Privacy-preserving Data Processing

Luca Vassio
Politecnico di Torino, Italy
luca.vassio@polito.it

April 10th, 2024

Who am I?

- **Luca Vassio**
- Assistant professor at Politecnico di Torino
- Area of Computer Engineering
- PhD in Electronic and Telecommunication Engineering
- MSc in Mathematical Engineering
- BSc in Mathematics

Outline

1. Private data generation and the need for anonymization
2. Privacy-preserving techniques
   - K-anonymity (and variants)
   - Differential Privacy
3. Tools for anonymization

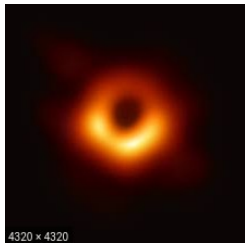
Outline

1. Private data generation and the need for anonymization
2. Privacy-preserving techniques
   - K-anonymity (and variants)
   - Differential Privacy
3. Tools for anonymization

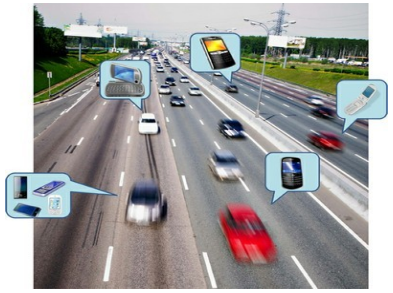
Private data generation and the need for anonymization

Who generates (big) data?

- User generated content on the Internet
- Health and scientific computing
- Log files
  - Web server log files, machine system log files
- Internet Of Things (IoT)
  - Sensor networks, RFID, smart meters

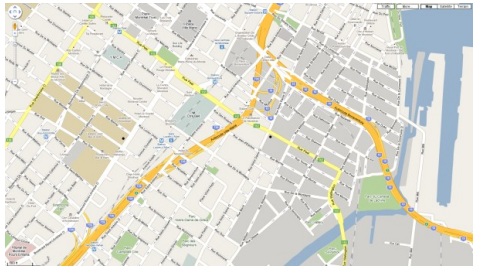


An example of data usage

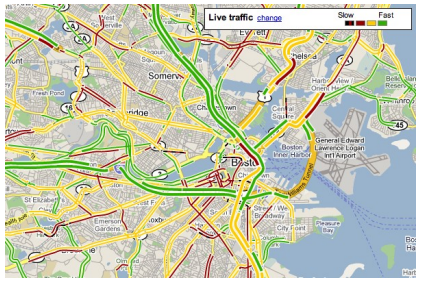


Crowdsourcing

Computing



Map data



Real time traffic info

Crowdsourcing data are generated by human beings

Who collects user data on the Internet

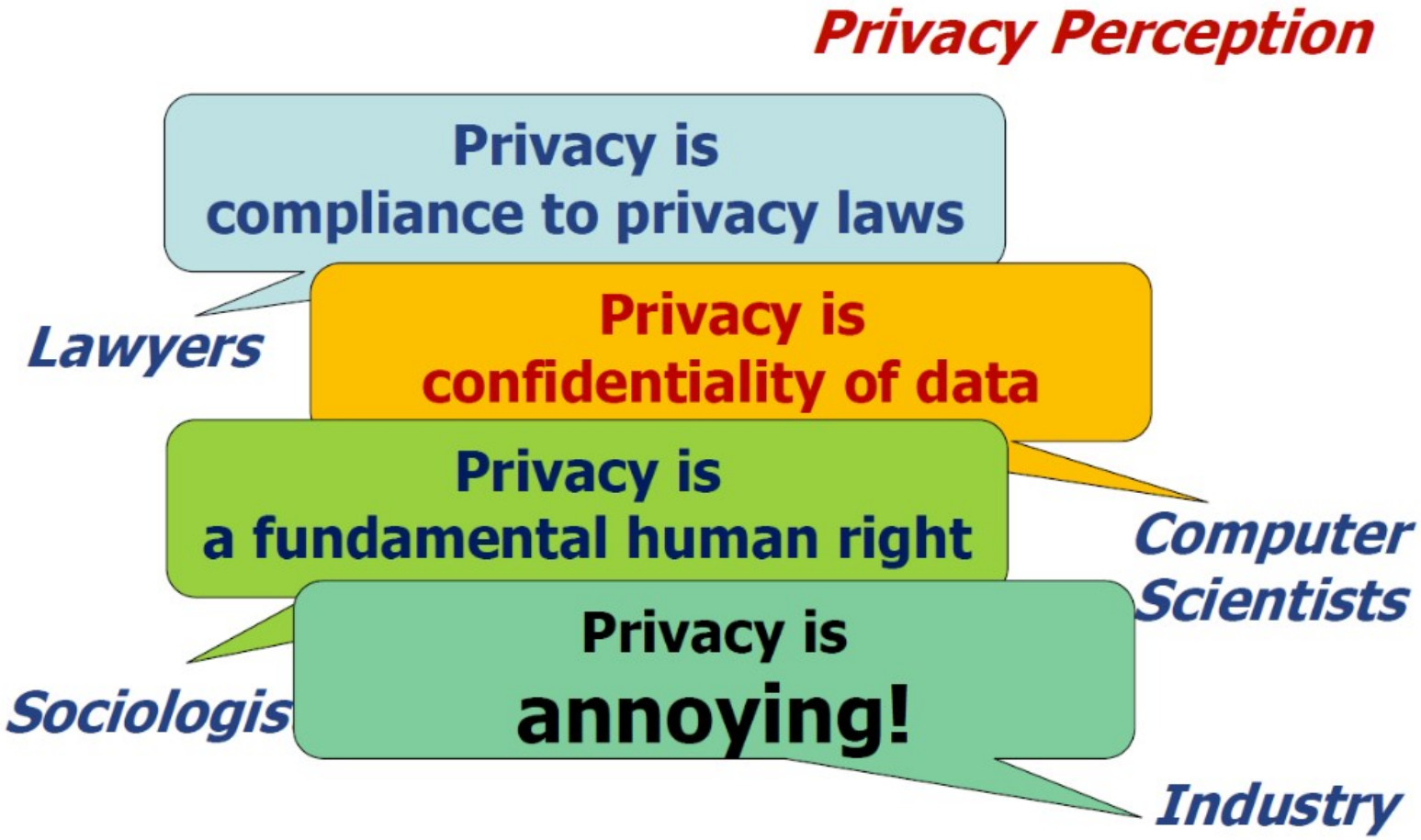
- Online shops
- Search engines
- Navigation systems
- ISPs
- (Almost) all websites for ads: web tracking

The (big) problem

A lot of these information are generated by human beings:

Data may threaten their privacy!

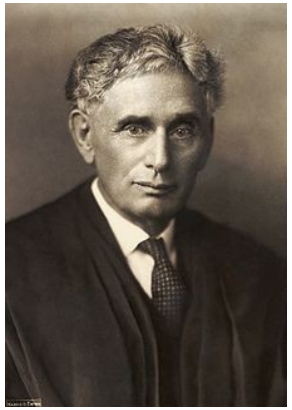
What do we mean by privacy?



Source: Meiko Jensen (German Research Center for Privacy and Security)
PhD Excellence Course: Data intensive approaches to urban and regional development. Politecnico di Torino. 14-20 October 2015

What do we mean by privacy?

- Samuel Warren and Louis Brandeis (1890) In *The Right to Privacy*
  - "Right to be left alone"



- Alan Westin (1967)
  - "Right to control, edit, manage, and delete information about themselves and decide when, how, and to what extent information is communicated to others"

Privacy and security of data

Deal with:

- how the **information is delivered or used** when **personal data** allowing to **identify the individual** are contained
- how the **information is anonymised** to ensure to not recognize the individual when data from multiple sources is combined

General Data Protection Regulation

EU Regulation 2016/6791 – Entered into Force in May 2018

- Goal: protect users' privacy, and punish violations
- It applies to data of EU citizen, worldwide

What are Personal Data?

