\boldsymbol{A}(\boldsymbol{D}') = \boldsymbol{O}]$$

- Method: Release the number + Laplace noise

$$pdf(\boldsymbol{x}) = \exp\left(-\frac{|\boldsymbol{x}|}{\boldsymbol{\lambda}}\right)/2\boldsymbol{\lambda}$$

- Rationale:
  - $\boldsymbol{D}$: the original dataset; # of diabetes patients = $\boldsymbol{h}$
  - $\boldsymbol{D}'$: modify a patient in $\boldsymbol{D}$; # of diabetes patients = $\boldsymbol{h}'$

$$\Pr[\boldsymbol{A}(\boldsymbol{D}) = \boldsymbol{y}] = pdf(\boldsymbol{y} - \boldsymbol{h}) = \exp(-|\boldsymbol{y} - \boldsymbol{h}|/\boldsymbol{\lambda})/2\boldsymbol{\lambda}$$



Differential Privacy via Laplace Noise

- Dataset: A set of patients
- Objective: Release # of diabetes patients with $\boldsymbol{\varepsilon}$-differential privacy

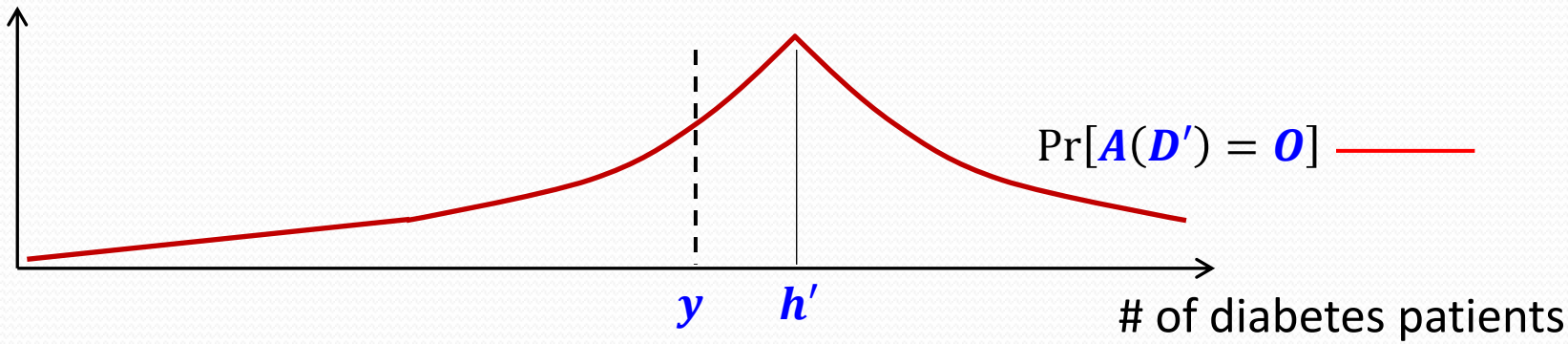
$$\Pr[\boldsymbol{A}(\boldsymbol{D}) = \boldsymbol{O}] \leq \exp(\boldsymbol{\varepsilon}) \cdot \Pr[\boldsymbol{A}(\boldsymbol{D}') = \boldsymbol{O}]$$

- Method: Release the number + Laplace noise

$$pdf(\boldsymbol{x}) = \exp\left(-\frac{|\boldsymbol{x}|}{\boldsymbol{\lambda}}\right)/2\boldsymbol{\lambda}$$

- Rationale:
  - $\boldsymbol{D}$: the original dataset; # of diabetes patients = $\boldsymbol{h}$
  - $\boldsymbol{D}'$: modify the height of an individual in $\boldsymbol{D}$; # of diabetes patients = $\boldsymbol{h}'$

$$\Pr[\boldsymbol{A}(\boldsymbol{D}') = \boldsymbol{y}] = pdf(\boldsymbol{y} - \boldsymbol{h}') = \exp(-|\boldsymbol{y} - \boldsymbol{h}'|/\boldsymbol{\lambda})/2\boldsymbol{\lambda}$$



Differential Privacy via Laplace Noise

- Dataset: A set of patients
- Objective: Release # of diabetes patients with $\boldsymbol{\varepsilon}$-differential privacy

$$\Pr[\boldsymbol{A}(\boldsymbol{D}) = \boldsymbol{O}] \leq \exp(\boldsymbol{\varepsilon}) \cdot \Pr[\boldsymbol{A}(\boldsymbol{D}') = \boldsymbol{O}]$$

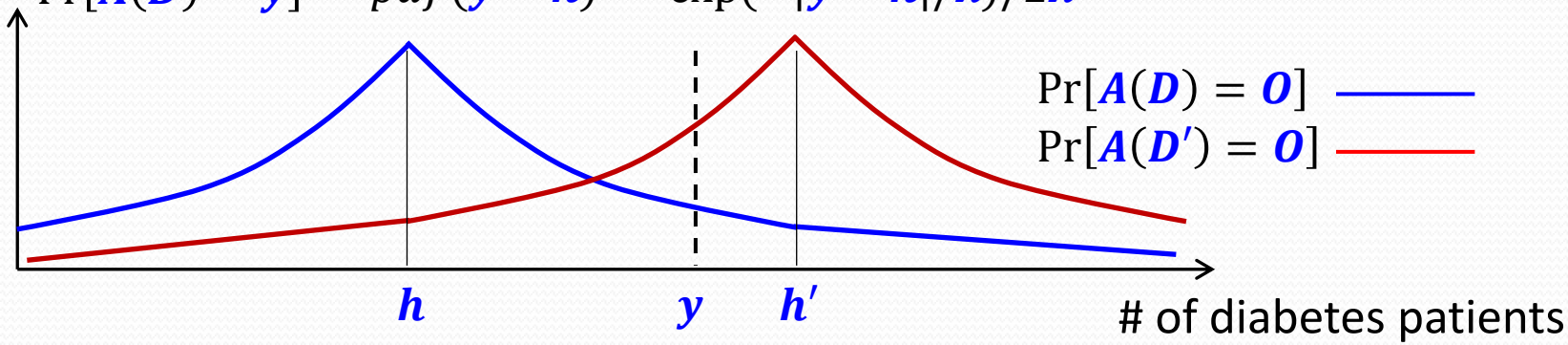
- Method: Release the number + Laplace noise

$$pdf(\boldsymbol{x}) = \exp\left(-\frac{|\boldsymbol{x}|}{\boldsymbol{\lambda}}\right)/2\boldsymbol{\lambda}$$

