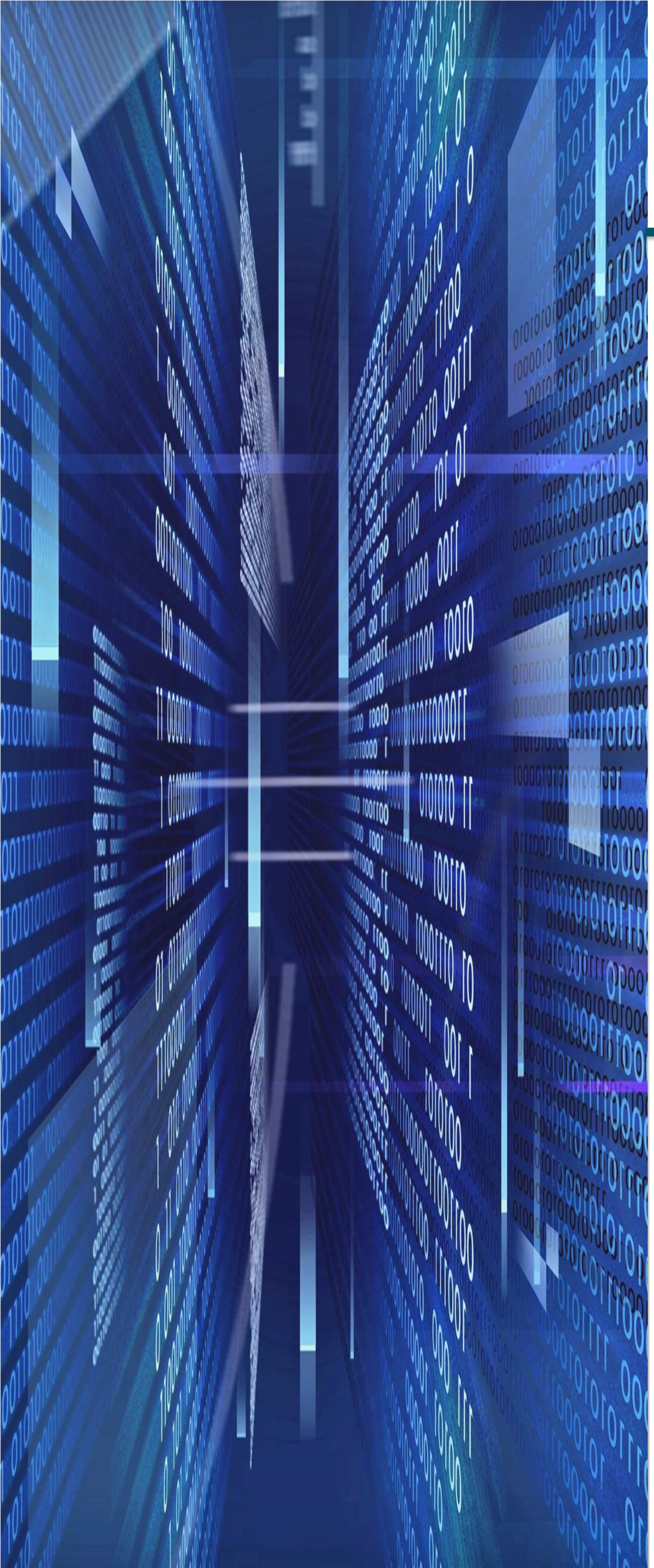


Data Anonymisation

Prof. Federica Paci

Today’ s Lecture

- Data Anonymisation techniques
  - k-anonimity
  - l-diversity
  - t-closeness
  - differential privacy

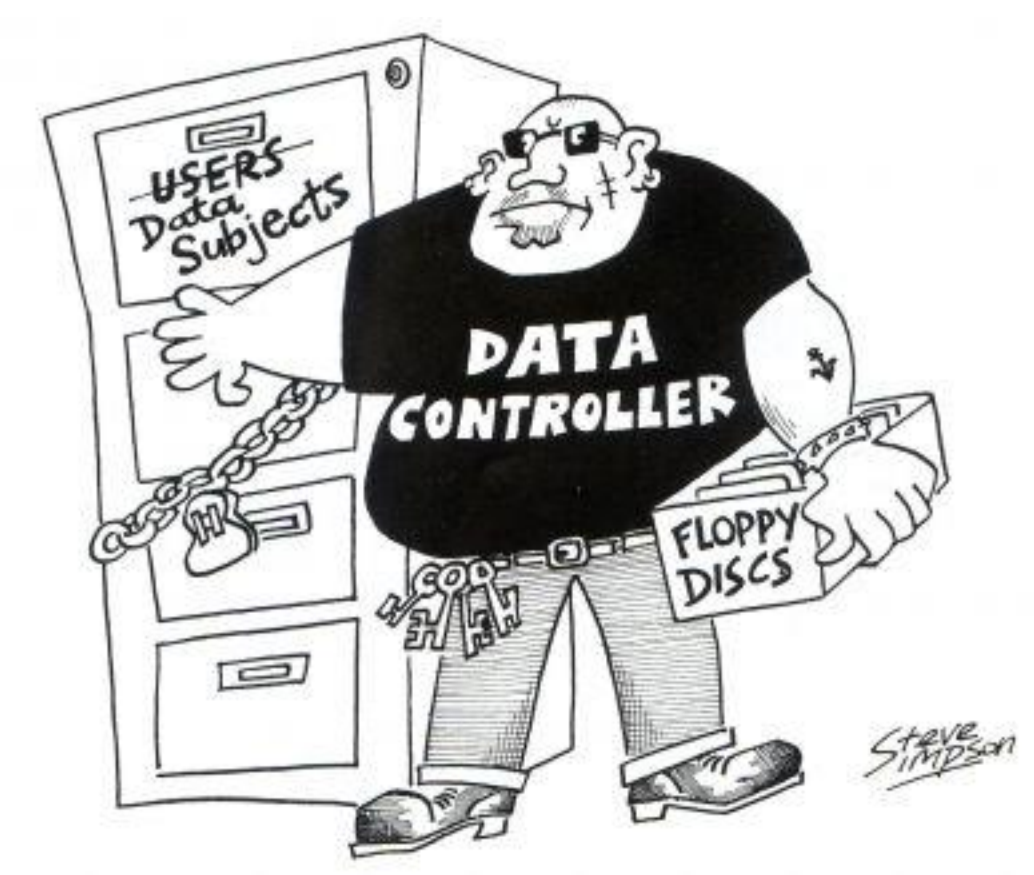
Learning Outcomes

- At the end of this lecture you should be able to:
  - Provide a definition of k-anonymity
  - Provide a definition of l-diversity
  - Provide a definition of t-closeness
  - Provide a definition of differential privacy
  - Discuss the limitations of these approaches to privacy

