

INFORMATION SECURITY

(WBCS004-05)

Fatih Turkmen, PhD



TODAY

- Data Privacy
- Anonymization Techniques
- Pseudonymization Techniques
- Differential Privacy
- Location Privacy

DATA PRIVACY

“Privacy forms the basis of our freedom. You have to have moments of reserve, reflection, intimacy, and solitude”

Dr. Ann Cavoukian

Many definitions are there... but in IT, it is related to

- *The collection, handling and dissemination of sensitive data*

What is considered as private? Typically;

- Personally identifiable information (PII)
- And personal health information (PHI)

“Some”
Sensitive

“Some” Non-
Sensitive

DATA PRIVACY

- From Personal Data : Any information that can be used to identify a person
 - Examples: Name/Surname, Phone Number, E-mail address, Passport Number, Social Security Number...
- To Private Data: Anything that “should not” be made available to the public
 - The gray area that appears increasingly more Friends/Posts/”Followee” on Social Networks, Places you visited, IP Addresses, Products you bought, Diagnosis/Treatments you received...

EXAMPLE PRIVACY BREACH

In 2007, Netflix released a dataset of their user ratings as part of a competition to see if anyone can outperform their collaborative filtering algorithm.

The dataset did not contain personally identifying information!!

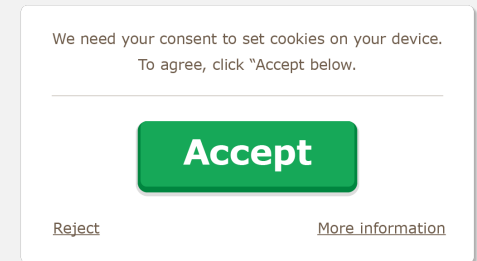
But researchers were still able to breach privacy; they recovered 99% of all the personal information that was removed from the dataset [3].

In this case, the researchers breached privacy using auxiliary information from IMDB.

- General Data Protection Regulation
 - cf: <http://www.eugdpr.org/>
 - Aim: protect all EU citizens from privacy and data breaches
 - subjects must have given their consent
 - data may only be used for intended purpose
 - no additional data may be collected
 - integrity, confidentiality are required



What is GDPR?



LEGAL BASIS (CONT.)



• EU-US Privacy Shield:

Protects the fundamental rights of anyone in the EU whose personal data is transferred to the United States for **commercial purposes**. It allows the free transfer of data to companies that are certified in the US under the Privacy Shield.



On July 16 2020, the European Court of Justice (ECJ) ruled to [invalidate the EU-US Privacy Shield agreement on data sharing](#), on the grounds that the US is not a safe haven for EU citizens' data due to disproportionate surveillance practices.



• Long live EU-US Data Privacy Framework (DPF)

One of the main differences between them is the emphasis on **transparency**. DPF requires participating companies to publicly disclose their privacy policies and the third-party service providers they use, and introduces new binding safeguards ensuring access by **U.S. intelligence is allowed** only to the extent necessary and proportionate ... to handle/resolve complaints from Europeans ... for national security purposes...

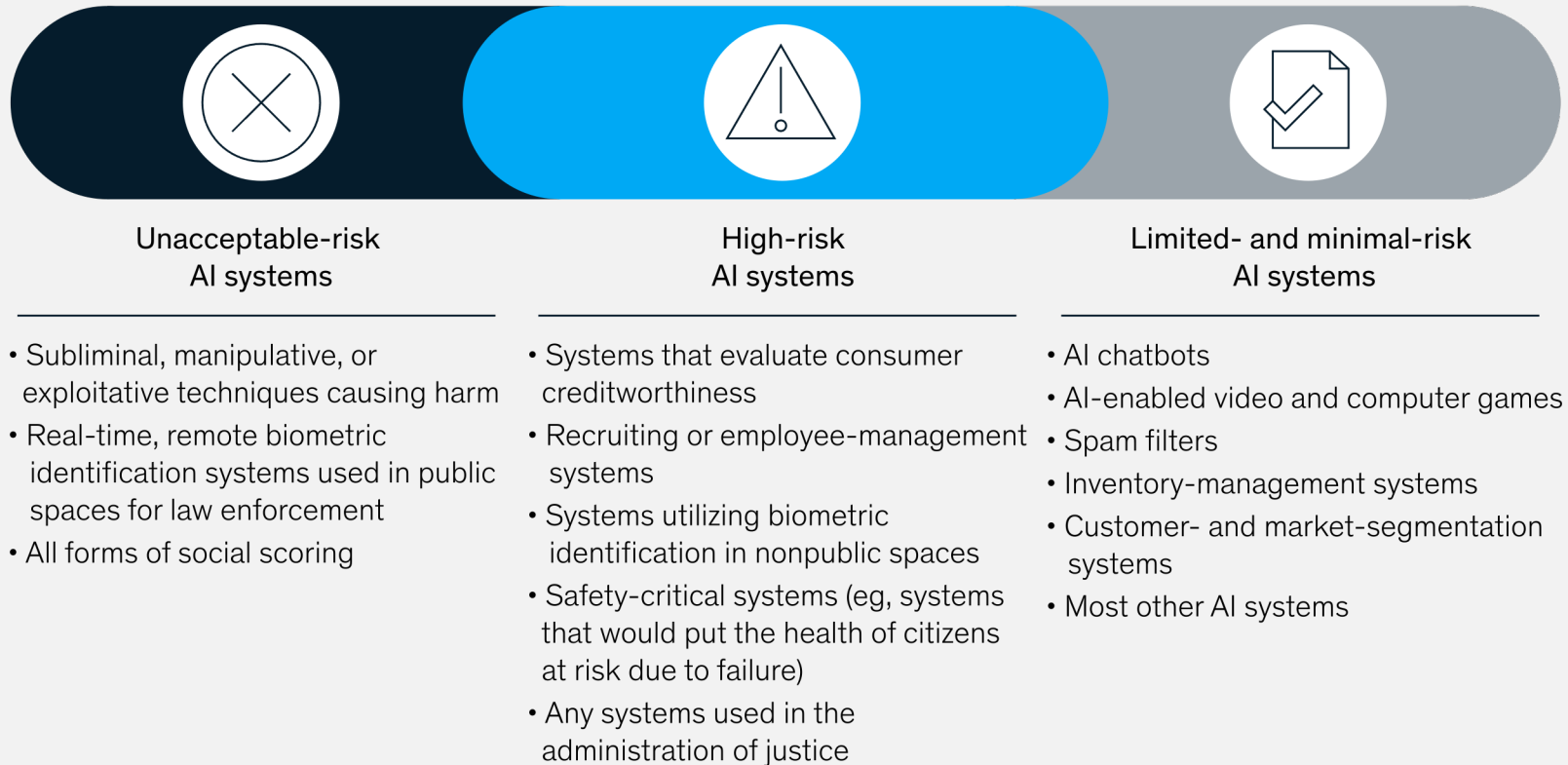
LEGAL BASIS (CONT.)

