An introduction to Anonymization

Benjamin NGUYEN & Sara TAKI
LIFO/INSA Centre Val de Loire & Univ Orléans
Inria PETSCRAFT
GDR Sécurité Informatique, Privacy WG

What am I working on with my group ?

Anonymization

- Anonymization algorithms evaluation
- Attacks on anonymization algorithms
- Anonymization algorithms for specific data (RDF)
- Organisation of anonymization and reidentification competitions

Privacy concepts

- Data minimization of data collected through forms (logic & game theory)
- Formal models of attackers and knowledge collected (modal logic)
- Attack models for distributed computations on graphs (graph theory & distribution)
- Better quantification of risk in differential privacy
- PETS for various applications (including medical examples)

Software

You may want to download :

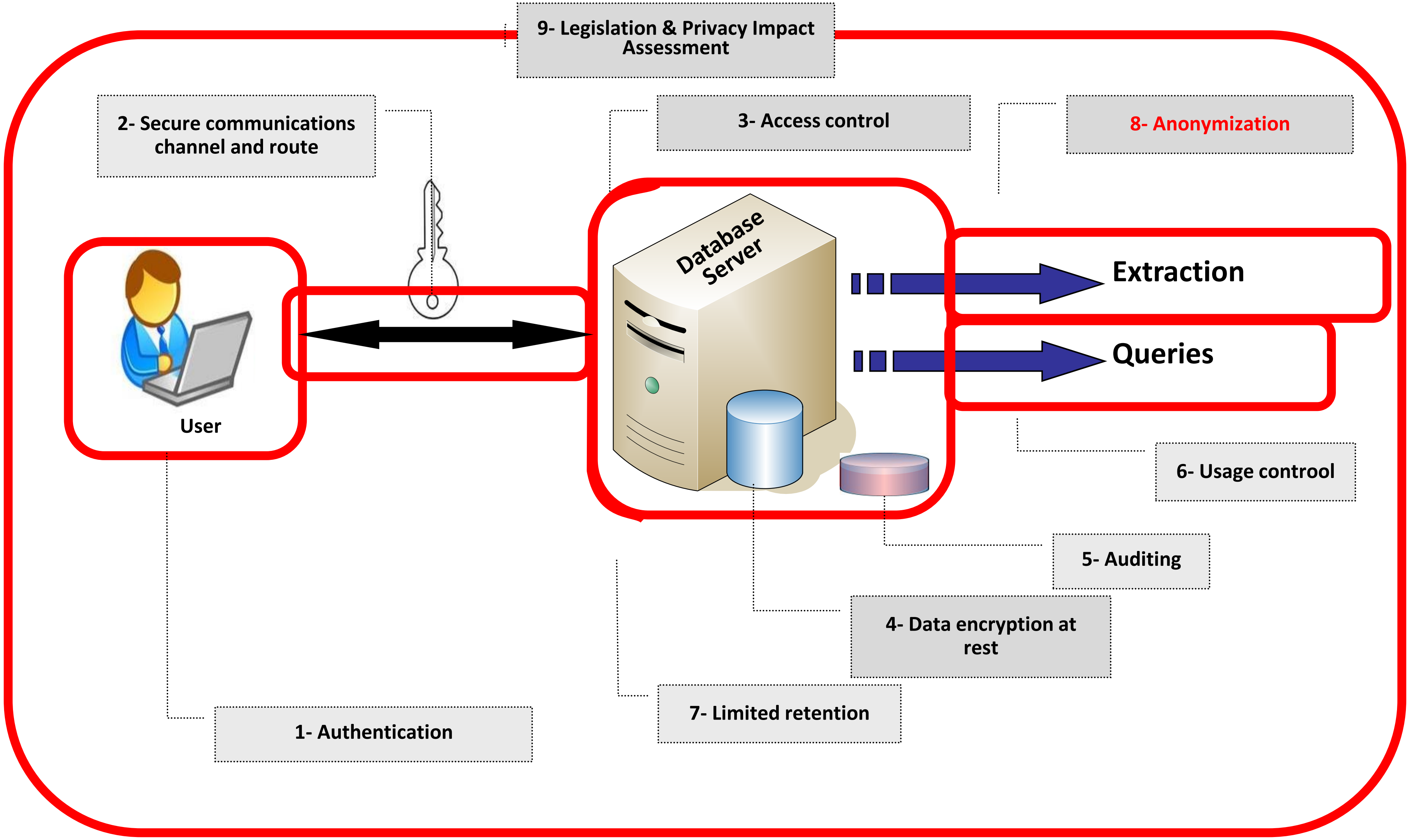
1. ARX Deidentification tool https://arx.deidentifier.org/downloads/
Open source software published by TU München (DE)