- Rationale:
  - $\boldsymbol{D}$: the original dataset; # of diabetes patients = $\boldsymbol{h}$
  - $\boldsymbol{D}'$: modify the height of an individual in $\boldsymbol{D}$; # of diabetes patients = $\boldsymbol{h}'$

$$\Pr[\boldsymbol{A}(\boldsymbol{D}') = \boldsymbol{y}] = pdf(\boldsymbol{y} - \boldsymbol{h}') = \exp(-|\boldsymbol{y} - \boldsymbol{h}'|/\boldsymbol{\lambda})/2\boldsymbol{\lambda}$$

$$\Pr[\boldsymbol{A}(\boldsymbol{D}) = \boldsymbol{y}] = pdf(\boldsymbol{y} - \boldsymbol{h}) = \exp(-|\boldsymbol{y} - \boldsymbol{h}|/\boldsymbol{\lambda})/2\boldsymbol{\lambda}$$



Differential Privacy via Laplace Noise

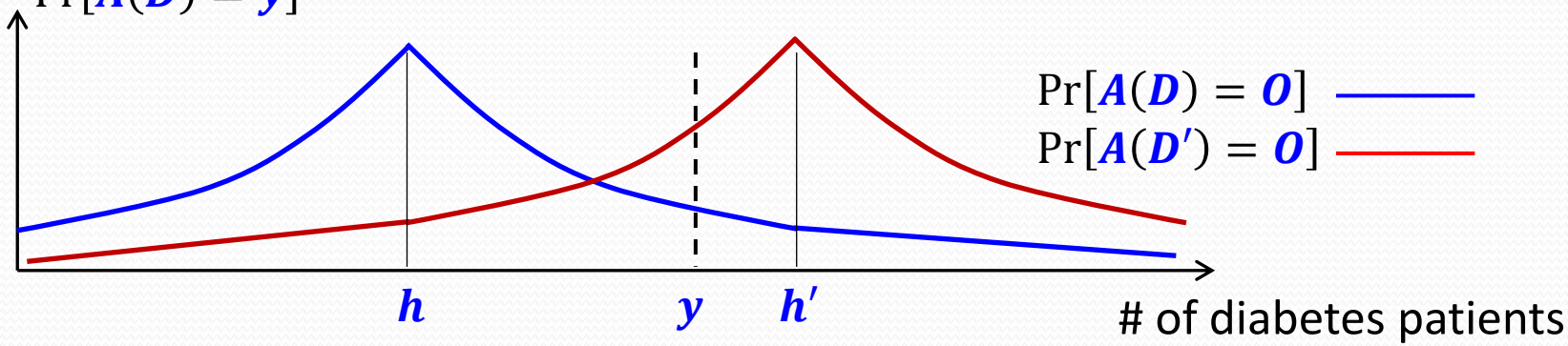
- Dataset: A set of patients
- Objective: Release # of diabetes patients with $\boldsymbol{\varepsilon}$-differential privacy

$$\Pr[\boldsymbol{A}(\boldsymbol{D}) = \boldsymbol{O}] \leq \exp(\boldsymbol{\varepsilon}) \cdot \Pr[\boldsymbol{A}(\boldsymbol{D}') = \boldsymbol{O}]$$

- Method: Release the number + Laplace noise

$$pdf(\boldsymbol{x}) = \exp\left(-\frac{|\boldsymbol{x}|}{\boldsymbol{\lambda}}\right)/2\boldsymbol{\lambda}$$

- Rationale:
  - $\boldsymbol{D}$: the original dataset; # of diabetes patients = $\boldsymbol{h}$
  - $\boldsymbol{D}'$: modify the height of an individual in $\boldsymbol{D}$; # of diabetes patients = $\boldsymbol{h}'$



Differential Privacy via Laplace Noise

- Dataset: A set of patients
- Objective: Release # of diabetes patients with $\boldsymbol{\varepsilon}$-differential privacy

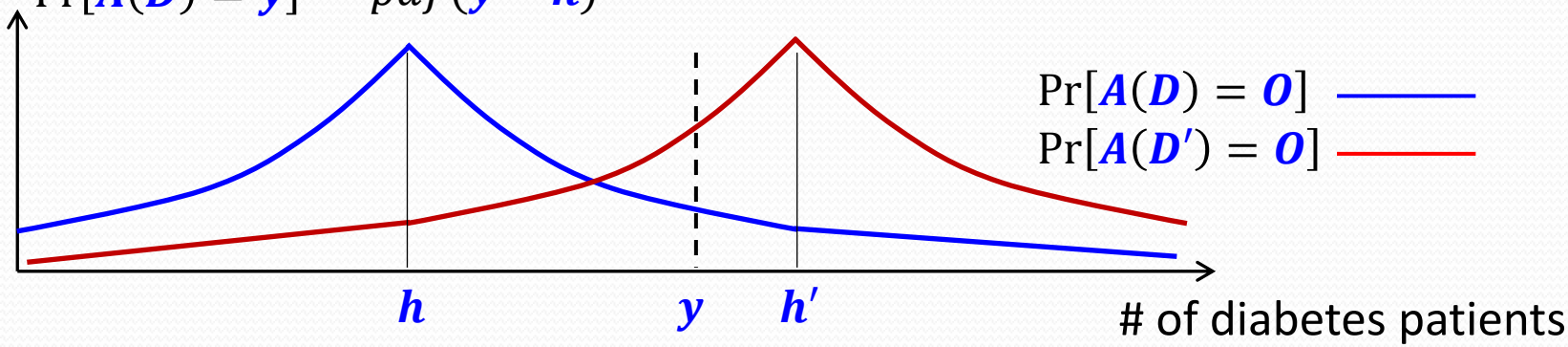
$$\Pr[\boldsymbol{A}(\boldsymbol{D}) = \boldsymbol{O}] \leq \exp(\boldsymbol{\varepsilon}) \cdot \Pr[\boldsymbol{A}(\boldsymbol{D}') = \boldsymbol{O}]$$