What is coming up?



- Artificial Intelligence Act (Upcoming)
 - One step ahead
 - More focus on algorithmic discrimination
 - Risk-based approach
- EU Recommendations on Ethics and data protection :
https://ec.europa.eu/info/funding-tenders/opportunities/docs/2021-2027/horizon/guidance/ethics-and-data-protection_he_en.pdf

A PEEK AT ARTIFICIAL INTELLIGENCE ACT

The European Union's draft AI regulations classify AI systems into three risk categories.



PSEUDONYMIZATION VS ANONYMIZATION

- **Anonymization**  What are they? irreversible
- **Pseudonymization**  requires additional information to re-identify the data subject


Name	Token/Pseudonym	Anonymized
Clyde	qOerd	XXXXX
Marco	Loqfh	XXXXX
Les	Mcv	XXXXX
Les	Mcv	XXXXX
Marco	Loqfh	XXXXX
Raul	BhQl	XXXXX
Clyde	qOerd	XXXXX

Courtesy of [5]

K-ANONYMITY

k - 1

- *Idea: Given a dataset with private info about individuals, k-anonymity requires that each individual/record cannot be distinguished from **k-1** individuals/record of the same dataset.*
 - *Each record has 1/k probability of **re-identification***
- *One of the earliest on **balance between utility and privacy***
- *Relevant Concepts:*
 - *Quasi-identifiers and key (unique) attributes*
 - *Sensitive attributes*
 - *Release Table*



Take a note of this,
“re-identification”!

K-ANONYMITY

k - 1

Google Cloud

Why Google Solutions Products Pricing Getting Started

Cloud Data Loss Prevention Overview **Guides** Reference Samples Support Resources

Using the command-line tool
Using a JSON request
Scheduling inspection scans
Creating an inspection template
De-identifying and re-identifying sensitive text

Samples
All Data Loss Prevention code samples
All code samples for all products

How-to guides
All how-to guides
Authentication
▸ Inspection
▸ InfoType detectors
▸ De-identification
▾ Re-identification risk analysis
 Overview
 Computing k-anonymity for a dataset
 Computing l-diversity for a dataset
 Computing k-map for a dataset
 Computing δ-presence for a dataset
 Computing numerical and categorical statistics
 Visualizing re-identification risk using Data Studio
▸ Scheduling jobs
▸ Working with DLP scan results
 Securing Cloud DLP resources

Concepts
All concepts
Actions
▸ De-identification
 Hybrid jobs and job triggers
▸ InfoTypes
 Job triggers

Cloud Data Loss Prevention > Documentation > Guides

Was this helpful?

Computing k-anonymity for a dataset

[Send feedback](#)

K-anonymity is a property of a dataset that indicates the re-identifiability of its records. A dataset is *k*-anonymous if quasi-identifiers for each person in the dataset are identical to at least $k - 1$ other people also in the dataset.

You can compute the *k*-anonymity value based on one or more columns, or fields, of a dataset. This topic demonstrates how to compute *k*-anonymity values for a dataset using Cloud Data Loss Prevention (DLP). For more information about *k*-anonymity or risk analysis in general, see the [risk analysis concept topic](#) before continuing on.