2. WEKA : https://waikato.github.io/weka-wiki/downloading_weka/
Open source software published by Univ. Waikato (NZ)

3. Diffprivlib : https://github.com/IBM/differential-privacy-library
MIT Licence (open source) published by IBM (US). Python library.

Suggested installation : executable JAR file for ARX & WEKA

Main defense mechanisms



Two approaches when processing personal data

- Keep identifiable data & respect the GDPR and other laws : Do a Privacy Impact Assessment (PIA, see : https://www.cnil.fr/fr/outil-pia-telechargez-et-installez-le-logiciel-de-la-cnil)
- Use anonymous data (the rest of this tutorial)

Outline

1. Pseudonymization
2. Anonymization architectures
3. K-anonymity : a historical anonymization technique
4. Reidentification risk
5. Aggregation based anonymization techniques
6. Statistical anonymization techniques
7. Differential privacy
8. Hands on *k*-anonymity using ARX (Sara)

1- Pseudonymization …

… is not anonymization

GDPR Recital 26

[1]The principles of data protection should apply to any information concerning an identified or identifiable natural person. [2]Personal data which have undergone pseudonymisation, which could be attributed to a natural person by the use of additional information should be considered to be information on an identifiable natural person. [3]To determine whether a natural person is identifiable, account should be taken of **all the means reasonably likely to be used**, such as singling out, either by the controller or by another person to identify the natural person directly or indirectly. [4]To ascertain whether means are reasonably likely to be used to identify the natural person, account should be taken of **all objective factors, such as the costs of and the amount of time required for identification, taking into consideration the available technology at the time of the processing and technological developments**. [5]The principles of data protection should therefore not apply to anonymous information, namely information which does not relate to an identified or identifiable natural person or to personal data rendered anonymous in such a manner that the data subject is not or no longer identifiable.

This means :

Anonymized data must provide *irreversible* de-indentification, assuming means reasonably used.

An original “semi decidable” definition.

Thus : if any entity is *easily* able to re-identify a dataset, this shows that the dataset was not anonymized (see DARC/INSAnonym)

Pseudonymization :
Is not anonymization for the GDPR

Individuals

Raw data

SSN	**Activity**	**Age**	**Diag**
123	"M2 Stud."	22	Flu
456	"M1 Stud."	25	HIV
789	"L3 Stud."	20	Flu

One way or
Two-way
Pseudonymization
function

Curator
(trusted)

Pseudonymized Data

Pseudo	**Activity**	**Age**	**Diag**
ABC	"M2 Stud."	22	Flu
MNO	"M1 Stud"	25	HIV
XYZ	"L3 Stud."	20	Flu

Third parties
(untrusted)

Trusted server ?