- Method: Release the number + Laplace noise

$$pdf(\boldsymbol{x}) = \exp\left(-\frac{|\boldsymbol{x}|}{\boldsymbol{\lambda}}\right)/2\boldsymbol{\lambda}$$

- Rationale:
  - $\boldsymbol{D}$: the original dataset; # of diabetes patients = $\boldsymbol{h}$
  - $\boldsymbol{D}'$: modify the height of an individual in $\boldsymbol{D}$; # of diabetes patients = $\boldsymbol{h}'$

$$\frac{\Pr[\boldsymbol{A}(\boldsymbol{D}') = \boldsymbol{y}]}{\Pr[\boldsymbol{A}(\boldsymbol{D}) = \boldsymbol{y}]} = \frac{pdf(\boldsymbol{y} - \boldsymbol{h}')}{pdf(\boldsymbol{y} - \boldsymbol{h})}$$



Differential Privacy via Laplace Noise

- Dataset: A set of patients
- Objective: Release # of diabetes patients with $\boldsymbol{\varepsilon}$-differential privacy

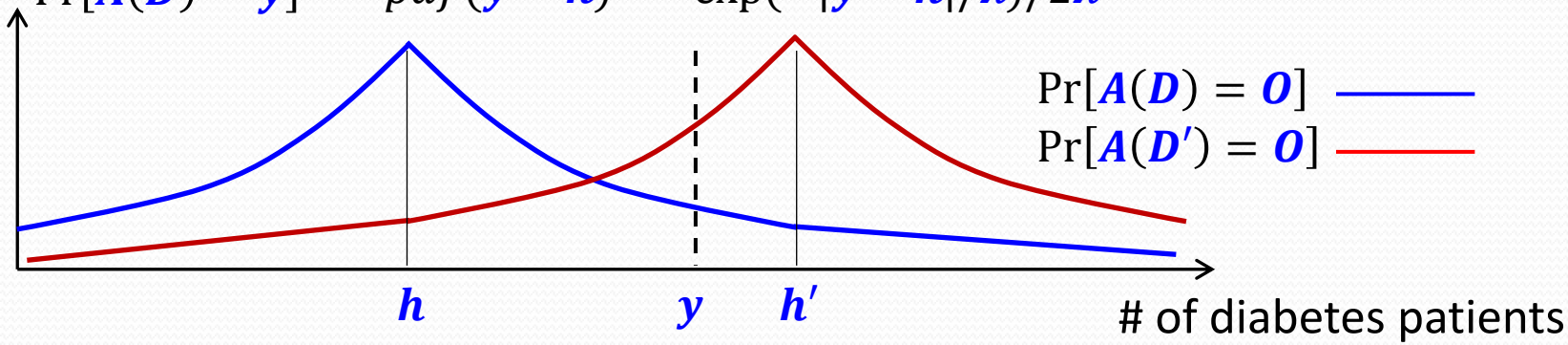
$$\Pr[\boldsymbol{A}(\boldsymbol{D}) = \boldsymbol{O}] \leq \exp(\boldsymbol{\varepsilon}) \cdot \Pr[\boldsymbol{A}(\boldsymbol{D}') = \boldsymbol{O}]$$

- Method: Release the number + Laplace noise

$$pdf(\boldsymbol{x}) = \exp\left(-\frac{|\boldsymbol{x}|}{\boldsymbol{\lambda}}\right)/2\boldsymbol{\lambda}$$

- Rationale:
  - $\boldsymbol{D}$: the original dataset; # of diabetes patients = $\boldsymbol{h}$
  - $\boldsymbol{D}'$: modify the height of an individual in $\boldsymbol{D}$; # of diabetes patients = $\boldsymbol{h}'$

$$\frac{\Pr[\boldsymbol{A}(\boldsymbol{D}') = \boldsymbol{y}]}{\Pr[\boldsymbol{A}(\boldsymbol{D}) = \boldsymbol{y}]} = \frac{pdf(\boldsymbol{y} - \boldsymbol{h}')}{pdf(\boldsymbol{y} - \boldsymbol{h})} = \frac{\exp(-|\boldsymbol{y} - \boldsymbol{h}'|/\boldsymbol{\lambda})/2\boldsymbol{\lambda}}{\exp(-|\boldsymbol{y} - \boldsymbol{h}|/\boldsymbol{\lambda})/2\boldsymbol{\lambda}}$$



Differential Privacy via Laplace Noise

- Dataset: A set of patients
- Objective: Release # of diabetes patients with $\boldsymbol{\varepsilon}$-differential privacy

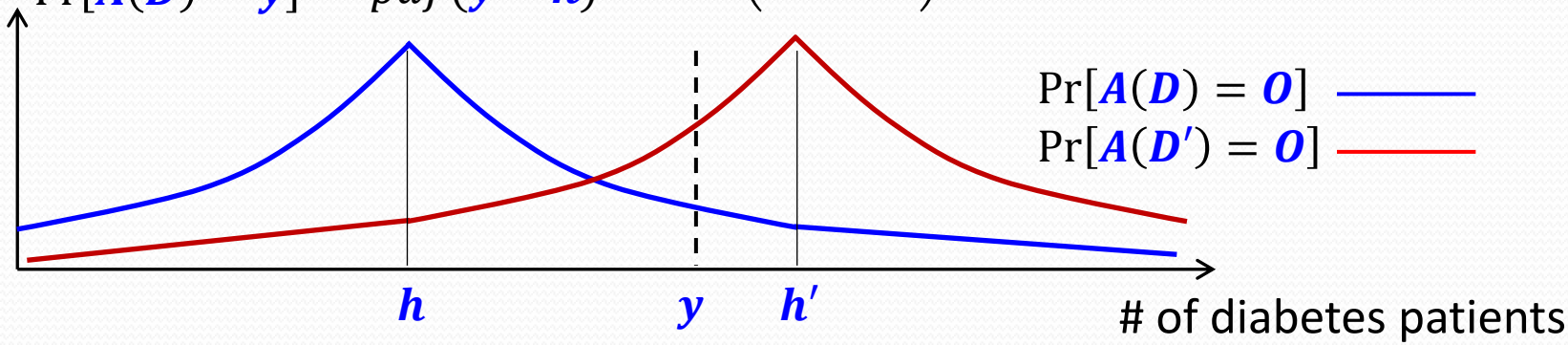
$$\Pr[\boldsymbol{A}(\boldsymbol{D}) = \boldsymbol{O}] \leq \exp(\boldsymbol{\varepsilon}) \cdot \Pr[\boldsymbol{A}(\boldsymbol{D}') = \boldsymbol{O}]$$