★ **Note:** Prematurely canceling an operation midway through a job still incurs costs for the portion of the job that was completed. For more information about billing, see [Cloud DLP pricing](#).

Before you begin

Before continuing, be sure you've done the following:

1. [Sign in](#) to your Google Account.
2. In the Google Cloud Console, on the project selector page, select or create a Google Cloud project.
[Go to the project selector](#)
3. Make sure that billing is enabled for your Google Cloud project. [Learn how to confirm billing is enabled for your project](#).
4. Enable Cloud DLP.
[Enable Cloud DLP](#)
5. Select a BigQuery dataset to analyze. Cloud DLP calculates the *k*-anonymity metric by scanning a BigQuery table.
6. Determine an identifier (if applicable) and at least one quasi-identifier in the dataset. For more information, see [Risk analysis terms and techniques](#).

K-ANONYMITY

k - 1

Unique ID		Quasi-identifiers (QIs)		Sensitive	
Name	Age	ZipCode	Diseases	...	
Betty	45	21344	Cancer		
Dan	45	21344	HIV		
Simon	32	21340	Flu		
Rachel	24	21330	Tuberculosis		
Ellen	45	21344	Hepatitis		
Bill	32	21340	Cancer		
Hugh	24	21330	Tuberculosis		
Axel	18	21330	Tuberculosis		
Curtis	32	21340	Hepatitis		
John	44	21340	Hepatitis		

Let's assume for a minute that the names are unique...

- Objective: Allow some useful information, e.g., statistics about diseases, without allowing the re-identification of individuals
 - The adversary has only access to QIs
- The table is released after applying certain anonymization methods
- Several methods exist:
 - Suppression: Hide individual attributes (e.g., replace with *)
 - Generalization: Generalize individual attributes to a broader category (e.g., ZipCode = 2134*).
 - Omission: Remove the individual record as a whole

K-ANONYMITY EXAMPLE

k - 1

- We want to release the table after some anonymization
- ✓ First **suppress** *unique* attributes

Name	Age	ZipCode	Diseases
*	45	21344	Cancer
*	45	21344	HIV
*	32	21340	Flu
*	24	21330	Tuberculosis
*	45	21344	Hepatitis
*	32	21340	Cancer
*	24	21330	Tuberculosis
*	18	21330	Tuberculosis
*	32	21340	Hepatitis
*	44	21340	Hepatitis

K-ANONYMITY EXAMPLE

k - 1

✓ Apply suppression and generalization to the rest

✓ Each quasi-identifier tuple appears in at least k records (build an equivalence class)



What anonymity do we have here wrt QIs shown before, i.e., $k=?$

$k=3$

Name	Age	ZipCode	Diseases
*	>40	2134*	Cancer
*	>40	2134*	HIV
*	3*	2134*	Flu
*	<30	2133*	Tuberculosis
*	>40	2134*	Hepatitis
*	3*	2134*	Cancer
*	<30	2133*	Tuberculosis
*	<30	2133*	Tuberculosis
*	3*	2134*	Hepatitis
*	>40	2134*	Hepatitis

As noted by [7]

- K-anonymity *ensures that each quasi-identifier tuple occurs in at least k records in the anonymized database and nothing more!*
- It **does not guarantee any privacy** in reality, because the values of sensitive attributes associated with a given quasi-identifier may not be sufficiently diverse (you release the whole table after all!)
- Or the adversary may know more than just the quasi identifiers [20].
- Furthermore, k-anonymization completely fails on high-dimensional datasets [2], such as the Netflix Prize dataset and most real-world datasets of individual recommendations and purchases.

ATTACKS ON K-ANONYMITY



- Further issues with k-anonymity

- Homogeneity: Attacker knows the ZipCode (21330) and Age (18) of a person

Name	Age	ZipCode	Diseases
*	<30	2133*	Tuberculosis
*	<30	2133*	Tuberculosis
*	<30	2133*	Tuberculosis

- Background Knowledge: Attacker knows that, Ellen has a very low risk of STD

Name	Age	ZipCode	Diseases
*	>40	2134*	Cancer
*	>40	2134*	Hepatitis

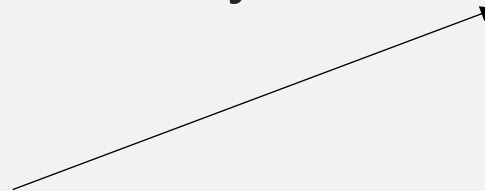
L-DIVERSITY

/

- Objective: Generate diversity of sensitive attributes in given quasi-identifier groups (denoted as q^*)
- Distinct l-diversity: There are at least l distinct **sensitive values** in each equivalence class
- Probabilistic l-diversity: The frequency of the most frequent value is bounded by $1/l$



Where do we mainly need diversity here?