Personal data are any information which are related to an identified or identifiable natural person

- Can be identifiers (e.g., name, phone number)
- Or special characteristics (e.g., genetic, mental, cultural info)

General Data Protection Regulation

Principles:

1. **Accountability**: be responsible and adopt a safe behavior and document it
2. **Privacy by design:** assess the privacy impact **before** data processing
3. **Privacy by default:** Cannot accumulate data without an objective
   - Can accumulate only the data strictly needed for the goal
   - Consent of users must be explicit
4. **Transparency**: simple, clear and complete information
   - Users should have access to data and be able to **download** it and delete it
   - Users and authorities must be informed of **data breaches** within **72 hours**

Data publishing and anonymization

- Data is published for different reasons:
  - For research, societal benefits, challenges, transparency,…
- Data need to be anonymized: no personal data of individuals
- **Anonymization** is harder than it appears

Anonymization

Hide identity → remove identifying info

Is this enough?

Popular cases of data de-anonymization

- In 1997 the Massachusetts Group Insurance Commission released **"anonymized" data** on state employees that showed every single hospital visit
    - **Identifiers** such as name, address, and Social Security number were **removed**
- Researcher **Latanya Sweeney:**
    - Purchased the voter rolls from Cambridge: a database with name, address, ZIP code, birth date, and sex of every voter
    - By combining data, she found **Governor Weld** hospital visits:
        - Only six people in Cambridge shared his birth date, only three of them were men, and of them, only he lived in his ZIP code.

born July 31, 1945
resident of 02138

Popular cases of data de-anonymization

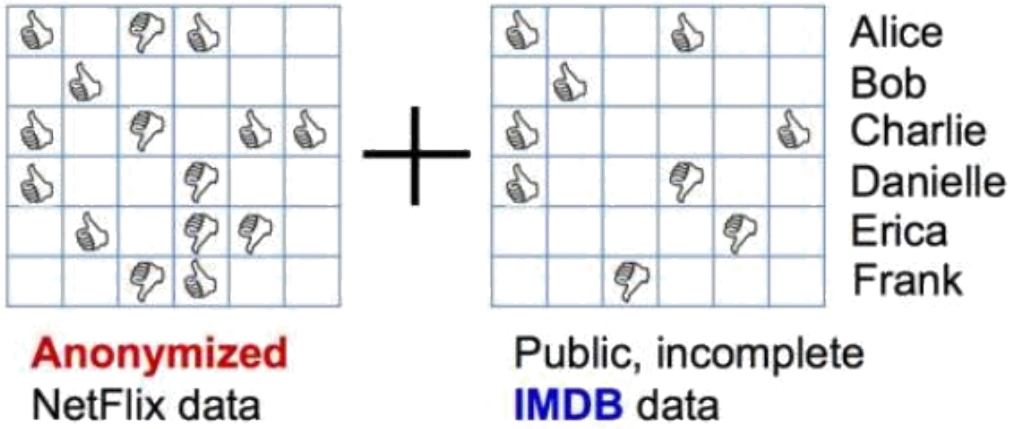
- Netflix Prize dataset, released 2006
  - 100,000,000 (private) ratings from 500,000 users
  - average 200 ratings per user
- Competition to improve recommendations
- Anonymized: usernames replaced by a number

Popular cases of data de-anonymization

- Problem: can combine “private” ratings from Netflix with public reviews from IMDB to identify users in dataset
- May expose private info about members…

Popular cases of data de-anonymization

Use Public Reviews from IMDB.com



Credit: Arvind Narayanan via Adam Smith

Narayanan, Shmatikov, Robust De-anonymization of Large Datasets (How to Break Anonymity of the Netflix Prize Dataset), 2008

Popular cases of data de-anonymization

- Fitness tracking app STRAVA showed heatmap of activities
- Built from one billion activities - some three trillion points of data, covering 27 billion km
- Published to enable people to check popular routes, etc.

https://www.theguardian.com/world/2018/jan/28/fitness-tracking-app-gives-away-location-of-secret-us-army-bases

Popular cases of data de-anonymization

- In remote locations, it was easy to de-anonymize the category of people staying there
- This data exposed the exercise routes of military personnel in bases around the world
- Active bases and maps were hence discovered

https://www.theguardian.com/world/2018/jan/28/fitness-tracking-app-gives-away-location-of-secret-us-army-bases

Popular cases of data de-anonymization

Other attacks:

- Su et al, De-anonymizing Web Browsing Data with Social Networks , 2017
  - Using links appearing in one's feed is unique
- Korolova, "Privacy Violations Using Microtargeted Ads: A Case Study", PADM
  - Attackers can instrument ad campaigns to identify individual users.
- Calandrino, Kilzer, Narayanan, Felten, Shmatikov, "You Might Also Like: Privacy Risks of Collaborative Filtering"
  - Attackers can infer customers' transaction with a limited amount of auxiliary data

Popular cases of data de-anonymization

- Lesson learnt: cannot always anonymize data simply by removing identifiers
- Vulnerable to aggregating data from multiple sources/networks

Outline

1. Private data generation and the need for anonymization
2. Privacy-preserving techniques
   - K-anonymity (and variants)
   - Differential Privacy
3. Tools for anonymization

Personal data attributes

- Name
- Home address
- Phone number
- E-mail address
- Age
- Biometrics (hair colour, eye colour, appearance, finger prints, etc.)

Other personal data attributes

- IP address on internet
- MAC address of sensors
- Identifiers of smartphones (IMSI; IMEI, etc.)
- GPS positions
- Photos showing me
- Photos I have made
- Text I wrote (on Facebook, in my blog, on Twitter, etc.)
- ...

Identifiers and quasi-indentifiers attributes

- **Identifiers**
  - Anything that might identify the person directly
  - E.g.: Name+Surname, phone number, email, ...
- **Quasi-identifiers**
  - Not unique, but sufficiently well correlated with a person such that they can be combined with other quasi-identifiers to create a unique identifier
  - Can be used for linking anonymized dataset with other datasets

Sensitive attributes

Data that the data subjects want to keep private, but often agencies/researchers/companies need to provide a service

- Physical conditions
  - Pregnancy, Diseases, Allergies, Physical limitations, …
- Social condition
  - Financial situation, Political opinions, Being recipient of government support, ...
- Sexual orientation
- ...

Classification of attributes

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

Re-identification by linking

[Latanya Sweeney, 1997]

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
		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

Public voter dataset

Voter List

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

k-Anonymity

And other cluster-based methods

k-Anonymity

- 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 person in the released table, and the information you have is his birth date, ZIP and gender. However there are k people in the table with the same combination of birth, ZIP and gender
- Any combination of quasi-identifier present in the released table must appear in at least k records

First defined:

[Samarati and Sweeney. "Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression", 1998]

k-Anonymity

[Samarati et al, 1998]

- In SQL, released table T is k-anonymous if each line of:

```
SELECT COUNT(*)
FROM T
GROUP BY quasi-identifier combination
```

is **≥ k**

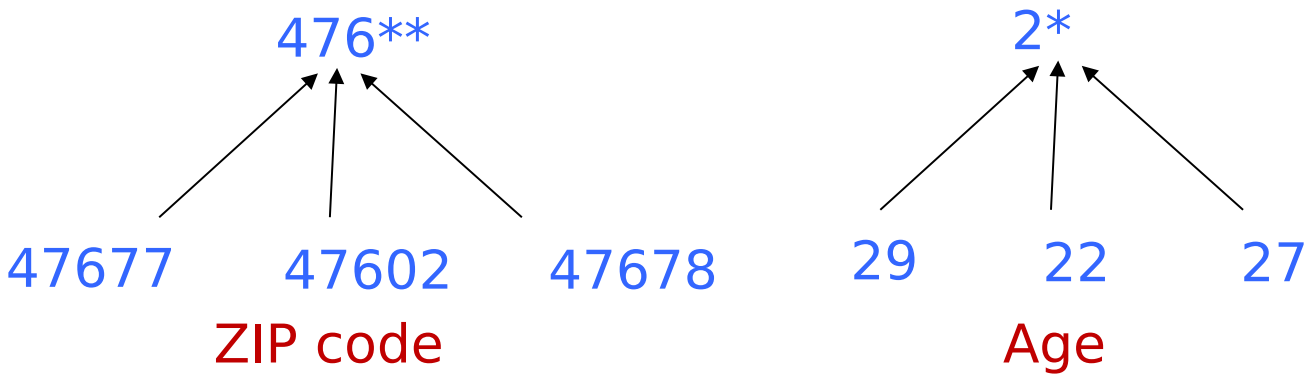
- Parameter k indicates the “degree” of anonymity

Achieving k-Anonymity

- Goal of k-Anonymity
  - Each record is indistinguishable from at least k-1 other records
  - These k records form an equivalence class
- **Generalize, modify, or supress** quasi-identifier values so that no individual is uniquely identifiable from a group of k