- **First problem :** maybe the editor has stored information in order to inverse the transformation.
- In this case, the data is not anonymized. GDPR mechanisms should apply to such (personal) data.
- **Secon problem :** Reidentification attacks on pseudonymization

2- Anonymization architectures

Investigating the anonymization process

Anonymization taxonomy

Privacy Preserving Data Publishing (PPDP)

Interactive methods
Adaptive query answering
On a hidden dataset

Non-interactive methods
Releasing a « sanitized » dataset

Centralized publication
The dataset is analysed
prior to anonymization

Local perturbation
Data is anonymized
Independently by all
data sources / providers

Centralized anonymized data publication

Context

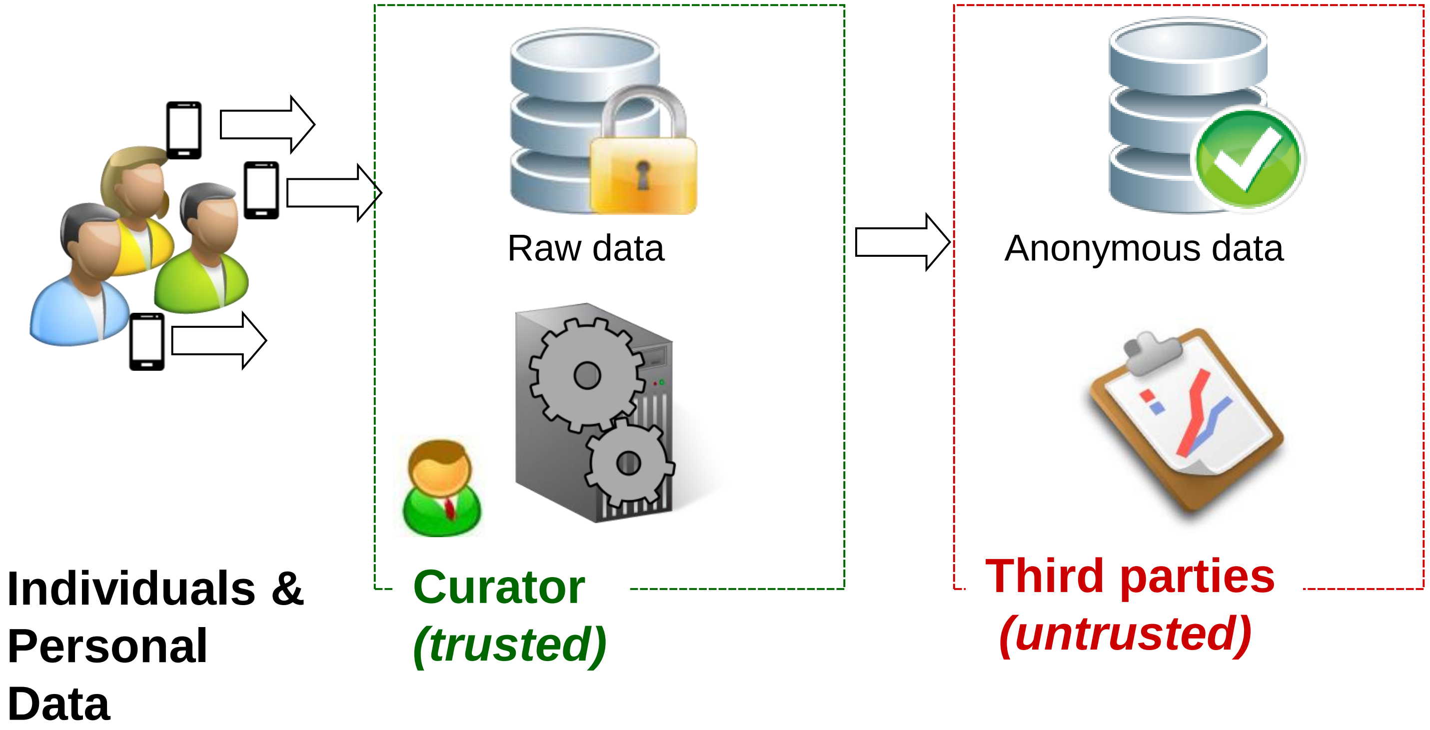
Personal data produced by sensors, forms, mobile phones, etc

Objective

Execute data intensive queries (agregates, AI, …)