The Privacy Problem

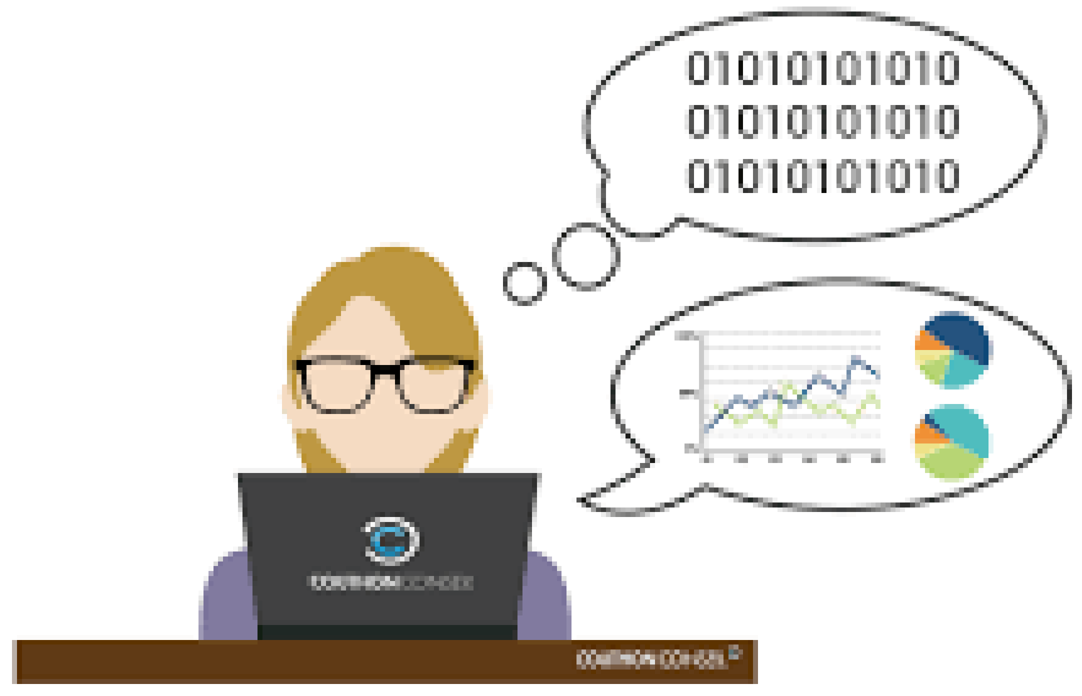
Given a dataset with sensitive personal information: how to compute and release functions of the dataset while protecting individual privacy?

An Example



ID	QID			SA
Name	Zipcode	Age	Sex	Disease
Izzy	47677	29	F	Ovarian Cancer
Rose	47602	22	F	Ovarian Cancer
Bob	47678	27	M	Prostate Cancer
John	47905	43	M	Flu
Alice	47909	52	F	Heart Disease
Fred	47906	47	M	Heart Disease

Medical Data



Data Analyst

Attacker

Bob had Prostate Cancer

Data Subject

Classification of Attributes

- Explicit identifiers
  - Identify a user
  - E.g name, lastname, passport number, etc.
- Quasi-identifiers
  - E.g Date of birth, Age, Zip code, phone number
- Sensitive attributes
  - E.g diseases, salaries, etc.

  These attributes is what the researchers need, so they are always released directly

An Example

Key Attributes | **Quasi-identifiers** | **Sensitive attributes**

ID	Name	DOB	Gender	Zipcode	Disease
12345	Andre	1/21/76	Male	53715	Heart Disease
56789	Beth	4/13/86	Female	53715	Hepatitis
52131	Carol	2/28/76	Male	53703	Brochitis
85438	Dan	1/21/76	Male	53703	Broken Arm
91281	Ellen	4/13/86	Female	53706	Flu
11253	Eric	2/28/76	Female	53706	Hang Nail

Protecting Explicit Identifiers

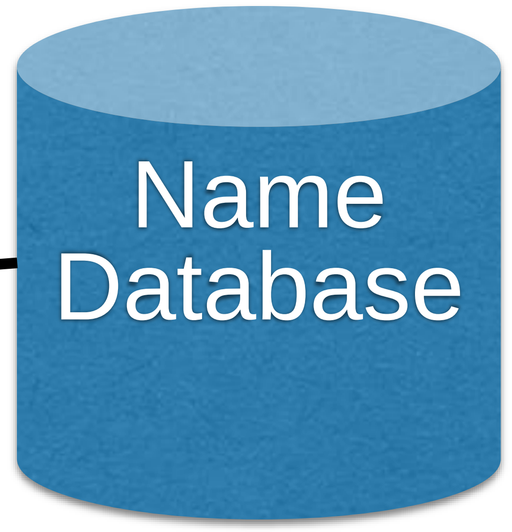
- **Tokenization**: generates a unique token for the input data
- **Substitution**: replaces an attribute value with alternative data values

Original Dataset

ID	Name
12345	Andre
56789	Beth
52131	Carol
85438	Dan
91281	Ellen
11253	Eric

Released Dataset

ID	Name
40011	Jack
81100	Sammy
62410	Mark
79820	Jane
14532	Singh
22244	Khan



Tokenization

Is it enough to protect explicit identifiers?

A Face Is Exposed for AOL Searcher No. 4417749

By MICHAEL BARBARO and TOM ZELLER Jr. AUG. 9, 2006

Buried in a list of 20 million Web search queries collected by AOL and recently released on the Internet is user No. 4417749. The number w assigned by the company to protect the searcher's anonymity, but it much of a shield.

No. 4417749 conducted hundreds of searches over a three-month pe topics ranging from "numb fingers" to "60 single men" to "dog that u on everything."

And search by search, click by click, the identity of AOL user No. 441 became easier to discern. There are queries for "landscapers in Lilbu Ga," several people with the last name Arnold and "homes sold in sh lake subdivision gwinnett county georgia."

It did not take much investigating to follow that data trail to Thelma Arnold, a 62-year-old widow who lives in Lilburn, Ga., frequently researches her friends' medical ailments and loves her three dogs. "Those are my searches," she said, after a reporter read part of the list to her.