Diseases
Cancer
HIV
Flu
Tuberculosis
Hepatitis
Cancer
Tuberculosis
Tuberculosis
Hepatitis
Hepatitis

L-DIVERSITY EXAMPLE

/

	Non-Sensitive			Sensitive
	Zip Code	Age	Nationality	Condition
1	1305*	≤ 40	*	Heart Disease
4	1305*	≤ 40	*	Viral Infection
9	1305*	≤ 40	*	Cancer
10	1305*	≤ 40	*	Cancer
5	1485*	> 40	*	Cancer
6	1485*	> 40	*	Heart Disease
7	1485*	> 40	*	Viral Infection
8	1485*	> 40	*	Viral Infection
2	1306*	≤ 40	*	Heart Disease
3	1306*	≤ 40	*	Viral Infection
11	1306*	≤ 40	*	Cancer
12	1306*	≤ 40	*	Cancer



What is the diversity level here (i.e., l)?

3-diverse Data set (l -distinct) [4]

- Entropy l-diversity: the entropy of the distribution of sensitive values in each q^* is at least $\log(l)$

Given a quasi-identifier group q^* with l distinct values for the sensitive attribute, s is a sensitive attribute value from domain S and $p(E, s)$ is the fraction of records in q^* with the sensitive attribute value s

$$\frac{n(q^*, s)}{\sum_{s' \in S} n(q^*, s')}$$

the fraction of tuples in the equivalence class E with sensitive attribute value equal to s

$$\text{Entropy}(q^*) = - \sum_{s \in S} p(q^*, s) \log(p(q^*, s))$$

We want $\text{Entropy}(q^*) \geq \log l$

Break

Tool to play with for the anonymity techniques: **ARX**
(<https://arx.deidentifier.org/>)

ATTACKS ON L-DIVERSITY



- Similarity Attack: If the l-diverse group has similar values
- Skewness Attack: If the l-diverse group sensitive values are skewed from other l-diverse groups.

	Zip Code	Age	Salary	Disease
1	476**	2*	3k	negative
2	476**	2*	4k	negative
3	476**	2*	5k	negative
4	476**	2*	6k	negative
5	4790*	>=40	7k	negative
6	4790*	>=40	8k	positive
7	4790*	>=40	9k	negative
8	4790*	>=40	10k	positive
9	476**	3*	11k	positive
10	476**	3*	12k	positive
11	476**	3*	13k	positive
12	476**	3*	14k	negative
13	4770*	4*	15k	negative
...
10,000	488**	>=60	16k	negative



Do you see something strange here?

	ZIP Code	Age	Salary	Disease
1	47677	29	3K	gastric ulcer
2	47602	22	4K	gastritis
3	47678	27	5K	stomach cancer
4	47905	43	6K	gastritis
5	47909	52	11K	flu
6	47906	47	8K	bronchitis
7	47605	30	7K	bronchitis
8	47673	36	9K	pneumonia
9	47607	32	10K	stomach cancer

Table 3. Original Salary/Disease Table

	ZIP Code	Age	Salary	Disease
1	476**	2*	3K	gastric ulcer
2	476**	2*	4K	gastritis
3	476**	2*	5K	stomach cancer
4	4790*	≥ 40	6K	gastritis
5	4790*	≥ 40	11K	flu
6	4790*	≥ 40	8K	bronchitis
7	476**	3*	7K	bronchitis
8	476**	3*	9K	pneumonia
9	476**	3*	10K	stomach cancer

Table 4. A 3-diverse version of Table 3

See [8] for more details

An equivalence class is t-close:

- If the distance between the distribution of a sensitive attribute in this class and the distribution of the attribute in the whole/original table is no more than a threshold t .
- A table is said to have t-closeness if all equivalence classes have t-closeness.

T-CLOSENESS

t

Original Table has a
distribution of values

Diseases
Cancer
HIV
Flu
Tuberculosis
Hepatitis
Cancer
Tuberculosis
Tuberculosis
Hepatitis
Hepatitis

Distribution of
sensitive attributes
(disease) in each
equivalence
group(X, Y, Z) should be
“ t ” close to the *original*
database

Name	Age	ZipCode	Diseases
Betty	45	21344	Cancer
Dan	45	21344	HIV
Simon	32	21340	Flu
Rachel	24	21330	Tuberculosis
Ellen	45	21344	Hepatitis
Bill	32	21340	Cancer
Hugh	24	21330	Tuberculosis
Axel	18	21330	Tuberculosis
Curtis	32	21340	Hepatitis
John	44	21340	Hepatitis
...	Cancer
...	HIV
...	Flu
...	Tuberculosis
...	Hepatitis
...	Cancer
...	Tuberculosis
...	Tuberculosis
...	Hepatitis