Achieving k-Anonymity: Generalization

- Generalization: replace quasi-identifiers with less specific, but semantically consistent values



Example of generalization

	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

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

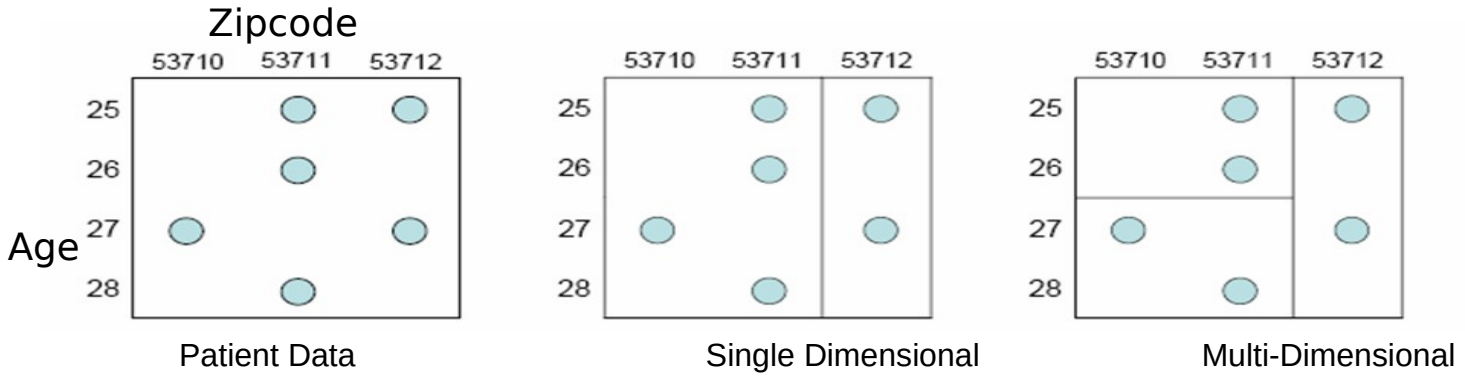
External data Source

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

Generalization algorithms

- There are tens of k-anonymization algorithms
- Example: *Greedy Partitioning Algorithm, Bucketization, ...*
- **Problem:** find multi-dimensional partitions of the data, where each partition has two or more data points (i.e. k=2)



- Optimal k-anonymous strict multi-dimensional partitioning is **NP-hard**
- Optimal = minimum information loss = maximum utility

Achieving k-Anonymity

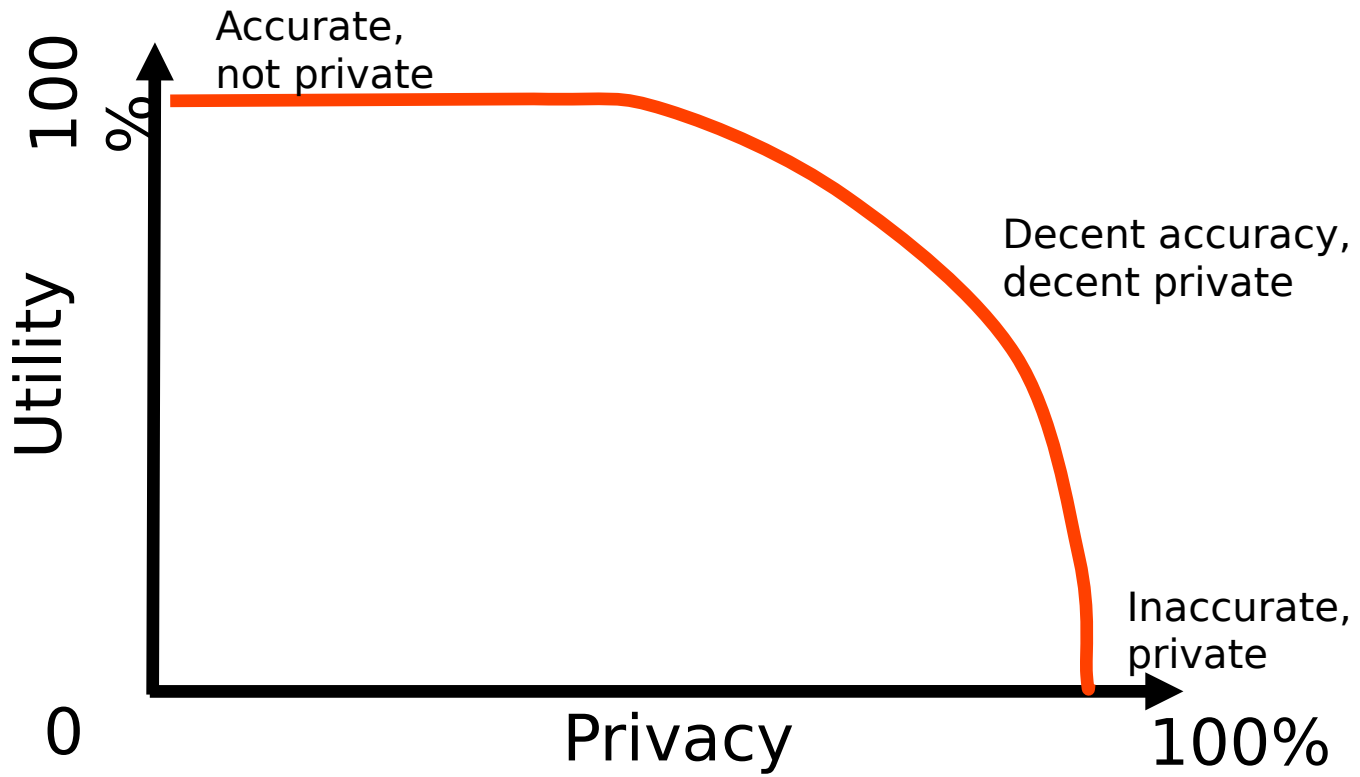
Generalization

- Replace specific quasi-identifiers with less specific values

Suppression

- When generalization causes too much information loss
  - This is common with “outliers”

Utility vs. Privacy



Determining the value of k

The value of k depends on

- Number of records in the table
- Number of quasi-identifiers
- The distribution of each quasi-identifier
- The relationship between quasi-identifier

Rule of thumb:

- k increases → privacy increases
- k increases → utility decreases

What is the best value of k?

2-Anonymized table

ID	Age	Sex	Zip	Disease
1	[25-26]	Male	53711	Flu
3	[25-26]	Male	53711	Brochities
2	[25-27]	Female	53712	Hepatitis
5	[25-27]	Female	53712	HIV
4	[27-28]	Male	[53710-53711]	Broken Arm
6	[27-28]	Male	[53710-53711]	Hang Nail

3-Anonymized table

ID	Age	Sex	Zip	Disease
1	[25-26]	*	5371*	Flu
3	[25-26]	*	5371*	Brochities
2	[25-26]	*	5371*	Hepatitis
5	[27-28]	*	5371*	HIV
4	[27-28]	*	5371*	Broken Arm
6	[27-28]	*	5371*	Hang Nail

4-Anonymized table

ID	Age	Sex	Zip	Disease
1	[25-28]	*	5371*	Flu
3	[25-28]	*	5371*	Brochities
2	[25-28]	*	5371*	Hepatitis
5	[25-28]	*	5371*	HIV
4	[25-28]	*	5371*	Broken Arm
6	[25-28]	*	5371*	Hang Nail

Curse of dimensionality

[Aggarwal, VLDB '05]

Generalization fundamentally relies on spatial locality

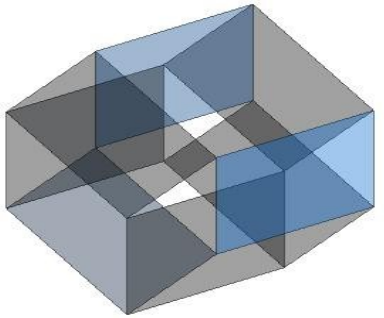
- Each record must have k close neighbors