- Method: Release the number + Laplace noise

$$pdf(\boldsymbol{x}) = \exp\left(-\frac{|\boldsymbol{x}|}{\boldsymbol{\lambda}}\right)/2\boldsymbol{\lambda}$$

- Rationale:
  - $\boldsymbol{D}$: the original dataset; # of diabetes patients = $\boldsymbol{h}$
  - $\boldsymbol{D}'$: modify the height of an individual in $\boldsymbol{D}$; # of diabetes patients = $\boldsymbol{h}'$

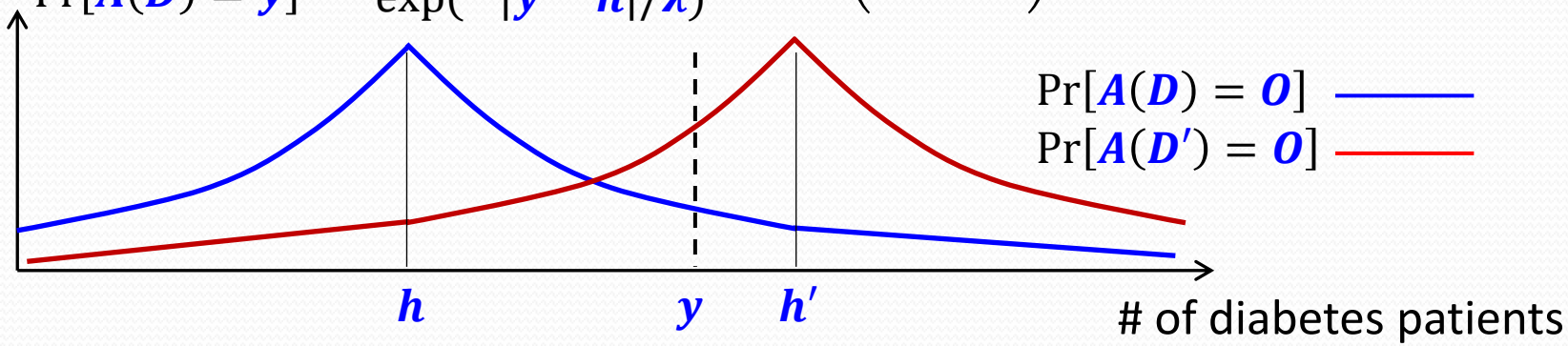
$$\frac{\Pr[\boldsymbol{A}(\boldsymbol{D}') = \boldsymbol{y}]}{\Pr[\boldsymbol{A}(\boldsymbol{D}) = \boldsymbol{y}]} = \frac{pdf(\boldsymbol{y} - \boldsymbol{h}')}{pdf(\boldsymbol{y} - \boldsymbol{h})} \leq \exp\left(\frac{|\boldsymbol{h} - \boldsymbol{h}'|}{\boldsymbol{\lambda}}\right)$$



Differential Privacy via Laplace Noise

- We aim to ensure $\boldsymbol{\varepsilon}$-differential privacy
- How large should $\boldsymbol{\lambda}$ be?
  - Change of a patient's data would change the number of diabetes patients by at most 1, i.e.,
- Conclusion: Setting $\lambda \geq \frac{|h - h'|}{\varepsilon}$ would ensure $\boldsymbol{\varepsilon}$-differential privacy

$$\frac{\Pr[A(D') = y]}{\Pr[A(D) = y]} = \frac{\exp(-|y - h'|/\lambda)}{\exp(-|y - h|/\lambda)} \leq \exp\left(\frac{|h - h'|}{\lambda}\right)$$



General Mechanism with Laplace Noise

- In general, if the query result v is a real number
  - Add Laplace noise into v
- To decide the scale λ of Laplace noise
  - Look at the maixmum change that can occur in v (when we change one tuple in the dataset)
  - Set λ to be proportional to the maximum change

Name	Gender	Age	Diabetes
Alice	F	28	Y
Bob	M	19	Y
Chris	M	25	N
Doug	M	30	N

Name	Gender	Age	Diabetes
Alice	F	28	Y
Bob	M	19	Y
Chris	M	23	Y
Doug	M	30	N

General via Laplace Noise

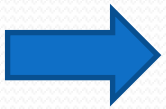
- What if we have multiple queries?
  - Add Laplace noise to each value
- How do we decide the noise scale?
  - Look at the *total change* that can occur in the values when we modify one tuple in the data
  - Total change: sum of the absolute change in each value (i.e., differences in L1 norm)
  - Set the scale of the noise to be proportional to the maximum total change
- The maximum total change is referred to as the *sensitivity* of the values
- Theorem [Dwork et al. 2006]: Adding Laplace noise of scale λ to each value ensures ε-differential privacy, if

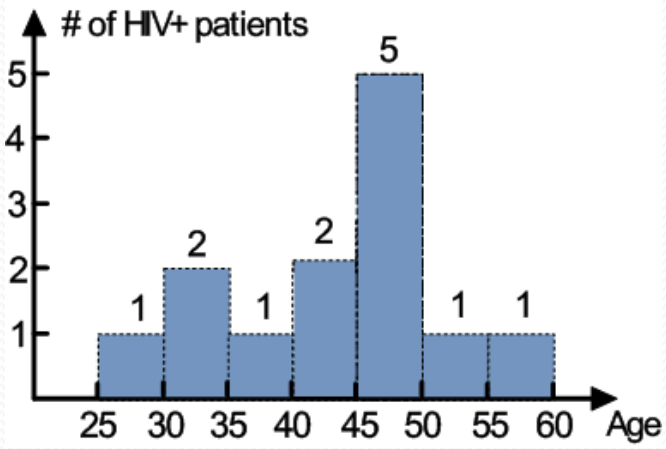
  $\lambda \geq$ (the sensitivity of the values)$/\varepsilon$