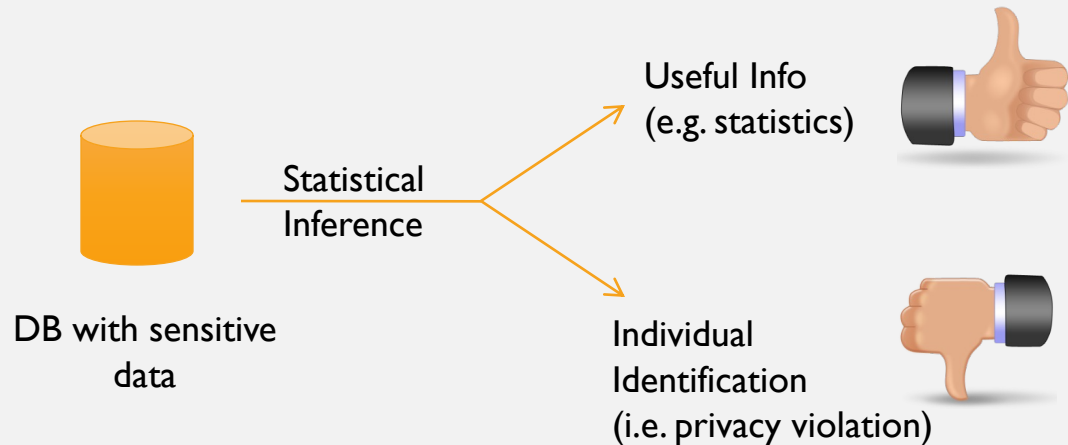
DIFFERENTIAL PRIVACY

ε

- Lots of publicly available data, many statistical studies...

- Concrete Examples:

- Medical records of a Governor
- IMDB + Netflix → user identification
- Individual Identification from AOL search queries



- **Paradoxical situation:** Learning nothing about an individual while learning useful information about a population

DIFFERENTIAL PRIVACY : DEFINITION

ϵ

- Differential privacy aims to solve the problem:
“The risk associated with privacy violation of an individual should not substantially increase as a result of participating in a statistical database”
- Here we need an algorithm/mechanism K (a differentially private mechanism) that for all pairs of very similar data sets D and D' , it will behave approximately the same on both data sets: If $K(D) = X$ and $K(D') = Y$ then X and Y should be indistinguishable.
- Most common method: Add controlled noise to data (e.g. Laplace, Gaussian..)

DIFFERENTIAL PRIVACY FORMALIZATION

ϵ

- A randomized function K gives ϵ -differential privacy if for all data sets D and D' differing on one entry and all $S \subseteq \text{Range}(K)$, s.t.

$$\Pr [K(D) \in S] \leq \exp(\epsilon) \cdot \Pr(K(D') \in S)$$

in other words, let $R \in S$ as above :

$$\frac{\Pr(K(D_k) = R)}{\Pr(K(D_{k \mp 1}) = R)} \leq e^\epsilon$$

- Function Sensitivity

$$\max_{D, D'} \|Q_D - Q_{D'}\|_1$$

DIFFERENTIAL PRIVACY FORMALIZATION

ϵ, δ

- A randomized function K gives (ϵ, δ) -differential privacy (i.e., Epsilon/Delta privacy) if for all data sets D and D' differing on one entry and all $S \subseteq \text{Range}(K)$, s.t.

$$\Pr [K(D) \in S] \leq \exp(\epsilon) \cdot \Pr(K(D') \in S) + \delta$$

DIFFERENTIAL PRIVACY FORMALIZATION

- Probability **Mass** Function (pmf) where X and Y are **discrete** random variables, i.e., $X, Y \in \{2.6, 2.8, 3.0, 3.3 \dots\}$
- Probability **Density** Function (pdf) where X and Y are **continuous** random variables, i.e., X and Y are in a range such as $2.8 \leq X \leq 3.0 \dots$
- So K is really about the distribution of values in its range ($K(D)$) for the data sets it is applied, i.e., the addition of noise.