Massachusetts hospital discharge dataset

Medical Data Released as Anonymous

SSN	Name	Ethnicity	Date Of Birth	Sex	ZIP	Marital Status	Problem
		asian	09/27/64	female	02139	divorced	hypertension
		asian	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
		[illegible]	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: Re-identifying anonymous data by linking to external data

Public voter dataset

K-Anonymity

- A record has to be indistinguishable from at least k-1 other records with the respect to the quasi-identifiers
- Each class of equivalence has to contain at least k records which have the same values for the quasi identifiers

Original Database

Name	Zipcode	Age	Disease
Hilary	47677	29	Heart Disease
Jenny	47602	22	Heart Disease
Bob	47678	27	Heart Disease
Izzy	47905	43	Flu
John	47909	52	Heart Disease
Fred	47906	47	Cancer
Sam	47605	30	Heart Disease
Carl	47673	36	Cancer
Sarah	47607	32	Cancer

Released Database

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

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

Example

Name	Zipcode	Age	Sex	Disease
Hilary	47677	29	F	Heart Disease
Jenny	47673	22	F	Heart Disease
Bob	47678	27	M	Heart Disease
Izzy	47905	43	F	Flu
John	47909	52	M	Heart Disease
Fred	47906	47	M	Cancer
Sam	47605	30	M	Heart Disease
Carl	47602	36	M	Cancer
Sarah	47607	32	F	Cancer

Example: Generalization

Zipcode	Age	Sex	Disease
47677	21-30	F	Heart Disease
47673	21-30	F	Heart Disease
47678	21-30	M	Heart Disease
47909	51-60	M	Heart Disease
47906	41-50	M	Cancer
47605	21-30	M	Heart Disease
47602	31-40	M	Cancer
47607	31-40	F	Cancer

Example: Generalization

Zipcode	Age	Sex	Disease
47677	10-29	F	Heart Disease
47673	10-29	F	Heart Disease
47678	10-29	M	Heart Disease
47909	50-69	M	Heart Disease
47906	50-69	M	Cancer
47605	30-49	M	Heart Disease
47602	30-49	M	Cancer
47607	30-49	F	Cancer

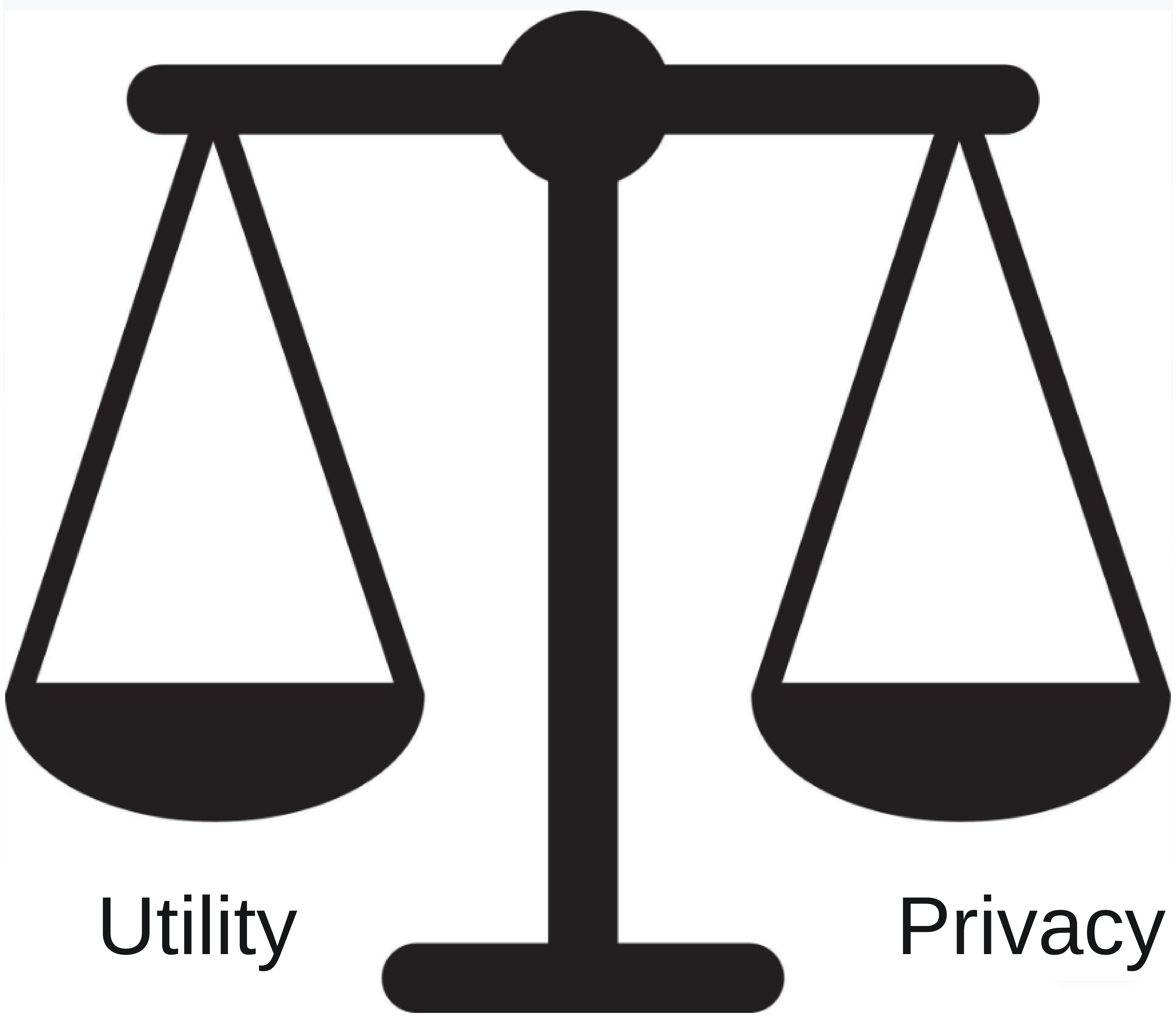
Example: Generalization + Suppression

Zipcode	Age	Sex	Disease
*	10-29	*	Heart Disease
*	10-29	*	Heart Disease
*	10-29	*	Heart Disease
*	50-69	M	Heart Disease
*	50-69	M	Cancer
*	30-49	*	Heart Disease
*	30-49	*	Cancer
*	30-49	*	Cancer

Example: Generalization + Suppression

Zipcode	Age	Sex	Disease
47670	10-29	*	Heart Disease
47670	10-29	*	Heart Disease
47670	10-29	*	Heart Disease
47900	50-69	M	Heart Disease
47900	50-69	M	Cancer
47600	30-49	*	Heart Disease
47600	30-49	*	Cancer
47600	30-49	*	Cancer

Privacy vs Utility



Exercise

Name	Age	Gender	State of domicile	Religion	Disease
Ramsha	29	Female	Tamil Nadu	Hindu	Cancer
Yadu	24	Female	Kerala	Hindu	Viral infection
Salima	28	Female	Tamil Nadu	Muslim	TB
Sunny	27	Male	Karnataka	Parsi	No illness
Joan	24	Female	Kerala	Christian	Heart-related
Bahuksana	23	Male	Karnataka	Buddhist	TB
Rambha	19	Male	Kerala	Hindu	Cancer
Kishor	29	Male	Karnataka	Hindu	Heart-related
Johnson	17	Male	Kerala	Christian	Heart-related
John	19	Male	Kerala	Christian	Viral infection

Exercise

Name	Age	Gender	State of domicile	Religion	Disease
*	$20 < Age \leq 30$	Female	Tamil Nadu	*	Cancer
*	$20 < Age \leq 30$	Female	Kerala	*	Viral infection
*	$20 < Age \leq 30$	Female	Tamil Nadu	*	TB
*	$20 < Age \leq 30$	Male	Karnataka	*	No illness
*	$20 < Age \leq 30$	Female	Kerala	*	Heart-related
*	$20 < Age \leq 30$	Male	Karnataka	*	TB
*	$Age \leq 20$	Male	Kerala	*	Cancer
*	$20 < Age \leq 30$	Male	Karnataka	*	Heart-related
*	$Age \leq 20$	Male	Kerala	*	Heart-related
*	$Age \leq 20$	Male	Kerala	*	Viral infection

Exercise

- What are the quasi identifiers?
- What is the value of k?