Sensitivity of Queries

- Histogram
  - Sensitivity of the bin counts: 2
  - Reason: When we modify a tuple in the dataset, at most two bin counts would change; furthermore, each bin count would change by at most 1
  - Scale of Laplace noise required:

Name	Age	HIV+
Frank	42	Y
Bob	31	Y
Mary	28	Y
Dave	43	N
…	…	…





- For more complex queries, the derivation of sensitivity can be much more complicated
  - Example: Parameters of a logistic model

Exponential Mechanism

- What if the query result is on discrete space?
  - Example: Which one is a more important factor to diabetic, age or gender?
- Given k items, each item is associated with a score $S(I, D)$, how to pick the one with maximal score under differential privacy?
- Adding Laplace noise is a feasible solution

Name	Gender	Age	Diabetes
Alice	F	28	Y
Bob	M	19	Y
Chris	M	25	N
Doug	M	30	N

S(*Gender*,D)=Corr(*Gender*,*Diabetes*)

S(*Age*,D)=Corr(*Age*,*Diabetes*)

Exponential Mechanism

- Using exponential mechanism, we can directly manipulate the probability of item pickup.
- For each item I_j, the probability is proportional to $\exp(S(I, D)/\lambda)$

S(*Gender*,D) = Corr(*Gender*,*Diabetes*) = 0.5

S(*Age*,D) = Corr(*Age*,*Diabetes*) = 0.3

Name	Gender	Age	Diabetes
Alice	F	28	Y
Bob	M	19	Y
Chris	M	25	N
Doug	M	30	N

Pr(*Gender*)=0.71

Pr(*Age*)=0.39

Exponential Mechanism

- Advantage: Improve skewedness on the probabilities
- Limitation: Needs to iterate all possible answers in the solution space. It is thus not applicable when the solution space is too large.
- Example: Pick up the best order of k items with maximal score. The number of possible orders is $k!$.

Variants of Differential Privacy

- Alternative definition of neighboring dataset:
  - Two datasets $\boldsymbol{D}$ and $\boldsymbol{D'}$, such that $\boldsymbol{D'}$ is obtained by adding/deleting one tuple in $\boldsymbol{D}$
- $\Pr[\boldsymbol{A}(\boldsymbol{D}) = \boldsymbol{O}] \leq \exp(\boldsymbol{\varepsilon}) \cdot \Pr[\boldsymbol{A}(\boldsymbol{D'}) = \boldsymbol{O}]$
  - Even if a tuple is added to or removed from the dataset, the output distribution of the algorithm is roughly the same
  - i.e., the output of the algorithm does not reveal the presence of a tuple
- Refer to this version as “unbounded” differential privacy, and the previous version as “bounded” differential privacy

Variants of Differential Privacy

- Bounded: $\boldsymbol{D'}$ is obtained by changing the values of one tuple in $\boldsymbol{D}$
- Unbounded: $\boldsymbol{D'}$ is obtained by adding/removing one tuple in $\boldsymbol{D}$
- Observation 1
  - Change of a tuple can be regarded as removing a tuple from the dataset and then inserting a new one
  - Indication: Unbounded $\boldsymbol{\varepsilon}$-differential privacy implies bounded $(2\boldsymbol{\varepsilon})$-differential privacy
  - Proof: $\Pr[\boldsymbol{A}(\boldsymbol{D}_1) = \boldsymbol{O}] \leq \exp(\boldsymbol{\varepsilon}) \cdot \Pr[\boldsymbol{A}(\boldsymbol{D}_2) = \boldsymbol{O}]$
    $\leq \exp(\boldsymbol{\varepsilon}) \cdot \exp(\boldsymbol{\varepsilon}) \cdot \Pr[\boldsymbol{A}(\boldsymbol{D}_3) = \boldsymbol{O}]$

Variants of Differential Privacy

- Bounded: $\boldsymbol{D'}$ is obtained by changing the values of one tuple in $\boldsymbol{D}$
- Unbounded: $\boldsymbol{D'}$ is obtained by adding/removing one tuple in $\boldsymbol{D}$
- Observation 2
  - Bounded differential privacy allows us to directly publish the number of tuples in the dataset

$$\Pr[\boldsymbol{A}(\boldsymbol{D}) = \boldsymbol{O}] \leq \exp(\boldsymbol{\varepsilon}) \cdot \Pr[\boldsymbol{A}(\boldsymbol{D'}) = \boldsymbol{O}]$$