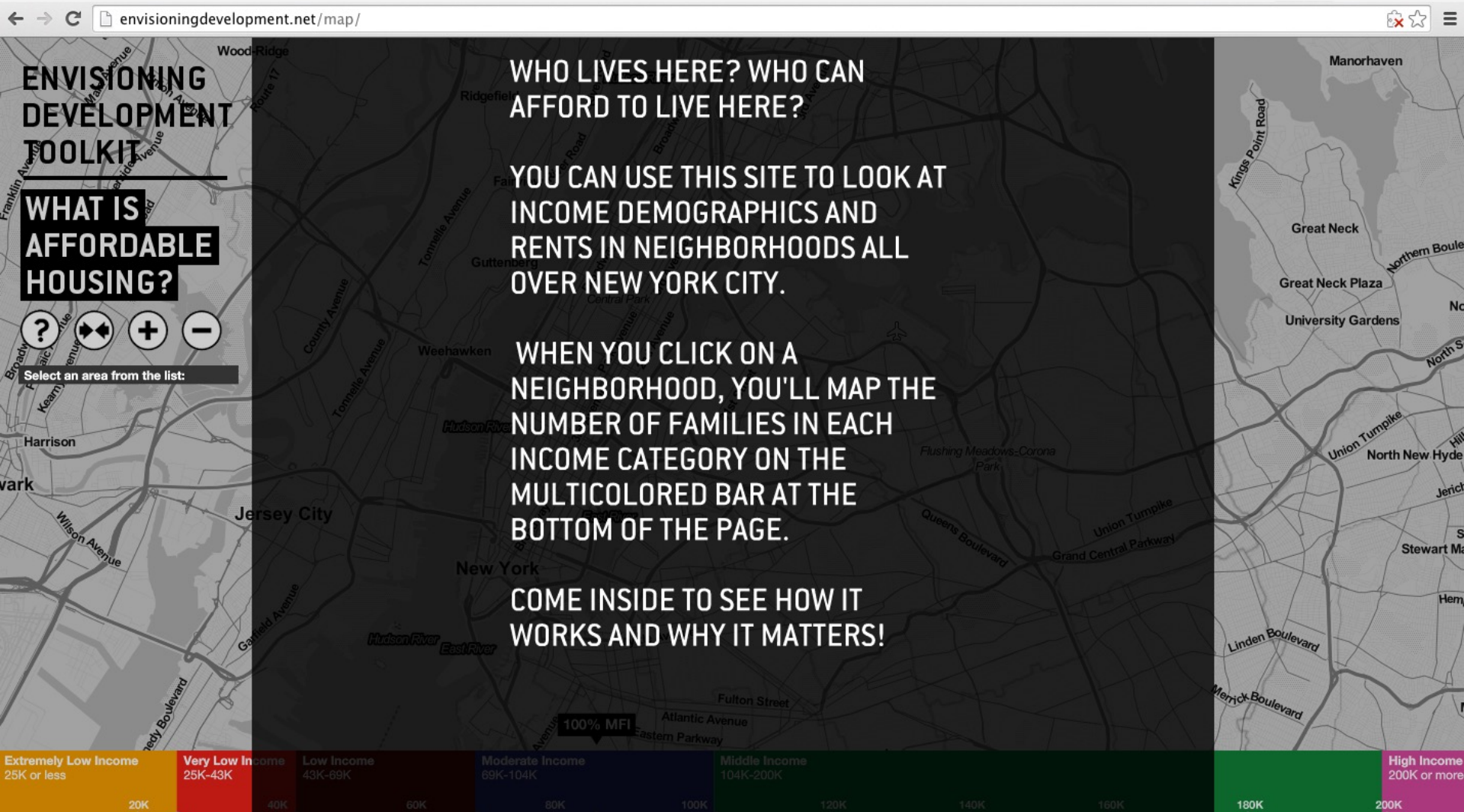
EXAMPLE

- Let's start with an example:

Suppose you have access to a database that allows you to compute the total income of all residents in a certain area. If you knew that Mr.White was going to move to another area from his current location, simply querying this database before and after his move would allow you to deduce his income.

Example by the courtesy of <http://research.neustar.biz/2014/09/08/differential-privacy-the-basics/>

PUBLIC DATABASES



PUBLIC DATABASES



BACK TO EXAMPLE

- Assume table represents the residents of the selected region (where Mr White is going to move)

→ 100 residents (+1 with Mr White)

- Adversary has a query ($Q(i)$) mechanism to get the sum of income up to the given row i . Run Q before and after Mr White moves...

- $Q(101) - Q(100) = \text{Mr White's income}$

Id	Name	Income
1	John Malkovich	80K
2	Jamal Malik	90K
3	Amelie Steiner	100K
4	Mirko Stanavic	75K
5	Mike Stanley	140K
6	Mehmet Uzun	90K
7	Stijn Neuer	60K
..
..

BACK TO EXAMPLE

- If K behaves as expected then we have a guarantee that whether an individual is in a given data set or not will not effect the outcome of a query significantly.
- $Q(5) = 485K$, $Q(6) = 575K$
- $K(Q(5)) = X$, $K(Q(6)) = Y$
- How do we define X and Y ?

Id	Name	Income
1	John Malkovich	80K
2	Jamal Malik	90K
3	Amelie Steiner	100K
4	Mirko Stanavic	75K
5	Mike Stanley	140K
6	Mehmet Uzun	90K
7	Stijn Neuer	60K
..
..

DIFFERENTIAL PRIVACY WITH LAPLACE

Let Δf denote the sensitivity of a function f .

(the maximum difference in the values $f(D)$ and $f(D')$,

for D and D' , a pair of databases that differ in only one row)

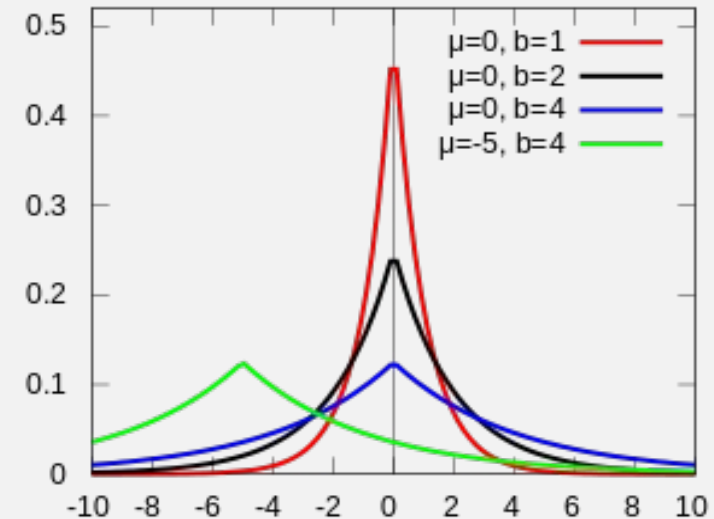
$$\max_{D, D'} \|Q_D - Q_{D'}\|_1$$

- Laplace Mechanism: Add controlled noise with $\text{Laplace}(\mu, b)$:

$$\text{Lap}(x \mid \mu, b) = \frac{1}{2b} e^{-\left(\frac{|x-\mu|}{b}\right)}$$

where

- $b = \Delta f / \epsilon$ (calibrating the noise to the function's sensitivity)
- μ refers to distance to function's true value (often set to 0)