	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

Attacks against k- Anonymity

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

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

Homogeneity attack

Bob	
Zipcode	***Age***
47678	27

Background knowledge attack

Umeko	
Zipcode	***Age***
47673	36

L-diversity: intuition

[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

....	Disease
	HIV
	HIV
	HIV
	HIV
	HIV
	HIV
	HIV
	Bronchitis
	Pneumonia

8 records have HIV

2 records have other values

Entropy l-diversity

- Each equivalence class not only must have enough different sensitive values, but also the different sensitive values must be distributed evenly enough
- The entropy of the distribution of the sensitive values in each equivalence class has to be at least log(l)

Entropy ℓ-diversity. The entropy of an equivalence class E is defined to be

$$Entropy(E) = -\sum_{s \in S} p(E,s) \log p(E,s)$$

in which S is the domain of the sensitive attribute, and $p(E,s)$ is the fraction of records in E that have sensitive value s.

Sensitive Attribute Disclosure

Similarity attack

Bob	
Zip	***Age***
47678	27

A 3-diverse patient table

Zipcode	Age	Salary	Disease
476**	2*	3K	Gastric Ulcer
476**	2*	4K	Gastritis
476**	2*	5K	Stomach Cancer
4790*	≥40	6K	Gastritis
4790*	≥40	11K	Flu
4790*	≥40	8K	Bronchitis
476**	3*	7K	Bronchitis
476**	3*	9K	Pneumonia
476**	3*	10K	Stomach Cancer

Conclusion

1. Bob's salary is in [3k,5k], which is relatively low
2. Bob has some stomach-related disease

l-diversity does not consider semantics of sensitive values!

Other limitations of l-diversity

- Example: sensitive attribute is HIV+ (1%) or HIV- (99%)
- Consider an equivalence class that contains an equal number of HIV+ and HIV- records
  - 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!

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 database

t-closeness

Zipcode	Age	Salary	Disease
476**	2*	3K	Gastric Ulcer
476**	2*	4K	Gastritis
476**	2*	5K	Stomach Cancer
4790*	≥40	6K	Gastritis
4790*	≥40	11K	Flu
4790*	≥40	8K	Bronchitis
476**	3*	7K	Bronchitis
476**	3*	9K	Pneumonia
476**	3*	10K	Stomach Cancer

The Earth Mover Distance

- Intuitively it estimates the effort to transform a distribution into another distribution
- one distribution is seen as a mass of earth spread in the space
- the other as a collection of holes in the same space.
- EMD measures the least amount of work needed to fill the holes with earth

t-closeness: Computing the Earth Mover Distance

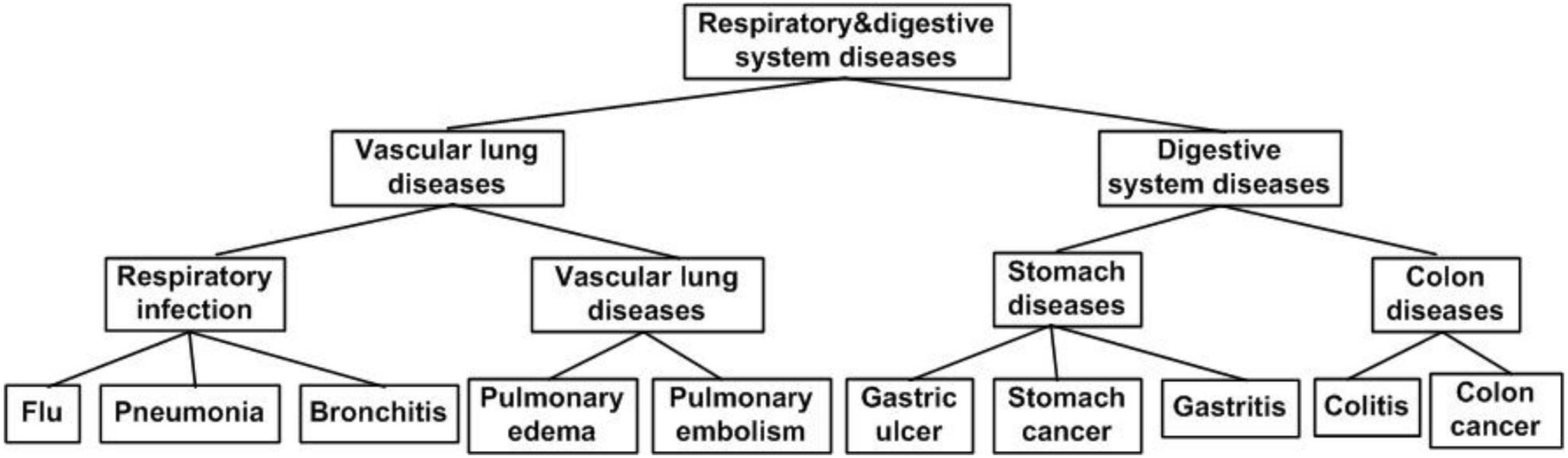
Disease
Gastric Ulcer
Gastritis
Stomach Cancer
Gastritis
Flu
Bronchitis
Bronchitis
Pneumonia
Stomach Cancer

{Gastric Ulcer, Gastritis, Stomach Cancer, Flu, Bronchitis, Pneumonia}

P1={Gastric Ulcer, Gastritis, Stomach Cancer} D[P1,Q]=0.5

P2={Gastric Ulcer, Stomach Cancer, Pneumonia} D[P2,Q]= 0.278

T-closeness : Computing the Earth Mover Distance



t-closeness

Zipcode	Age	Salary	Disease
476**	20-40	3K	Gastric Ulcer
476**	20-40	4K	Gastritis
476**	20-40	5K	Stomach Cancer
4790*	40-60	6K	Gastritis
4790*	40-60	11K	Flu
4790*	40-60	8K	Bronchitis
476**	20-40	7K	Bronchitis
476**	20-40	9K	Pneumonia
476**	20-40	10K	Stomach Cancer

t-closeness

Zipcode	Age	Salary	Disease
476**	20-40	3K	Gastric Ulcer
476**	20-40	9K	Pneumonia
476**	20-40	5K	Stomach Cancer
4790*	40-60	6K	Gastritis
4790*	40-60	11K	Flu
4790*	40-60	8K	Bronchitis
476**	20-40	7K	Bronchitis
476**	20-40	4K	Gastritis
476**	20-40	10K	Stomach Cancer

Simply anonymizing data is **unsafe!**

Lessons Learned

- Supposedly de-identified data often contain alternative ways of identification (a.k.a. quasi identifiers)
- Access to the appropriate auxiliary information can then result in re-identification
- This is not ’purely theoretical’ but has been demonstrated many times with real-world datasets

The Aggregation Dream

A common intuitive idea: counts, averages, statistical models, classifiers, … are `structurally' safe

Science and practice have shown this to be often wrong

Reconstruction Attacks

Name/Id	age	weight	sex	disease	...
Mario Rossi	65	82	M	yes	...
Daniele Bianchi	35	120	M	yes	...
Lucia Verdi	40	45	F	no	...
...	...	...	...	...	...