  - Unbounded differential privacy does not allow this

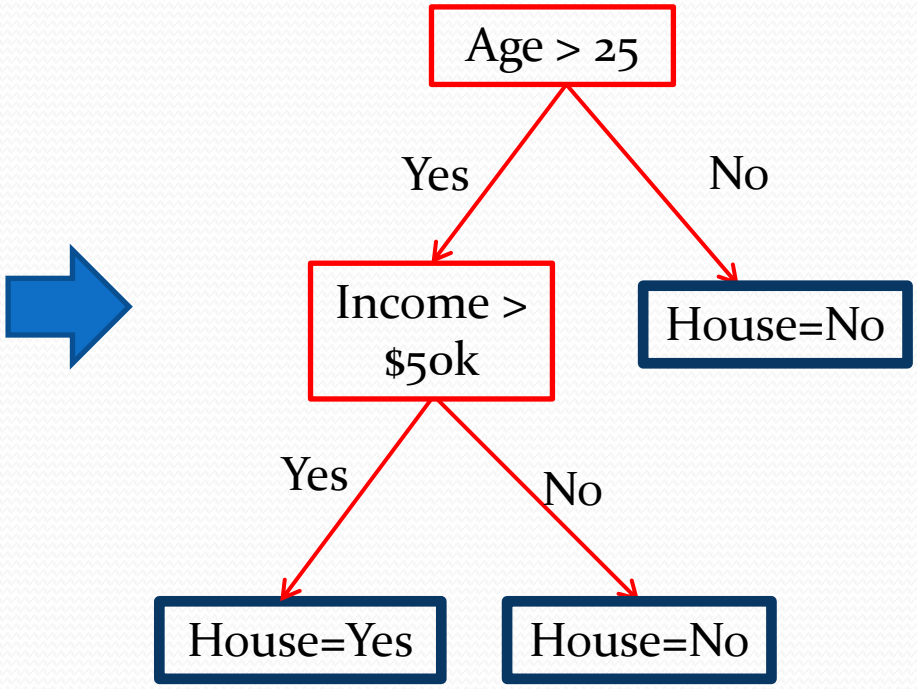
Limitations of Differential Privacy

- Differential privacy tends to be less effective when there exist correlations among the tuples
- Example (from [Kifer and Machanavajjhala 2011]):
  - Bob's family includes 10 people, and all of them are in a database
  - There is a highly contagious disease, such that if one family member contracts the disease, then the whole family will be infected
  - Differential privacy would underestimate the risk of disclosure
- Summary: Amount of noise needed depends on the correlations among the tuples, which is not captured by differential privacy

Decision Tree Classification

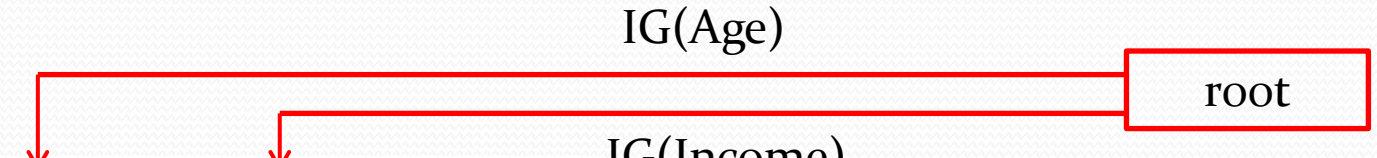
- Problem Definition

User	Age	Income	House
Alice	25	$50k	No
Bob	51	$40k	No
Chris	44	$100k	Yes
Doug	28	$60k	Yes
…	…	…	…



Decision Tree Classification

- Attribute Selection *[Friedman, 2010]*



User	Age	Income	House
Alice	25	$50k	No
Bob	51	$40k	No
Chris	44	$100k	Yes
Doug	28	$60k	Yes
…	…	…	…

Pick up a splitting attributes by maximizing the information gain

$$IG(T,a) = H(T) - \sum_{v \in vals(a)} \frac{|\{\mathbf{x} \in T | x_a = v\}|}{|T|} \cdot H(\{\mathbf{x} \in T | x_a = v\})$$

Decision Tree Classification

- How to enforce differential privacy in the selection?
  - Laplace Mechanism
  - Exponential Mechanism

Attribute	Info. Gain
Age	3.5
Income	2.2
…	…

Laplace

Attribute	Info. Gain
Age	2.9
Income	2.7
…	…

Budget consumption:
$\varepsilon \times m$

Decision Tree Classification

- How to enforce differential privacy in the selection?
  - Laplace Mechanism
  - Exponential Mechanism

Attribute	Info. Gain
Age	3.5
Income	2.2
…	…

Exponential →

Attribute	Probability
Age	0.7
Income	0.2
…	…

Budget consumption: ε

Conclusion

- Differential Privacy is a new and robust criterion of privacy detection
- There are simple algorithms enforcing differential privacy
- For a specific query engine, we need to carefully pick up the appropriate place to insert noise.

Source of Privacy Problem

- Even if we do not publish the identities of individuals, there are some (non-sensitive) fields that may *uniquely* identify some individuals
  - These attributes form the *quasi identifier*

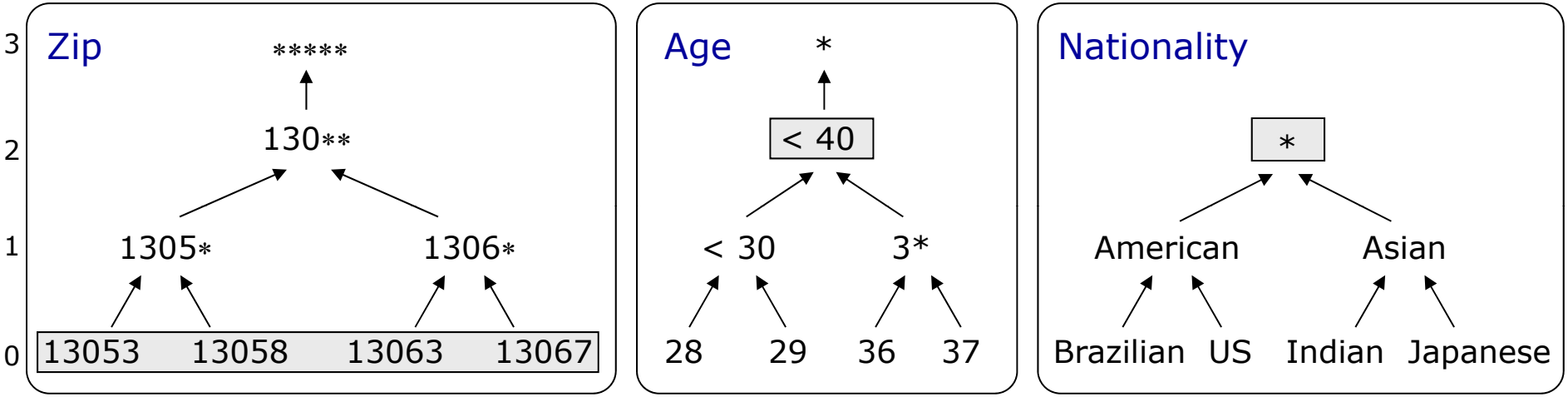
	Non-Sensitive Data (Quasi identifier)			*Other non-sensitive attributes (non-QI)*	*Sensitive Data*
#	***Zip***	***Age***	***Nationality***	***…***	***Condition***
…	…	…	…		…