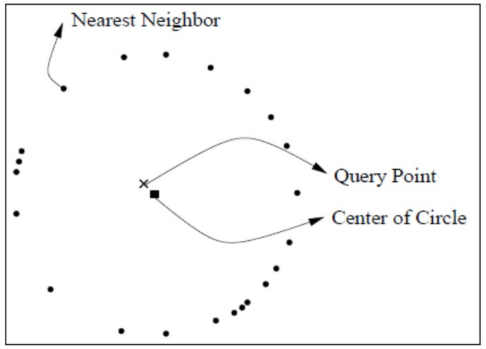
Real-world datasets are very sparse

- Many attributes (dimensions)
  - Netflix Prize dataset: 17,000 dimensions
  - Amazon customer records: several million dimensions
- "Nearest neighbor" is very far

Projection to low dimensions might lose interesting info

- k-anonymized datasets might results useless





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

Carl does not have heart disease

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

A solution: l-Diversity

[Machanavajjhala et al., ICDE '06]

Principle

- Each equi-class contains at least *l* well-represented sensitive values

Zipcode	Age	Disease
476**	2*	Acne
476**	2*	Heart disease
476**	2*	Flu
476**	2*	Heart disease
476**	2*	Flu
476**	2*	Flu
476**	3*	Flu
476**	3*	Acne
476**	3*	Cancer
476**	3*	Acne
476**	3*	Flu
476**	3*	Cancer

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

l-Diversity

Distinct l-Diversity

- Each equivalence class contains l distinct sensitive values

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)

…

Neither sufficient, nor necessary

Original dataset

…	HIV-
…	HIV-
…	HIV-
…	HIV+
…	HIV-
…	HIV-
…	HIV-
…	HIV-
…	HIV-
…	HIV-
…	HIV-
…	HIV+
…	HIV-

99% are HIV-

Anonymization A

Q1	HIV+
Q1	HIV-
Q1	HIV+
Q1	HIV+
Q1	HIV-
Q1	HIV-
Q2	HIV-
Q2	V-
Q2	
Q2	
Q	

50% HIV+ quasi-identifier group is “diverse”

This leaks a ton of information

Neither sufficient, nor necessary

Original dataset

…	HIV-
…	HIV-
…	HIV-
…	HIV+
…	HIV-
…	HIV-
…	HIV-
…	HIV-
…	HIV-
…	HIV-
…	HIV-
…	HIV+
…	HIV-

99% are HIV-

Anonymization A

Q1	HIV+
Q1	HIV-
Q1	HIV+
Q1	HIV+
Q1	HIV-
Q1	HIV-
Q2	HIV-
Q2	HIV+

Anonymization B

Q1	HIV-
Q1	HIV-
Q1	HIV-
Q1	HIV+
Q1	HIV-
Q1	HIV-
Q2	HIV-
Q2	HIV+

99% HIV- : quasi-identifier group is <u>not</u> “diverse”

…yet anonymized database does not leak anything

50% HIV+ quasi-identifier group is “diverse”

This leaks a ton of information

Limitations of l-Diversity

- Example: sensitive attribute is HIV+ (1%) or HIV- (99%)
  - Very different degrees of sensitivity!
- l-diversity is unnecessary
  - 2-diversity is unnecessary for an equivalence class that contains only HIV- 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

Skewness attack

- ~ probabilistic inference attacks
- 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!

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!

Further privacy: t-Closeness

k-anonymity prevents identity disclosure but not attribute disclosure

- To solve that problem **l-Diversity** requires that each eq. class has at least l values for each sensitive attribute
- **t-Closeness** requires that the distribution of a sensitive attribute in any eq. class is close to the distribution of a sensitive attribute in the overall table

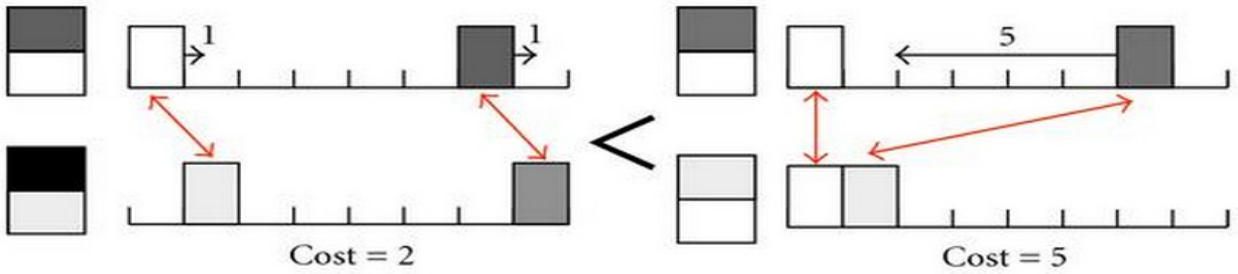
t-Closeness protects against attribute disclosure but not identity disclosure

t-Closeness

Distribution of sensitive attribute value in each equi-class should be close to that of the overall dataset (within a distance t)

- Earth Mover Distance to capture semantic relationship among sensitive attribute values
- Privacy is measured by the information gain of an observer regarding distribution of sensitive attributes

Information Gain = Posterior Belief – Prior Belief



[Li, Li, Venkatasubramanian, ICDE'2007]

Similarity attack example

	ZIP Code	Age	Salary	Disease
1	4767*	≤ 40	3K	gastric ulcer
3	4767*	≤ 40	5K	stomach cancer
8	4767*	≤ 40	9K	pneumonia
4	4790*	≥ 40	6K	gastritis
5	4790*	≥ 40	11K	flu
6	4790*	≥ 40	8K	bronchitis
2	4760*	≤ 40	4K	gastritis
7	4760*	≤ 40	7K	bronchitis
9	4760*	≤ 40	10K	stomach cancer

Table 5. Table that has 0.167-closeness w.r.t. Salary and 0.278-closeness w.r.t. Disease

Question: Why publish quasi-identifiers at all??

k-Anonymous, l-Diverse 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…

It is secure, right?

What does an attacker know?

Bob is Caucasian and I heard he was admitted to hospital with flu…

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

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

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

k-Anonymity can be harmful

- 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 the target
- Relies on locality
  - Destroys utility of many real-world datasets

Differential Privacy

A rigorous framework for privacy-preserving analysis of datasets

Why is anonymization hard?

It’s hard to guess what capabilities attackers will have, especially in the future

- Future datasets
- Future techniques
- Future computational power

Analogy with cryptography: cryptosystems today are designed based on what quantum computers might be able to do in 30 years

Why Differential Privacy (DP)?

- Strong, quantifiable, composable mathematical privacy **guarantee**
- **It is (by design) resilient** to known and unknown attack modes!
- DP enables many computations with personal data while preserving personal privacy

First defined:

[Dwork, McSherry, Nissim, and Smith, Calibrating Noise to Sensitivity in Private Data Analysis, in Third Theory of Cryptography Conference, TCC 2006.]

Earlier roots:

[*Warner*, Randomized Response 1965]

What to learn from data?

- **Differential privacy**: no harm in **participation**
  - Outcome of any analysis is essentially equally likely, **independent of whether any individual joins or not the dataset**

- **Usually, we want our observations to generalize to other data points → avoid overfitting**

Differential Privacy: Scenario

- **Worst-case scenario**: an attacker that knows everything, but not if the user is in the dataset
  - The attacker knows the value of every record, **even all the target user's one**
- The attacker can perform **queries** to the dataset
  - The answer of the dataset should not hint the presence of the user

Overly accurate answers to too many questions destroys privacy

[Dinur and Nissim, 2003]

Example: is Luca in the dataset?