Queries we would like to permit

How many people have the disease ?
Average age and weight of men who have the disease ?

aggregate

Queries that are dangerous for the privacy

Does Daniele Bianchi have the disease?
What is the name of the last record inserted in the database?
What is the age / weight of the last record inserted in the database?

individual

Reconstruction Attack

Name/Id	age	weight	sex	disease	...
Mario Rossi	65	82	M	yes	...
Daniele Bianchi	35	120	M	yes	...
Lucia Verdi	40	45	F	no	...
...	...	...	...	...	...

insertion of a new record

Name/Id	age	weight	sex	disease	...
Mario Rossi	65	82	M	yes	...
Daniele Bianchi	35	120	M	yes	...
Lucia Verdi	40	45	F	no	...
Sergio Neri	20	140	M	yes	...

How many men have the disease ? 2

What is the average age / weight of men who have the disease ? 50 / 101

How many men have the disease ? 3

What is the average age / weight of men who have the disease ? 40 / 114

We can deduce the exact age / weight of the new record

Problem

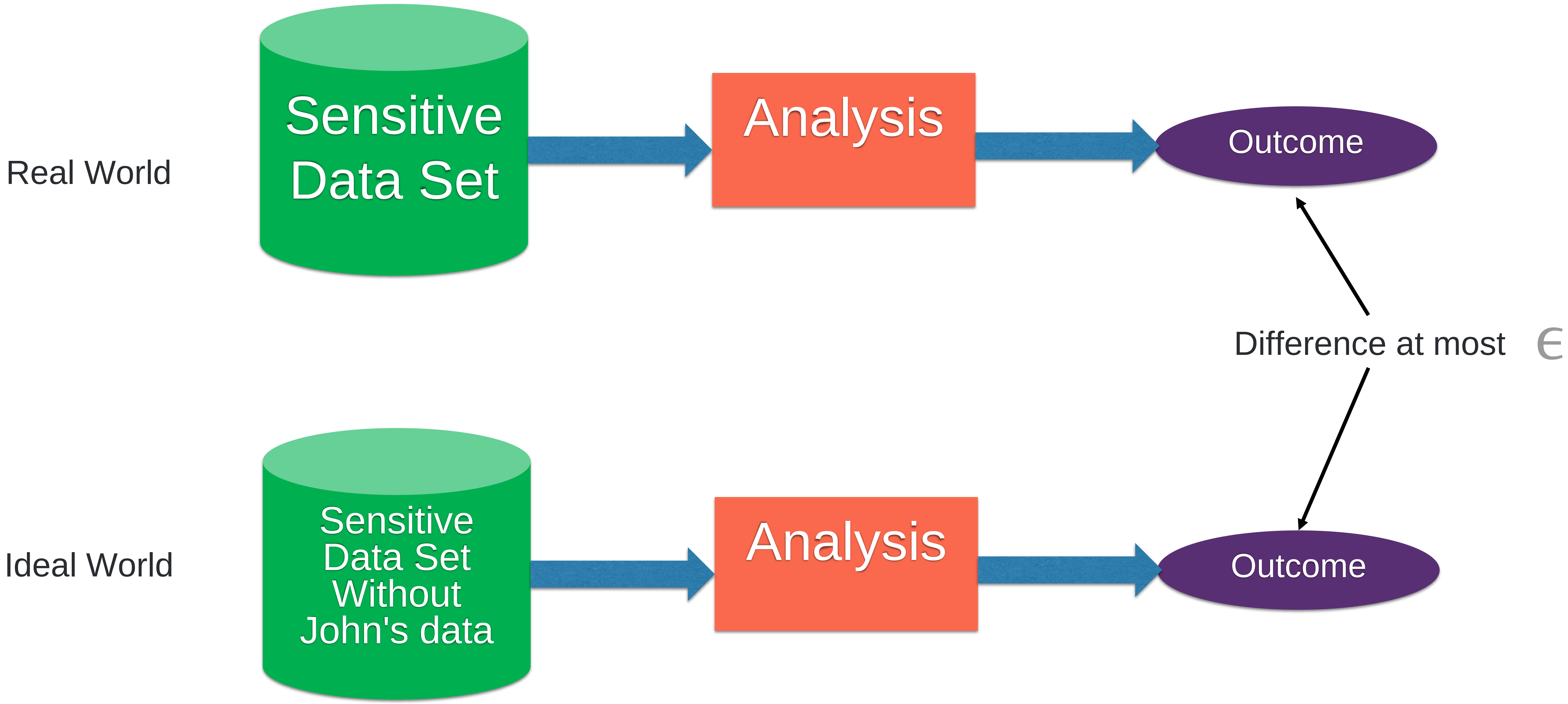
The restriction to aggregate queries is not sufficient: also these queries may leak information about individuals !

Differential privacy

It expresses a specific desiderata of an analysis:

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

Differential privacy: intuition



Example

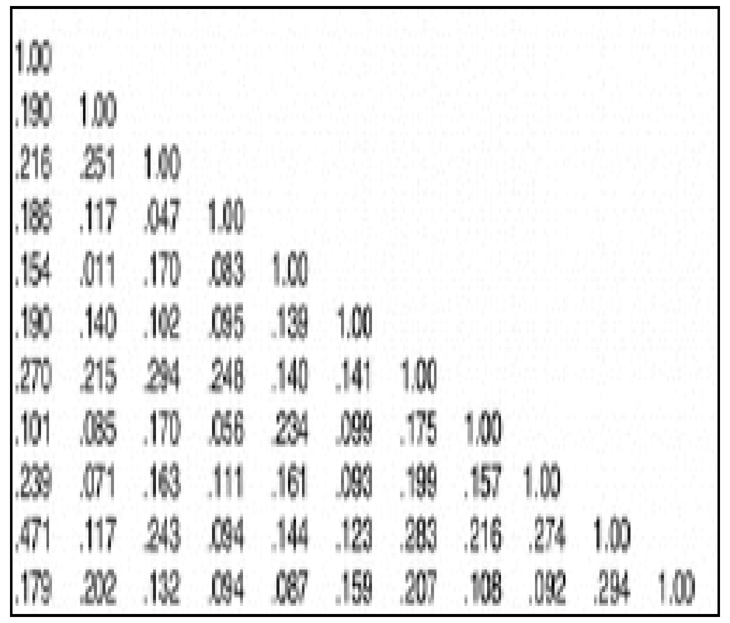
Adversary

Prior Knowledge:

A's Genetic profile

A smokes

Case 1: Study



1.00										
.190	1.00									
.216	.251	1.00								
.186	.117	.047	1.00							
.154	.011	.170	.083	1.00						
.190	.140	.102	.095	.139	1.00					
.270	.215	.294	.248	.140	.141	1.00				
.101	.085	.170	.056	.234	.099	.175	1.00			
.239	.071	.163	.111	.161	.093	.199	.157	1.00		
.471	.117	.243	.094	.144	.123	.283	.216	.274	1.00	
.179	.202	.132	.094	.087	.159	.207	.108	.092	.294	1.00

Cancer

A has cancer

[Study violates A's privacy]

Case 2: Study

Smoking causes cancer

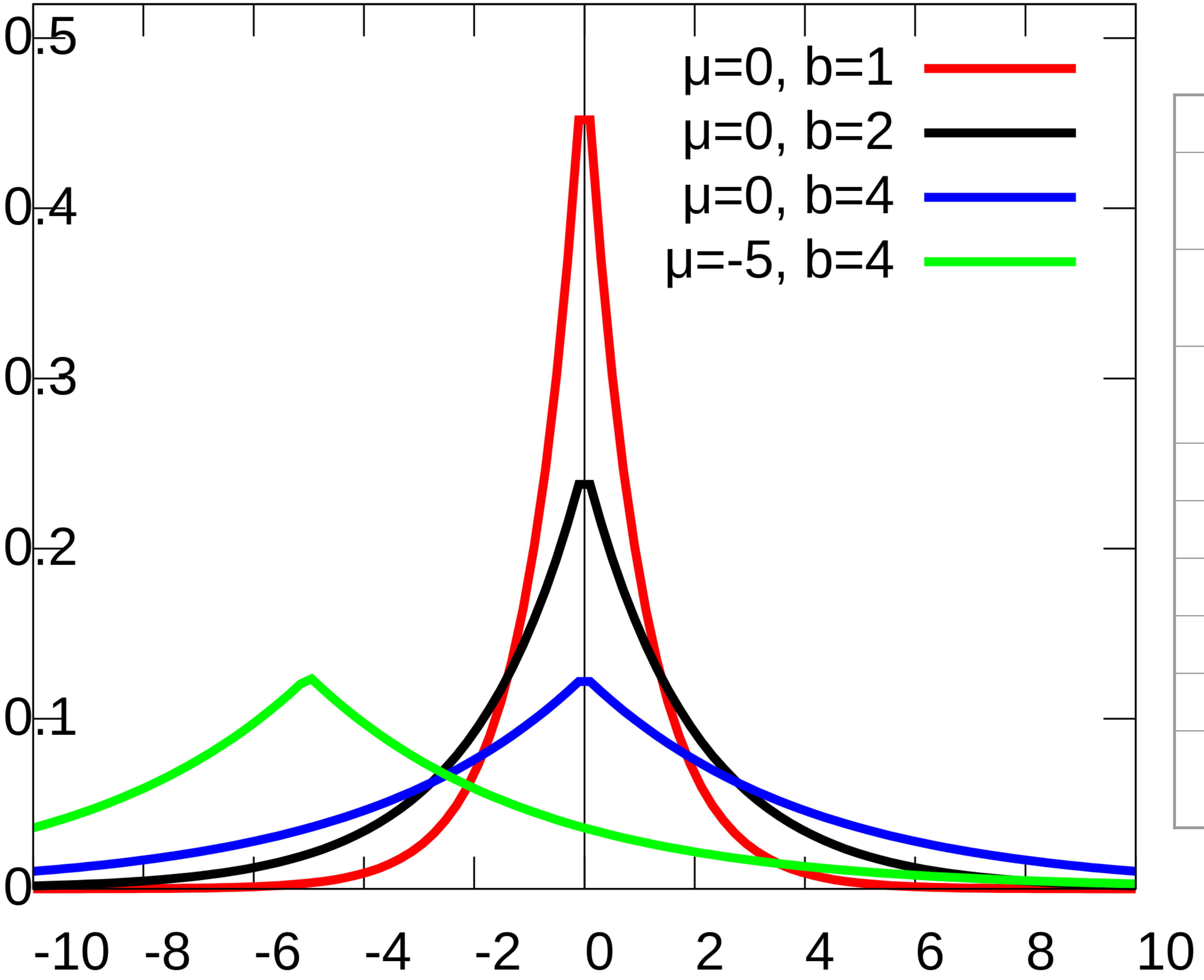
A probably has cancer

[Study does not violate privacy]

Differential Privacy: The Definition

$$\Pr[\mathcal{M}(\mathcal{D}) \in \mathcal{S}] \leq e^{\varepsilon} \cdot \Pr[\mathcal{M}(\mathcal{D}') \in \mathcal{S}]$$

Differential privacy: the intuition



Zipcode	Age	Salary	Disease	Noise
476**	20-40	3K	Gastric Ulcer	1
476**	20-40	9K	Pneumonia	2
476**	20-40	5K	Stomach Cancer	0
4790*	40-60	6K	Gastritis	- 3
4790*	40-60	11K	Flu	0
4790*	40-60	8K	Bronchitis	0
476**	20-40	7K	Bronchitis	-1
476**	20-40	4K	Gastritis	5
476**	20-40	10K	Stomach Cancer	0

Implementing differential privacy

Problem:

Given function f, sensitive dataset D

Find a differentially private approximation to f(D)

Example: f(D) = mean of data points in D

The Global Sensitivity Method

Given: A function f, sensitive dataset D

Define: dist(D, D') = #records that D, D' differ by

Global Sensitivity of f:

$$S(f) = \max_{dist(D, D') = 1} | f(D) - f(D')|$$

The Global Sensitivity Method

D

Name	Age
Alice	29
Bob	22
Charly	27
Dave	43
Eve	52
Ferris	47
George	30
Harvey	36
Iris	32

D’