Constraints

- Impossible to use an interactive query response system
- Publish the resulting *sanitized* dataset
- Choose an anonymization mechanism



http://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2014/wp216_en.pdf

Local perturbation data publication

Context

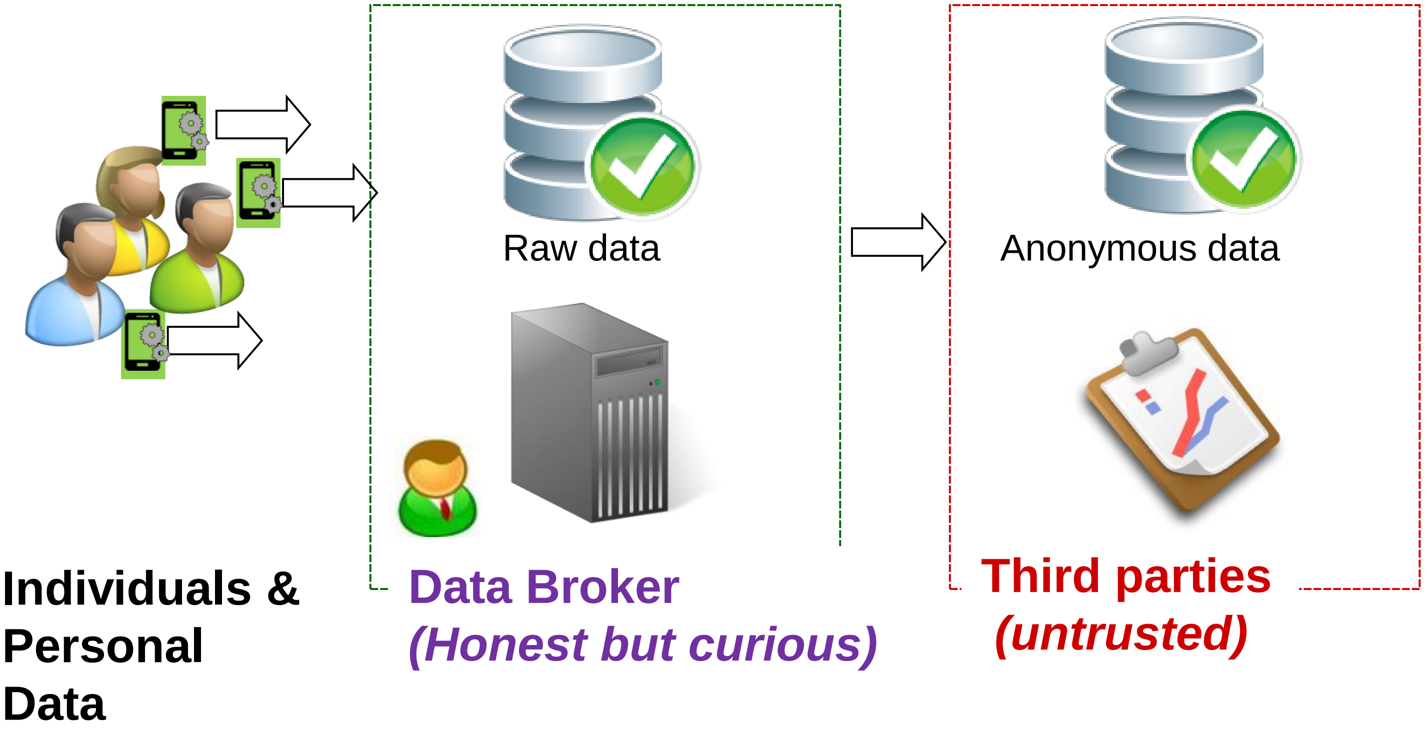
Personal data produced by sensors, forms, mobile phones, etc

Objective

Execute data intensive queries (agregates, AI, …)

Constraints

- Impossible to use an interactive query response system
- Publish the resulting *sanitized* dataset
- Choose an anonymization mechanism
- **Run in local perturbation mode**



Anonymization process components

(1) A privacy definition and metric answering the question :
What protection to propose and how to measure this protection ?