Quasi Identifier

- The attacker can use them to *join* with other sources and identify the individuals

How to k-anonymize a dataset? Generalization Hierarchies

- Generalization Hierarchies: Data owner defines how values can be generalized



- Table Generalization: A table generalization is created by generalizing all values in a column to a specific level of generalization

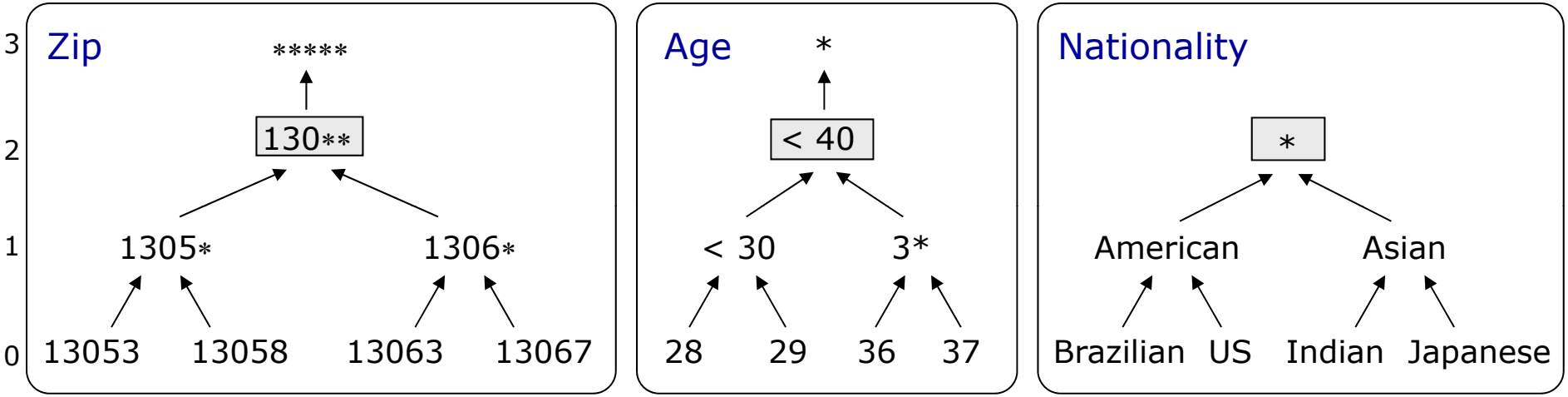
e.g.

2-anonymization

#	Zip	Age	Nationality	Condition
1	13053	< 40	*	Heart Disease
2	13067	< 40	*	Heart Disease
3	13053	< 40	*	Cancer
4	13067	< 40	*	Cancer

Generalization Hierarchies

- Generalization Hierarchies: Data owner defines how values can be generalized



- Table Generalization: A table generalization is created by generalizing all values in a column to a specific level of generalization

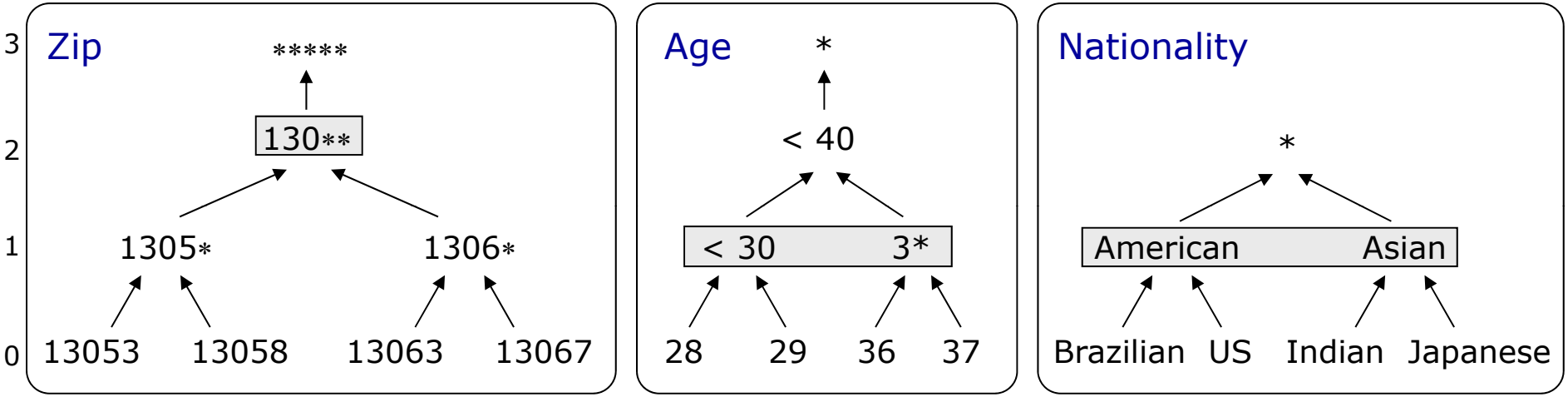
e.g.

2-anonymization

#	Zip	Age	Nationality	Condition
1	130**	< 40	*	Heart Disease
2	130**	< 40	*	Heart Disease
3	130**	< 40	*	Cancer
4	130**	< 40	*	Cancer

Generalization Hierarchies

- Generalization Hierarchies: Data owner defines how values can be generalized



- Table Generalization: A table generalization is created by generalizing all values in a column to a specific level of generalization

e.g.

2-anonymization

#	Zip	Age	Nationality	Condition
1	130**	< 30	American	Heart Disease
2	130**	< 30	American	Heart Disease
3	130**	3*	Asian	Cancer
4	130**	3*	Asian	Cancer

k-minimal Generalizations

- There are *many* k-anonymizations. Which to pick?
  The ones that do not generalize the data more than needed

k-minimal Generalization: A k-anonymization that is not a generalization of another k-anonymization

e.g.