Name	Age
Alice	29
Bob	22
Charly	27
Dave	43
Ferris	47
George	30
Harvey	36
Iris	32

The Global Sensitivity Method

If f is the count function

count(D) =9 count (D') =9 S(f) = 1

If f is the mean function

mean (D) = 35.3   mean (D') = 33.3   S(f) = 2

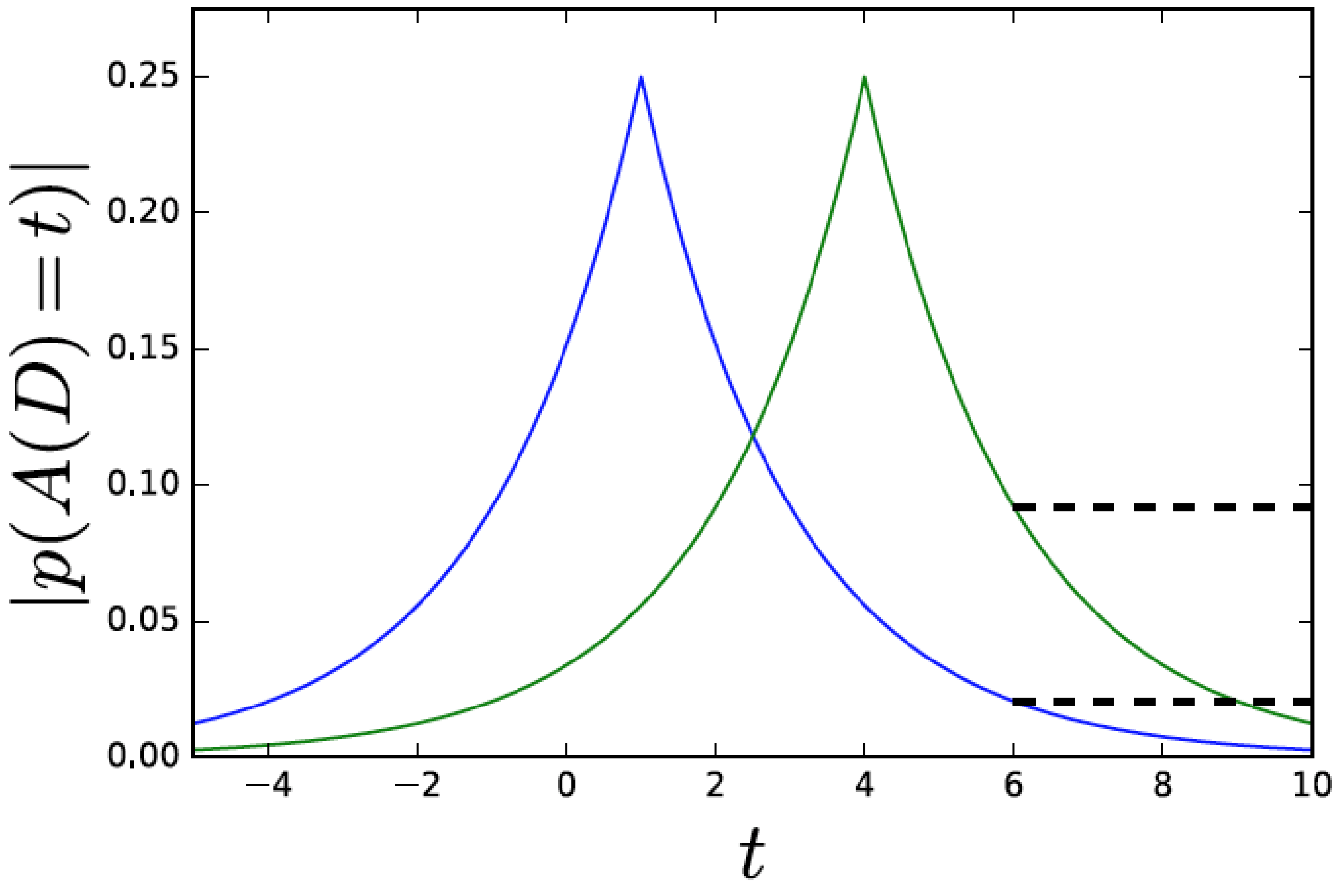
The Global Sensitivity Method

Global Sensitivity of f is $S(f) = \max_{\text{dist}(D, D') = 1} |f(D) - f(D')|$

Output $f(D) + Z$, where

$$Z \sim \frac{S(f)}{\epsilon}\text{Lap}(0, 1)$$

ϵ-differentially private



Laplace distribution:

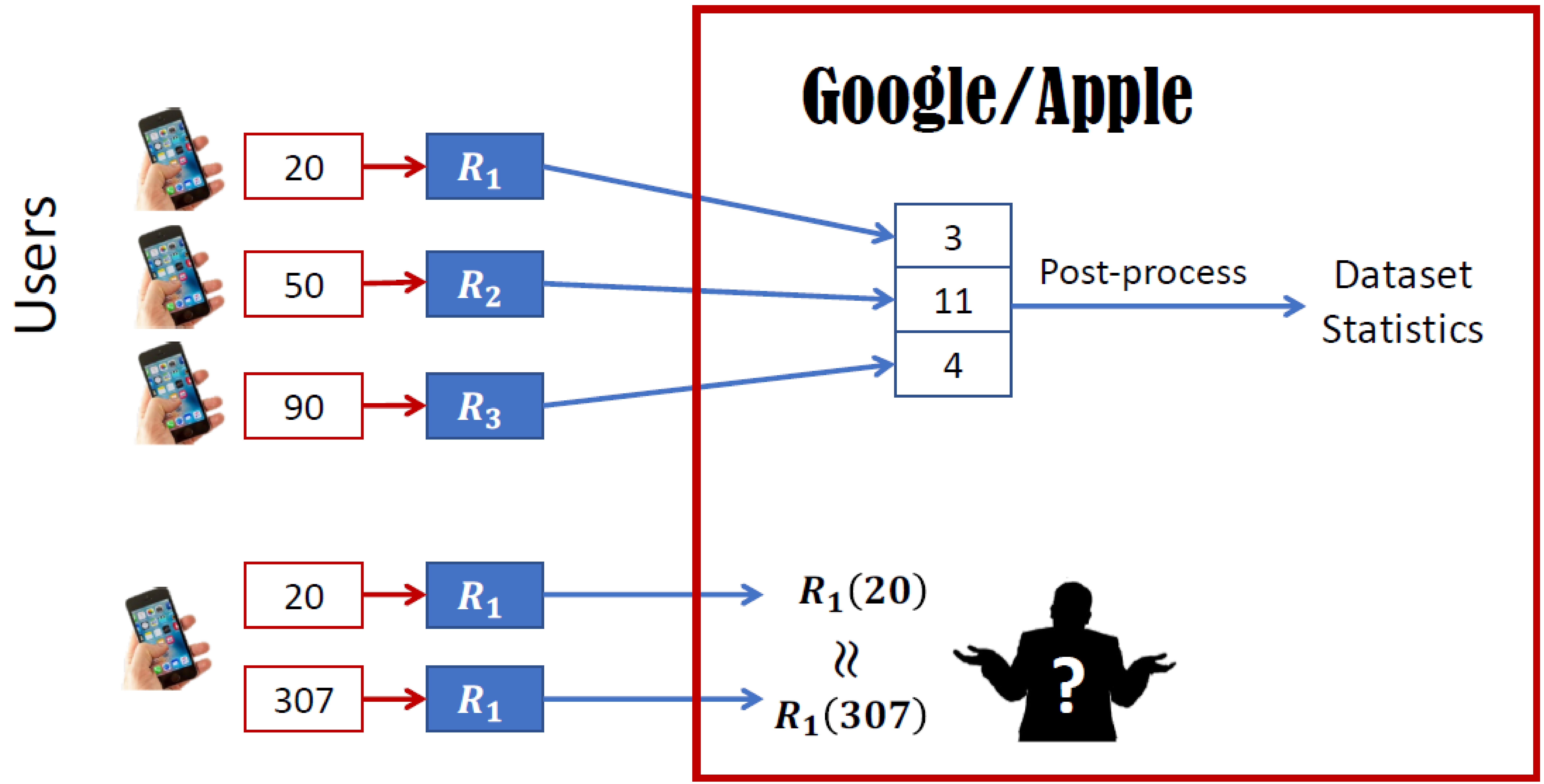
$$p(z|\mu, b) = \frac{1}{2b}\exp\left(-\frac{|z-\mu|}{b}\right)$$

What can be computed with differential privacy?