Name	Height [m]
Amir	1.65
Esteban	1.73
Xin	1.76

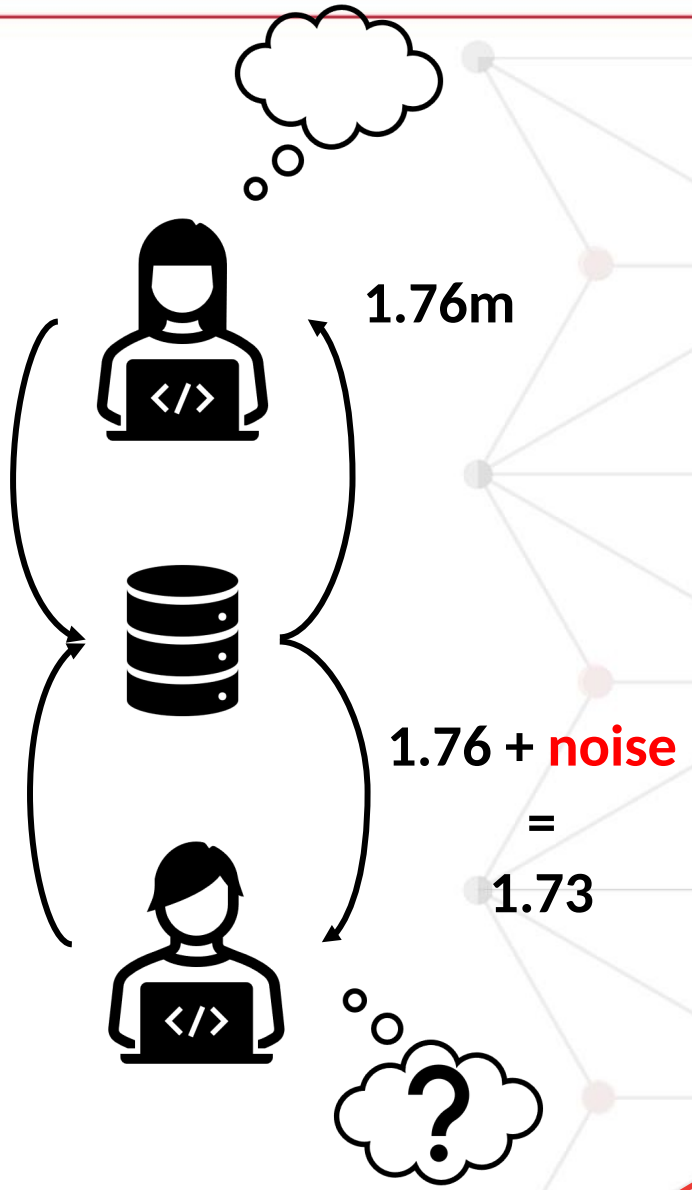
avg(height) → 1.71m

avg(height)?

Name	Height [m]
Amir	1.65
Esteban	1.73
Xin	1.76
Luca	**1.90**

avg(height) → 1.76m

avg(height)?

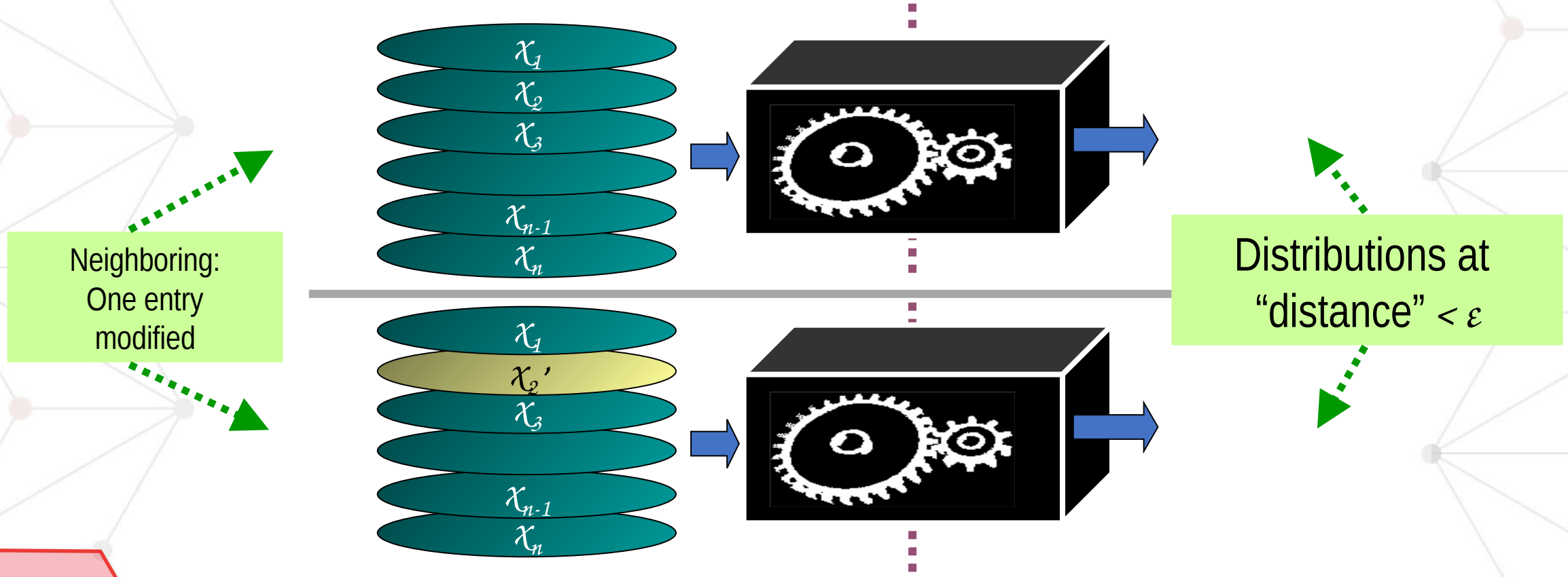


Differential privacy basic setup

- There is a database 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**
- Two databases D and D’ are **neighbouring** if they agree except for a single entry

Differential Privacy

[Dwork et al., 2006]

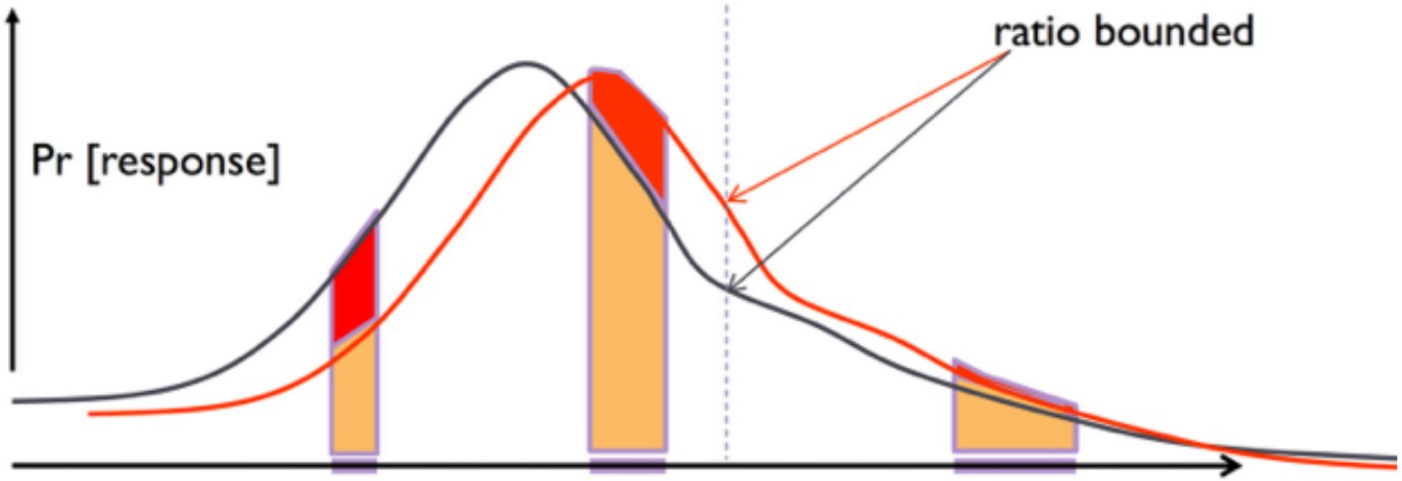


Formally...