✓ *2-minimal Generalization*

#	Zip	Age	Nationality	
1	13053	< 40	*	He
2	13067	< 40	*	He
3	13053	< 40	*	Ca
4	13067	< 40	*	Ca

✓ *2-minimal Generalization*

#	Zip	Age	Nationality	
1	130**	< 30	American	He
2	130**	< 30	American	He
3	130**	3*	Asian	Ca
4	130**	3*	Asian	Ca

#	Zip	Age	Nationality	
1	130**	< 40	*	H
2	130**	< 40	*	H
3	130**	< 40	*	C
4	130**	< 40	*	C

✘ *Non-minimal 2-anonymization*

k-Anonymity Attack Example

Original Data

	Quasi-Identifier			*Sensitive Data*
#	**ZIP**	**Age**	**Nationality**	**Condition**
1	13053	28	Russian	Heart Disease
2	13068	29	American	Heart Disease
3	13068	21	Japanese	Viral Infection
4	13053	23	American	Viral Infection
5	14853	50	Indian	Cancer
6	14853	55	Russian	Heart Disease
7	14850	47	American	Viral Infection
8	14850	49	American	Viral Infection
9	13053	31	American	Cancer
10	13053	37	Indian	Cancer
11	13068	36	Japanese	Cancer
12	13068	35	American	Cancer

The attacker knows:

- About quasi-identifiers:

Umeko		
Zip	**Age**	**National**
13068	21	Japanese

Bob		
Zip	**Age**	**National**
13053	31	American

- Other background knowledge:

  Japanese have low incidence of heart disease

k-Anonymity Attack Example

4-anonymization

	Quasi-Identifiers			*Sensitive Data*
#	***ZIP***	***Age***	***Nationality***	***Condition***
1	130**	< 30	*	Heart Disease
2	130**	< 30	*	Heart Disease
3	130**	< 30	*	Viral Infection
4	130**	< 30	*	Viral Infection
5	1485*	> = 40	*	Cancer
6	1485*	> = 40	*	Heart Disease
7	1485*	> = 40	*	Viral Infection
8	1485*	> = 40	*	Viral Infection
9	130**	3*	*	Cancer
10	130**	3*	*	Cancer
11	130**	3*	*	Cancer
12	130**	3*	*	Cancer

k-Anonymity Attack Example

4-anonymization

	Quasi-Identifiers			*Sensitive Data*
#	***ZIP***	***Age***	***Nationality***	***Condition***
1	130**	< 30	*	Heart Disease
2	130**	< 30	*	Heart Disease
3	130**	< 30	*	Viral Infection
4	130**	< 30	*	Viral Infection
5	1485*	> = 40	*	Cancer
6	1485*	> = 40	*	Heart Disease
7	1485*	> = 40	*	Viral Infection
8	1485*	> = 40	*	Viral Infection
9	130**	3*	*	**Cancer**
10	130**	3*	*	**Cancer**
11	130**	3*	*	**Cancer**
12	130**	3*	*	**Cancer**

Bob		
Zip	***Age***	***National***
13053	31	American

Bob has Cancer!

k-Anonymity Attack Example

4-anonymization

	Quasi-Identifiers			*Sensitive Data*
#	***ZIP***	***Age***	***Nationality***	***Condition***
1	130**	< 30	*	~~Heart Disease~~
2	130**	< 30	*	~~Heart Disease~~
3	130**	< 30	*	**Viral Infection**
4	130**	< 30	*	**Viral Infection**
5	1485*	> = 40	*	Cancer
6	1485*	> = 40	*	Heart Disease
7	1485*	> = 40	*	Viral Infection
8	1485*	> = 40	*	Viral Infection
9	130**	3*	*	**Cancer**
10	130**	3*	*	**Cancer**
11	130**	3*	*	**Cancer**
12	130**	3*	*	**Cancer**

Umeko

Zip	***Age***	***National***
13068	21	Japanese

Umeko has Viral Infection!

Data Leak !

Bob

Zip	***Age***	***National***
13053	31	American

Bob has Cancer!

Return a k-anonymization with the additional property that: For each distinct value of the quasi-identifier there exists ***l*** *different values for the sensitive attributes (i.e.,* ***l****-diversed)*

3-diversified

	Quasi-Identifiers			*Sensitive Data*
#	***ZIP***	***Age***	***Nationality***	***Condition***
1	1305*	<= 40	*	Heart Disease
2	1306*	<= 40	*	~~Heart Disease~~
3	1306*	<= 40	*	Viral Infection
4	1305*	<= 40	*	Viral Infection
5	1485*	>= 40	*	Cancer
6	1485*	>= 40	*	Heart Disease
7	1485*	>= 40	*	Viral Infection
8	1485*	>= 40	*	Viral Infection
9	1305*	<= 40	*	Cancer
10	1305*	<= 40	*	Cancer
11	1306*	<= 40	*	Cancer
12	1306*	<= 40	*	Cancer

Attack does not work!

Umeko

Zip	***Age***	***National***
13068	21	Japanese

Umeko has Viral Infection or Cancer

Bob

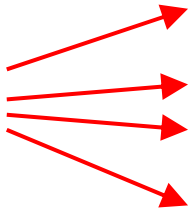
Zip	***Age***	***National***
13053	31	American

Bob has Viral Infection or Cancer or Heart Disease

What *l*-diversity guarantees

- From an *l*-diverse generalized table, an adversary (without any prior knowledge) can infer the sensitive value of each individual with confidence at most 1/***l***

Name	Age	Sex	Zipcode
Bob	23	M	11000



A 2-diverse generalized table

Age	Sex	Zipcode	Disease
[21, 60]	M	[10001, 60000]	pneumonia
[21, 60]	M	[10001, 60000]	dyspepsia
[21, 60]	M	[10001, 60000]	dyspepsia
[21, 60]	M	[10001, 60000]	pneumonia
[61, 70]	F	[10001, 60000]	flu
[61, 70]	F	[10001, 60000]	gastritis
[61, 70]	F	[10001, 60000]	flu
[61, 70]	F	[10001, 60000]	bronchitis

Limitations of *l*-Diversity

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

New notion of privacy needed to factor in the data distribution – t-closeness!