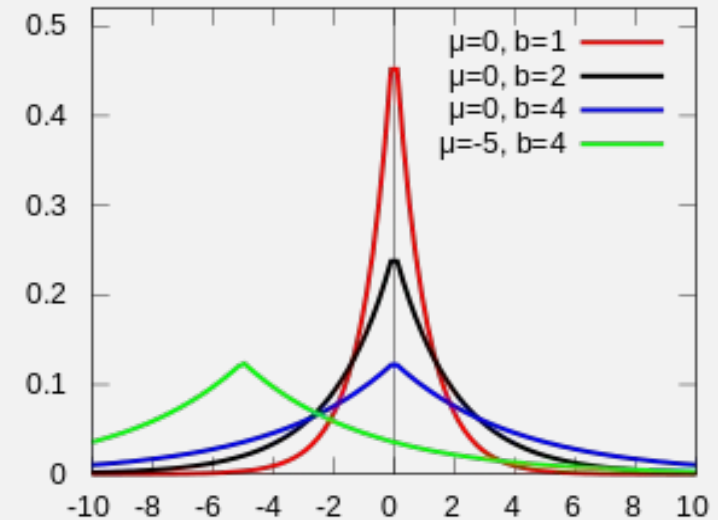


Courtesy of
https://en.wikipedia.org/wiki/Laplace_distribution

DIFFERENTIAL PRIVACY WITH LAPLACE (CONT.)

- Now, given the noise η drawn from the Laplace distribution, the result of the function is:

$$K(D) = f(D) + \eta$$



What happens with the change of ϵ ? (i.e., more privacy and/or more accuracy?)

Courtesy of
https://en.wikipedia.org/wiki/Laplace_distribution

LOCATION PRIVACY

- Your locations (history) tell a lot about you: habits, interests, health conditions, relations, political views...
- Pervasive devices, customized services, social connections, simplicity and instinct of information sharing
- Zillions of applications/services based on location, i.e. location-based services

See [6] for more details

MANY LBSS



shopkick™



foursquare



Path



facebook®

loopt®



NEARBY

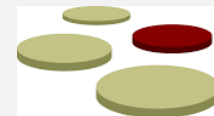
Geolqi

tinder™



GROUPON

place
pop

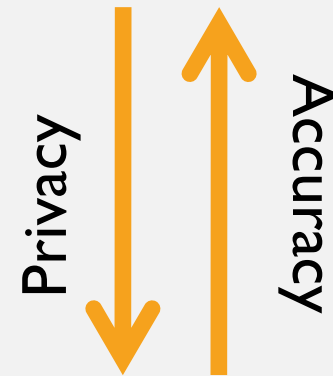


OpenTable™

LOCATION PRIVACY

- Location information from sensing devices
 - Accuracy
 - Privacy
- Trade-off between benefit from service and protection of location info from adversary

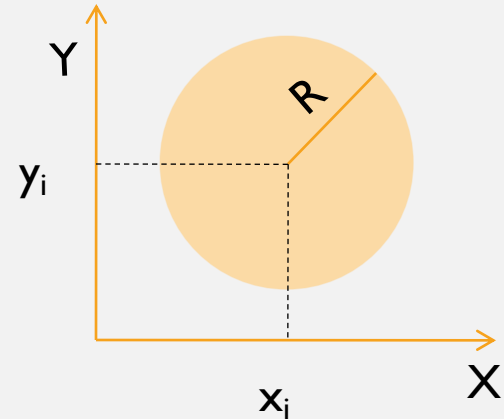
Bad **Privacy** Protection



Bad Service Experience
(**Utility**)

MODELING LOCATION INFORMATION

- Several models exist, for instance planar circular areas : (x_i, y_i, R)
 - Exact user location (x_i, y_i)
 - R is the radius introduced by the sensor initially
- Temporal dimension (x_i, y_i, R, t) when relevant: user is at (x_i, y_i, R) at time t



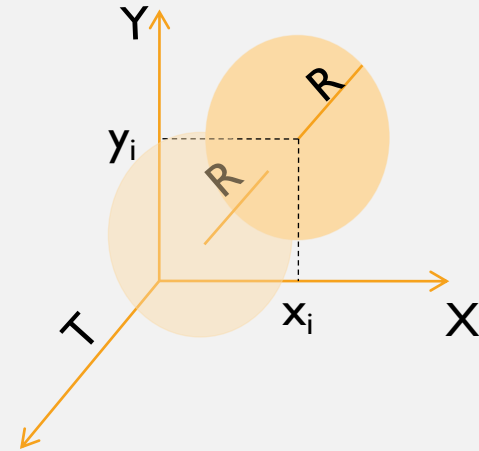
LOCATION OBFUSCATION (SPATIAL)