[Dwork et al. 2006]

A query mechanism M is ϵ-differentially private if, for any two adjacent databases D and D' (differing in just one entry) and $C \subseteq range(M)$

$$\Pr(M(D) \in C) \leq e^{\epsilon} \cdot Pr(M(D') \in C)$$



- Output does **not overly depend on any single tuple**
- Participation in the dataset poses **no additional risk**

Privacy parameter ε

A query mechanism M is ϵ-differentially private if, for any two adjacent databases D and D' (differing in just one entry) and $C \subseteq range(M)$

$$\Pr(M(D) \in C) \leq e^{\epsilon} \cdot Pr(M(D') \in C)$$

ε is an arbitrary parameter and controls the privacy of the system

- Low ε, the two quantities are forced to be similar → more privacy
- High ε, the two quantities are allowed to diverge → less privacy

However, all queries are not the same

- Some of them may be more intrusive
- What differentiate them is ***sensitivity***

The sensitivity

- A query can be more or less intrusive on the privacy of the users
- Each query f has a **sensitivity**
  - Depends on the **difference** in output between f(D) and f(D')
  - **Global Sensitivity** $GS_f = \max_{\text{neighbors } D,D'} ||f(D) - f(D')||_1$
- The added noise depends on the query sensitivity
  - Sensitive query → more noise
  - Insensitive query → less noise

Implications

Global sensitivity: $GS_f = \max_{\text{neighbors } D,D'} ||f(D) - f(D')||_1$

Some queries, such as counting queries, can be answered relatively accurately

- Since one tuple affects the result by at most 1 (GS=1)
- A small amount of noise can be added to achieve DP

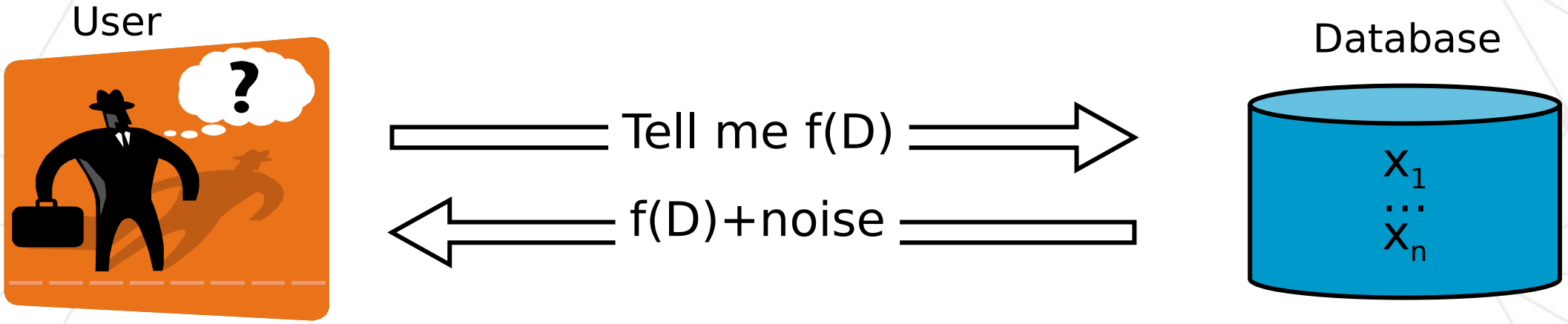
Some queries are hard to answer

- E.g., max, since it can be greatly affected by a single tuple (GS unbounded)

Challenge in using it

- Find suitable queries to ask so that noisy answers provide most utility

A DP mechanism: output perturbation



- Intuition: f(D) can be released accurately when f is insensitive to individual entries x_1, … x_n
- We want f(D) + noise to be ε-indistinguishable
- How this noise should be generated depending on ε and GS_f ?

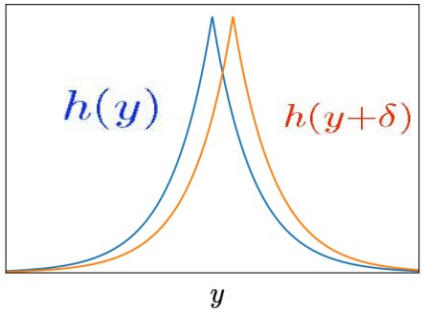
Laplace noise

Theorem

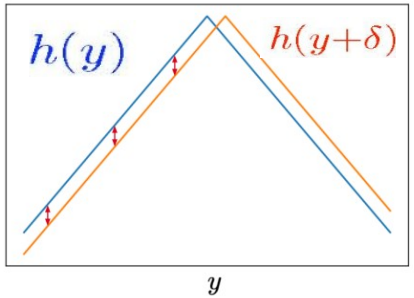
If $A(x) = f(x) + \mathsf{Lap}\left(\frac{\mathsf{GS}_f}{\varepsilon}\right)$ *then* A *is* ε*-indistinguishable.*

Laplace distribution $\mathsf{Lap}(\lambda)$ has density $h(y) \propto e^{-\frac{\|y\|_1}{\lambda}}$

h(y)



log(h(y))



It quantifies:

- Noise to add when function is more sensitive (higher GS_f)
- Noise to add if we want more privacy (lower ε)

Properties of Differential Privacy

Composability

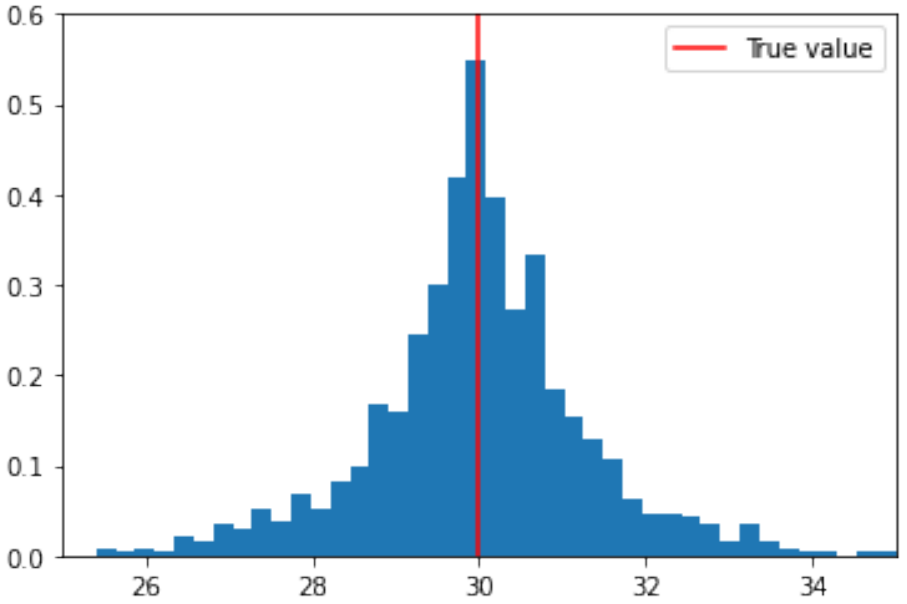
- Applying the sanitization several time yields a graceful degradation
- If A_1 satisfies ε_1-DP, and A_2 satisfies ε_2-DP, then outputting both A_1 and A_2 satisfies $(\varepsilon_1+\varepsilon_2)$-DP

Robustness to side information

- No need to specify **exactly** what the adversary knows
- Any **post-processing cannot improve** the attacker's knowledge

Multiple queries

- What if we allow people to perform the same query over and over again?
- Eventually, the noise will cancel out and the true value will emerge



Privacy budget

- To avoid noise cancelling, ε becomes a ***privacy budget*** **to query** the dataset
- The privacy consumption is <u>additive</u>
  - With budget 10, I can choose
    - 10x queries with $\varepsilon = 1$
    - 5x queries with $\varepsilon = 2$
    - 1x query with $\varepsilon = 10$

How to increase number of queries?

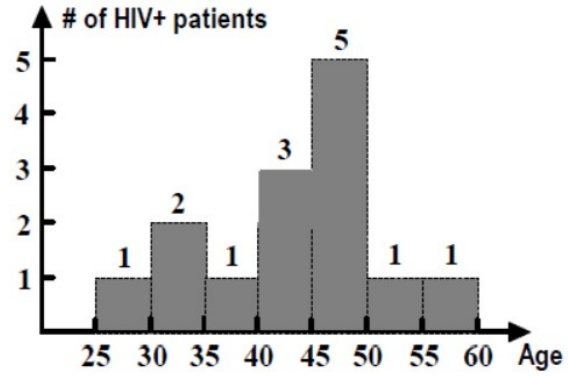
[Blum-Ligget-Roth, 2008]

Use **Coordinated Noise**

- If noise is added in with careful coordination, **rather than independently,** DP can save on the budget

Example: histogram queries with d bins:

- You can treat query independently: use d times the mechanism, hence noise Lap(d/ε)
- But actually only need Lap(1/ε), since sensitivity generalizes as max l_1 distance



A glimpse inside the rabbit hole

- What if we don't trust the data curator?
  - **Local** differential privacy, where data are anonymized before being uploaded to a platform
- What if we want to publish the dataset at once, without having to continuously answer to the queries?
  - **Non-interactive** differential privacy
- What if we have categorical values?
  - Other mechanisms exist: the **exponential mechanism** allows to choose the most-suited value inside a categorical set
- What if we assume the attacker does not know everything?
  - **Relaxed differential privacy**: only statistical knowledge of D and D', keeping precise knowledge on the element changing

DP: Pros & Cons

Pros 

- **Rigorous** mathematical definition of privacy
- **Flexible**: several mechanisms are available
- **Robust** to postprocessing
- The level of privacy ε can be **chosen** by the system administrator

Cons 

- **Precision** of the queries is affected
- Hard to **explain**
- Many non-trivial DP algorithms require really **large datasets** to be practically useful
- What ε and what privacy budget is **reasonable** for a dataset?