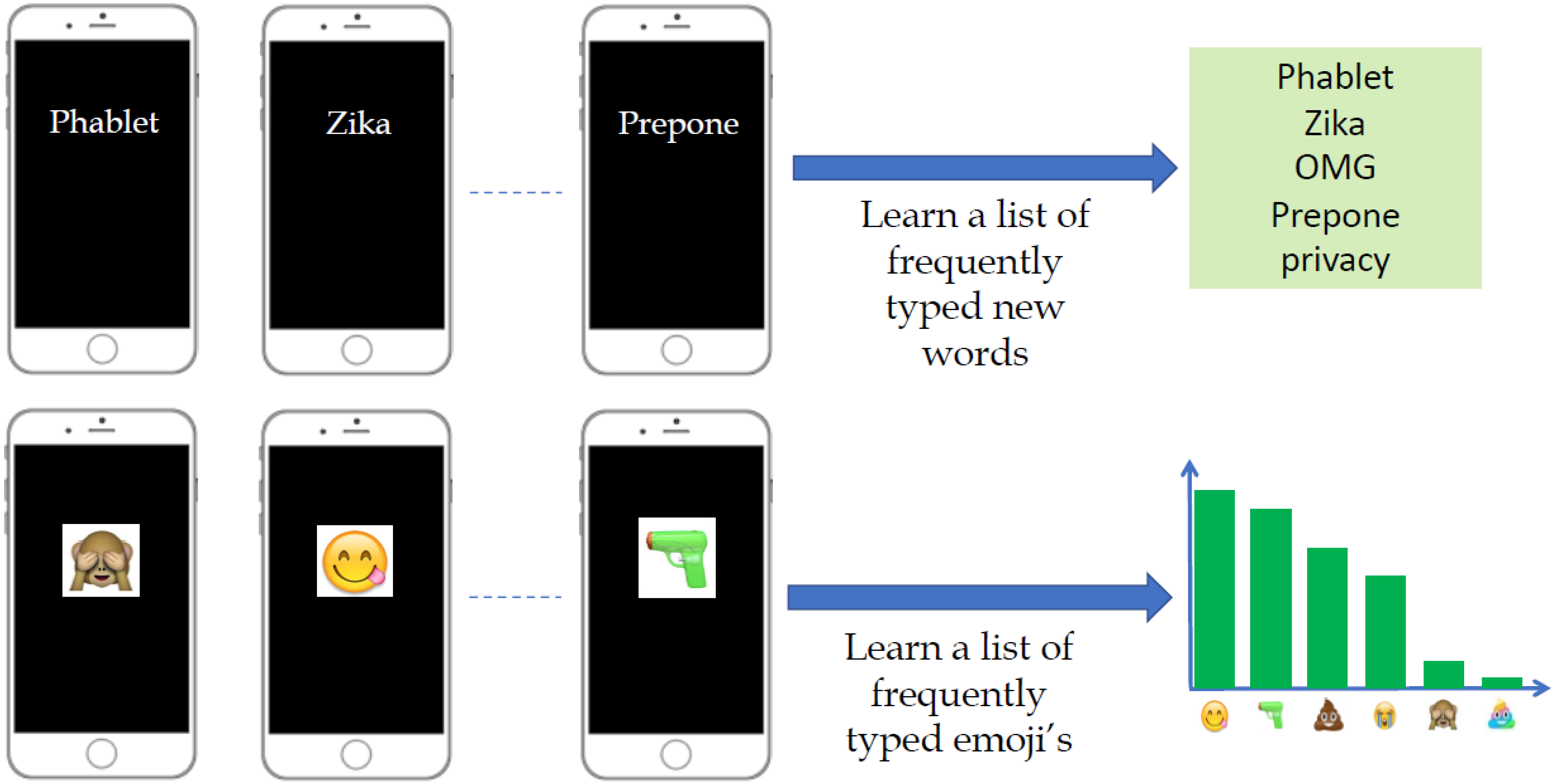
- Descriptive statistics: counts, mean, median, histograms, boxplots, etc.
- Supervised and unsupervised ML tasks: classification, regression, clustering, distribution learning, etc.
- Generation of synthetic data

US Census Bureau 2020

Local differential privacy



Learning Heavy Hitters



Introducing TensorFlow Privacy: Learning with Differential Privacy for Training Data

March 06, 2019

Posted by Carey Radebaugh (Product Manager) and Ulfar Erlingsson (Research Scientist)

Today, we're excited to announce TensorFlow Privacy (GitHub), an open source library that makes it easier not only for developers to train machine-learning models with privacy, but also for researchers to advance the state of the art in machine learning with strong privacy guarantees.

Facebook's Opacus

DEVELOPER TOOLS | OPEN SOURCE

Introducing Opacus: A high-speed library for training PyTorch models with differential privacy

August 31, 2020

We are releasing Opacus, a new high-speed library for training PyTorch models with differential privacy (DP) that's more scalable than existing state-of-the-art methods. Differential privacy is a mathematically rigorous framework for quantifying the anonymization of sensitive data. It's often used in analytics, with growing interest in the machine learning (ML) community. With the release of Opacus, we hope to provide an easier path for researchers and engineers to adopt differential privacy in ML, as well as to accelerate DP research in the field.

Summary

- k-anonymity only prevents identity disclosure
- L-diversity does not protect from attribute disclosure
- t-closeness protects against attribute disclosure
- Differential privacy guarantees that what can be learned about an individual is limited to what could be learned about him from everyone else's data without his own data being included in the computation

Recommended Readings

- Static Data Anonymization Part I: Multidimensional Data. Available at https://secure.ecs.soton.ac.uk/noteswiki/w/File:L05-Anonymization.pdf
- l-diversity: Privacy Beyond k-Anonymity: Available at: https://dl.acm.org/citation.cfm?id=1217302
- t-closeness: Privacy Beyond K-anonymity and l-diversity. Available at http://ieeexplore.ieee.org/document/4221659/
- Algorithmic foundations of differential privacy. Available at: https://www.cis.upenn.edu/~aaroth/Papers/privacybook.pdf
- Differential privacy: A primary for a Non-Technical Audience

  https://oconnell.fas.harvard.edu/files/salil/files/differential_privacy_primer_nontechnical_audience.pdf