- *Location obfuscation (blurring, cloaking)* is one way: Exact user location is hidden within a region
- By using several obfuscation operators
 - Enlarge: according to some metric, increase the radius
 - Reduce: according to some metric reduce the radius
 - Shift: Change the exact location info

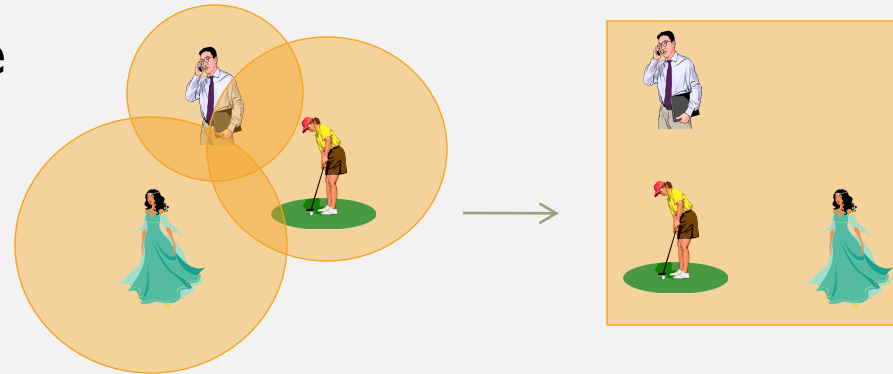


OTHER OBFUSCATION TECHNIQUES

- Spatio-temporal : Instead of (or in addition to) pure location obfuscation, a (time) delay can be introduced to location information



- Data-dependent obfuscation:
Calculate the region by considering $k-1$ nearest subject locations (to the requested location), i.e. generate a region that contains k subjects.

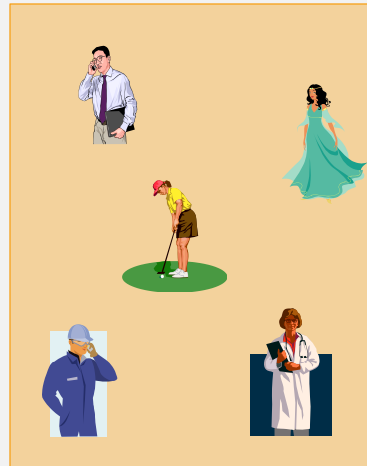
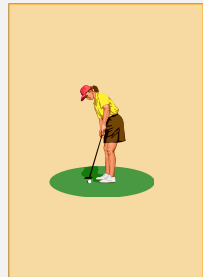


OTHER OBFUSCATION TECHNIQUES

- K-anonymity: at least K users in the obfuscated region

Dense places (stadium) → small region

Scarce places (desert) → large region

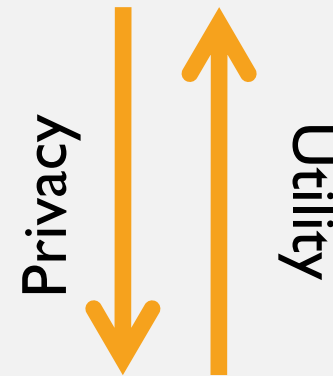


What anonymity
do we have?

WHAT DID WE LEARN?

- Inference from statistical databases
- k-anonymity, l-diversity, t-closeness
- Differential Privacy
- Location Privacy

Bad **Privacy** Protection



Bad Service Experience

REFERENCES

- [1] <https://www.protegrity.com/pseudonymization-vs-anonymization-help-gdpr/>
- [2] <https://desfontain.es/privacy/k-anonymity.html>
- [3] Latanya Sweeney, k-Anonymity: A Model for Protecting Privacy. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems 10(5): 557-570 (2002)
- [4] Ashwin Machanavajjhala, Daniel Kifer, Johannes Gehrke, Muthuramakrishnan Venkatasubramanian, L-diversity: Privacy beyond k-anonymity. TKDD 1(1) (2007)
- [5] Ninghui Li, Tiancheng Li, Suresh Venkatasubramanian, t-Closeness: Privacy Beyond k-Anonymity and l-Diversity. ICDE 2007: 106-115

REFERENCES

[6] Claudio Agostino Ardagna, Marco Cremonini, Sabrina De Capitani di Vimercati, Pierangela Samarati: An Obfuscation-Based Approach for Protecting Location Privacy. IEEE Trans. Dependable Sec. Comput. 8(1): 13-27 (2011)

[7] Arvind Narayan, Vitaly Shmatikov: Robust De-anonymization of Large Sparse Datasets. IEEE Symposium on Security and Privacy 2008 : 111-125,
https://www.cs.utexas.edu/~shmat/shmat_oak08netflix.pdf

[8] <https://www.ics.uci.edu/~projects/295d/presentations/295d-tcloseness#:~:text=t%2Dcloseness%20definition,equivalence%20classes%20have%20t%2Dcloseness>