Differential privacy for ML

Protect Training data

- Train a model on private data → add noise to the training data
- Noise in the training set can destroy the utility of the model

Protect Training data in the ML model

- Against membership inference attacks
- Applying differential privacy on model outputs

Do I need DP if I don’t care about privacy?

- Statistics to generalize well should **not be dependent on single instances**
- As for machine learning: it should **not overfit** on training
- We can use DP to **ensure statistical validity of exploratory data**

Better Privacy = Better Data

Outline

1. Private data generation and the need for anonymization
2. Privacy-preserving techniques
   - K-anonymity (and variants)
   - Differential Privacy
3. Tools for anonymization

Tools for anonymization

1. ARX
2. IBM DiffPrivLib
3. Google Differential Privacy
4. P-PPA
5. TensorFlow Privacy
6. …

ARX Data Anonymization Tool

Open-source software for anonymizing sensitive personal data

Has a graphical interface and APIs

It supports:

- privacy models (K-anonymity, l-diversity, t-closeness)
- methods for transforming data
- methods for analyzing the usefulness of output data

ARX Data Anonymization Tool

Can be used as a library

- Written in Java
- All functionalities (privacy models and metrics) available to developers
- Documentation: https://arx.deidentifier.org/wp-content/uploads/javadoc/current/api/index.html

There is a project called ARX as a Service that offers ARX API as Web API:

- https://navikt.github.io/arxaas/

ARX Data Anonymization Tool

```
curl 'http://localhost:8080/api/anonymize' -i -X POST \
    -H 'Content-Type: application/json' \
    -d '{
  "data" : [ [ "age", "gender", "zipcode" ], [ "34", "male", "81667" ], [ "35", "female", "81668" ], [ "36", "male",
"81669" ], [ "37", "female", "81670" ], [ "38", "male", "81671" ], [ "39", "female", "81672" ], [ "40", "male", "81673" ],
[ "41", "female", "81674" ], [ "42", "male", "81675" ], [ "43", "female", "81676" ], [ "44", "male", "81677" ] ],
  "attributes" : [ {
    "field" : "age",
    "attributeTypeModel" : "IDENTIFYING",
    "hierarchy" : null
  }, {
    "field" : "gender",
    "attributeTypeModel" : "SENSITIVE",
    "hierarchy" : null
  }, {
    "field" : "zipcode",
    "attributeTypeModel" : "QUASIIDENTIFYING",
    "hierarchy" : [ [ "81667", "8166*", "816**", "81***", "8****", "*****" ], [ "81668", "8166*", "816**", "81***",
"8****", "*****" ], [ "81669", "8166*", "816**", "81***", "8****", "*****" ], [ "81670", "8167*", "816**", "81***",
"8****", "*****" ], [ "81671", "8167*", "816**", "81***", "8****", "*****" ], [ "81672", "8167*", "816**", "81***",
"8****", "*****" ], [ "81673", "8167*", "816**", "81***", "8****", "*****" ], [ "81674", "8167*", "816**", "81***",
"8****", "*****" ], [ "81675", "8167*", "816**", "81***", "8****", "*****" ], [ "81676", "8167*", "816**", "81***",
"8****", "*****" ], [ "81677", "8167*", "816**", "81***", "8****", "*****" ] ]
  } ],
  "privacyModels" : [ {
    "privacyModel" : "KANONYMITY",
    "params" : {
      "k" : "5"
    }
  }, {
    "privacyModel" : "LDIVERSITY_DISTINCT",
    "params" : {
      "column_name" : "gender",
      "l" : "2"
    }
  } ],
}'
```

```
HTTP/1.1 200 OK
Vary: Origin
Vary: Access-Control-Request-Method
Vary: Access-Control-Request-Headers
Content-Type: application/json
Content-Length: 7813

{
  "anonymizeResult" : {
    "data" : [ [ "age", "gender", "zipcode" ], [ "*", "male", "816**" ], [ "*", "female", "816**" ],
[ "*", "male", "816**" ], [ "*", "female", "816**" ], [ "*", "male", "816**" ], [ "*", "female",
"816**" ], [ "*", "male", "816**" ], [ "*", "female", "816**" ], [ "*", "male", "816**" ], [ "*",
"female", "816**" ], [ "*", "male", "816**" ] ],

....

  "riskProfile" : {
    "reIdentificationRisk" : {
      "measures" : {
        "estimated_journalist_risk" : 0.09090909090909091,
        "records_affected_by_highest_prosecutor_risk" : 1.0,
        "sample_uniques" : 0.0,
        "lowest_risk" : 0.09090909090909091,
....
}
```

ARX Data Anonymization Tool

How to use it in Python:

1. Launch an ARX Server, using the Dockerized version

```
docker run -p 8080:8080 navikt/arxaas
```

2. Install pyarxaas

```
pip install pyarxaas
```

3. Use it in Python

```
arxaas = ARXaaS(url)
…
anon_result = arxaas.anonymize(dataset, [kanon])
```

IBM DiffPrivLib

Library written in Python

Developed by IBM

Dedicated to **differential privacy** *and* ***machine learning***

Allow experimentation, simulation, and implementation of differentially private models using a common codebase and building blocks

Documentation at: https://diffprivlib.readthedocs.io/en/latest/

IBM DiffPrivLib

Supports:

- Basic Mechanisms for Differential Privacy
  - Laplacian mechanism
  - Geometric Mechanism
  - Exponential Mechanism
- Tools for Differential Privacy functions
  - Mean, Sum, Histogram

Differentially-private machine learning models

- Classification models
  - Gaussian Naive Bayes
  - Logistic Regression
  - Random Forest
- Regression models
  - Linear Regression
- Clustering models
  - K-Means
- Dimensionality reduction models
  - PCA
- Preprocessing
  - Standard Scaler

IBM DiffPrivLib

Differentially Private Histograms

Plotting the distribution of ages in `Adult`

In [1]:
```
import numpy as np
from diffprivlib import tools as dp
import matplotlib.pyplot as plt
```

We first read in the list of ages in the Adult UCI dataset (the first column).

In [2]:
```
ages_adult = np.loadtxt("https://archive.ics.uci.edu/ml/machine-learning
-databases/adult/adult.data",
                        usecols=0, delimiter=", ")
```

From:
https://github.com/IBM/differential-privacy-library/blob/main/notebooks/histograms.ipynb

Differentially private histograms

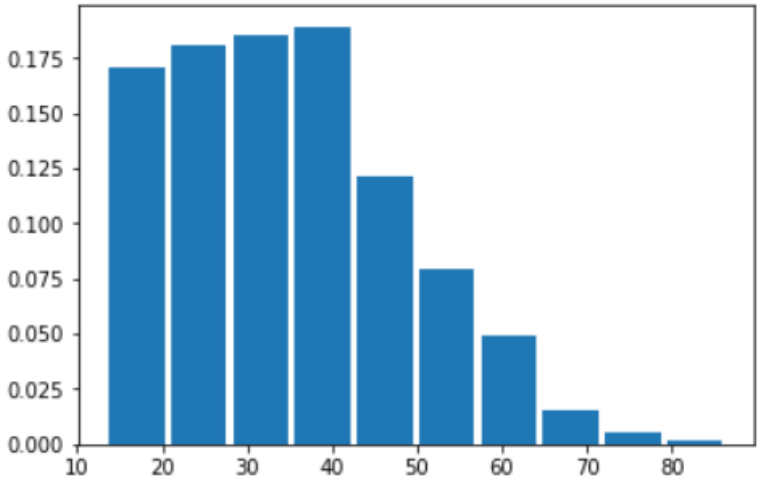
Using `diffprivlib`, we can calculate a differentially private version of the histogram. For this example, we use the default settings:

- `epsilon` is 1.0
- `range` is not specified, so is calculated by the function on-the-fly. This throws a warning, as it leaks privacy about the data (from `dp_bins`, we know that there are people in the dataset aged 17 and 90).

```
dp_hist, dp_bins = dp.histogram(ages_adult)
dp_hist = dp_hist / dp_hist.sum()

plt.bar(dp_bins[:-1], dp_hist, width=(dp_bins[1] - dp_bins[0]) * 0.9)
plt.show()
```

```
.../site-packages/diffprivlib/tools/histograms.py:131: PrivacyLeakWarnin
g: Range parameter has not been specified. Falling back to taking range
from the data.
To ensure differential privacy, and no additional privacy leakage, the r
ange must be specified independently of the data (i.e., using domain kno
wledge).
  "specified independently of the data (i.e., using domain knowledge).",
PrivacyLeakWarning)
```



IBM DiffPrivLib

Differentially Private Histograms

Plotting the distribution of ages in `Adult`

In [1]:
```
import numpy as np
from diffprivlib import tools as dp
import matplotlib.pyplot as plt
```

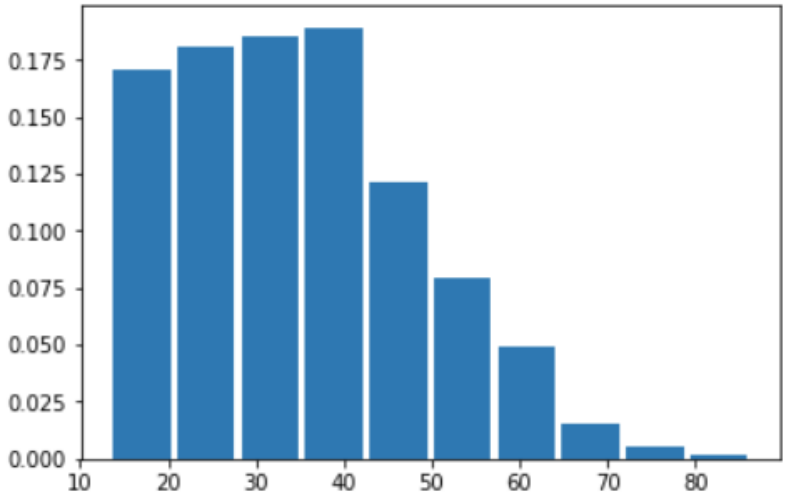
We first read in the list of ages in the Adult UCI dataset (the first column).

In [2]:
```
ages_adult = np.loadtxt("https://archive.ics.uci.edu/ml/machine-learning
-databases/adult/adult.data",
                        usecols=0, delimiter=", ")
```

:
```
hist, bins = np.histogram(ages_adult)
hist = hist / hist.sum()
```

Using `matplotlib.pyplot`, we can plot a barchart of the histogram distribution.

:
```
plt.bar(bins[:-1], hist, width=(bins[1]-bins[0]) * 0.9)
plt.show()
```



Google Differential Privacy

Library by Google to generate DP statistics over datasets

https://github.com/google/differential-privacy

Written in C

Implements various mechanisms

- Laplace, Gaussian, etc.

Various Statistics:

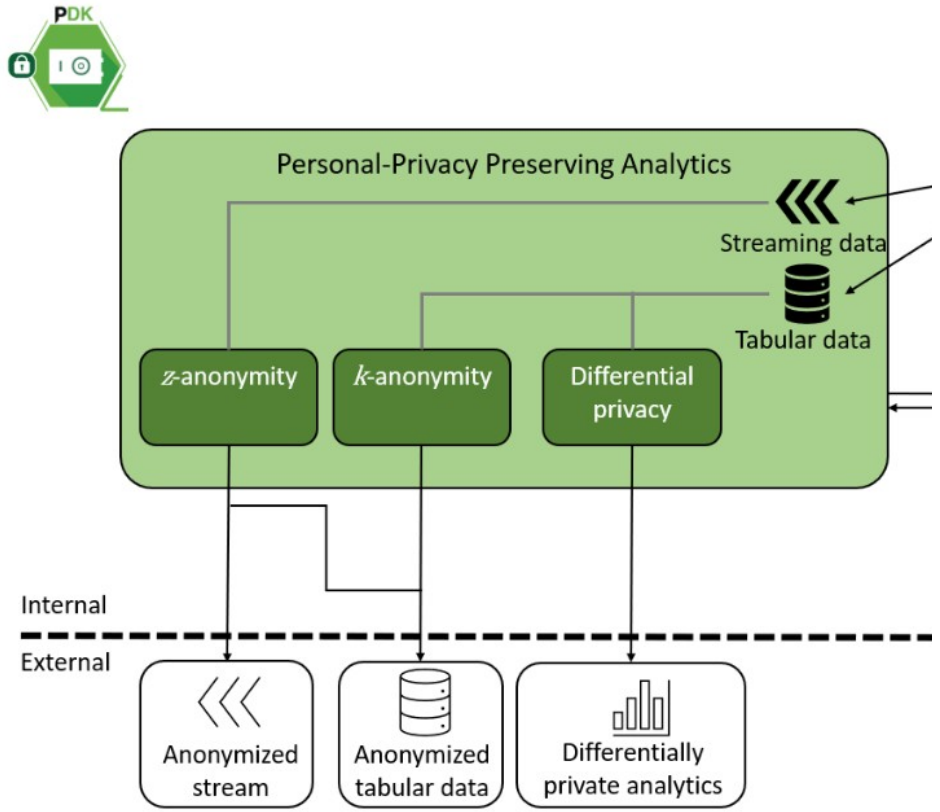
- Count, Sum, Mean, Variance, Quantiles…

There exist Python bindings:

- https://github.com/OpenMined/PyDP

Personal Privacy Preserving Analytics

Developed by Politecnico di Torino
In the context of the EU Project PIMCity (https://www.pimcity-h2020.eu/)
It is a simple Python module that offers easy-to-use privacy models



Personal Privacy Preserving Analytics

- Written in Python
- Simple installation, only a few Python `requirements`
- Usage a Python module

```
from algorithms.kanonymity.mondrian.mondrian import Mondrian
mondrian = Mondrian(3, user_choice_index=[2,3,4,5])
k_anonymized_dataframe = mondrian.perform(input_dataframe)
```

Personal Privacy Preserving Analytics

- The P-PPA offers Web APIs to allow remote use

Personal-Pivacy Preserving Analytics 1.0.0 OAS3

The PIMCity P-PPA, it's a tool to allow data analysts and stakeholders to retrieve useful information from the data, while preserving the privacy of the users whose data are in the studied datasets. It follows some query example.

k-anonymity Provides k-anonymity privacy guarantees.

GET `/kanon` According to the passed parameters, it permits internal PPPA modules to collect data from the requested source and provide k-anonymized data.

differential privacy Provides data statistics in a differentially private flavour.

GET `/diffpriv` According to the passed parameters, it permits internal PPPA modules to collect data from the requested source and provide differentially private data statistics.

Tools summary

TOOL	Language	K-anon & similar	Diff Priv	Web API	Input
ARX	Java	Y	Partially	Y	CSV, Excel, SQL, JSON, Java API
DiffPrivLib	Python		Y		Numpy
Google Differential Privacy	C/C++		Y		C vars
P-PPA	Python	Y	Y	Y	Pandas, CSV, SQL

Credits

- Martino Trevisan – Università di Trieste
- Nikhil Jha – Politecnico di Torino
- Ashwin Machanavajjhala - Duke University
- Vitaly Shmatikov - Cornell Tech
- Chris Clifton – Purdue
- Muhamad Felemban – KFUPM
- Moni Naor- Weizmann Institute of Science
- Adam Smith - Penn State
- Roger Grosse - University of Toronto
- Katrina Ligett - California Institute of Technology
- Cynthia Dwork - Harvard University
- Matteo Maffei – TU Wien
- Eduardo Cuervo – Oculos
- Amre Shakimov - Duke University
- Frank McSherry - Materialize, Inc.
- Krishnaram Kenthapadi - AI @ LinkedIn
- Ilya Mironov - Google AI
- Abhradeep Thakurta - UC Santa Cruz


Savage Chickens
by Doug Savage
I'M TIRED OF THE ANONYMITY OF LIVING IN THE CITY
WHERE DO YOU LIVE?
AROUND
©2006 BY DOUG SAVAGE
www.savagechickens.com


Perguntas
Fragen
Domande
Galdera
Otázky
Questions
Spørgsmål
Pertanyaan
kysymykset
Frågor
Spørsmål
Cwestiynau
вопросы
Preguntes
Sorular
Въпроси
Vragen
Pytania