(2) A utility definition and metric answering the question :
How to measure utility loss due to using anonymous data and not real data ?

(3) An anonymization algorithm answering the question :
How to protect the data (1) while maximizing its utility (2) ?

(4) An anonymization process answering the question :
How to execute the algorithm (3) in a safe and secure manner ?

3- *k*-anonymity : a historical anonymization technique

A first non trivial definition of anonymization

External attacks on pseudonymization

Sweeney 2002, *k*-anonymity: a model for protecting privacy (IJUFK-BS)

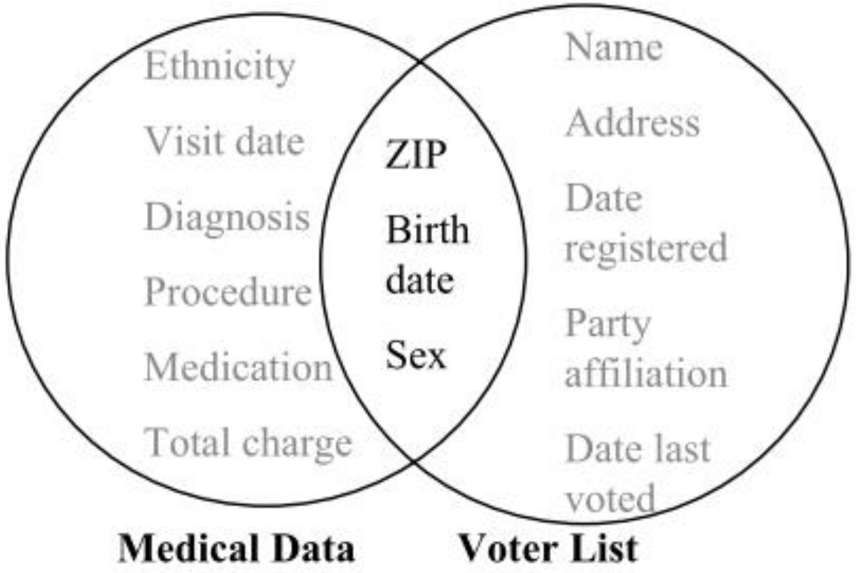
Sweeney showed the existence of quasi identifiers (QIDs):

1- Medical data was « anonymized » and published (sold) by a hospital in Massachussets

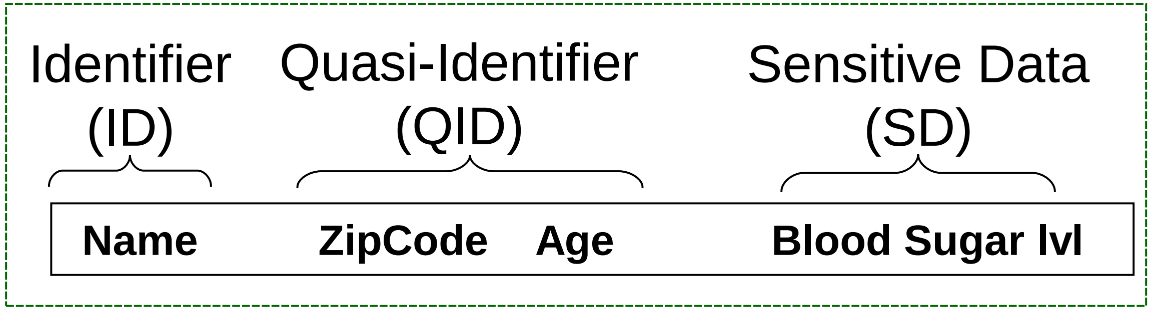
2- A nominative list of voters of Massachussets was publicly available

→Identification of Gov. Weld was possible by performing a simple *join* on both datasets using the following QID, the triplet (ZIP, Birthdate, Sex)

US 1990 census : « 87% of the population in the US had **characteristics that likely made them unique** based only on {5-digit Zip, gender, date of birth} »



Birth of *k*-anonymity

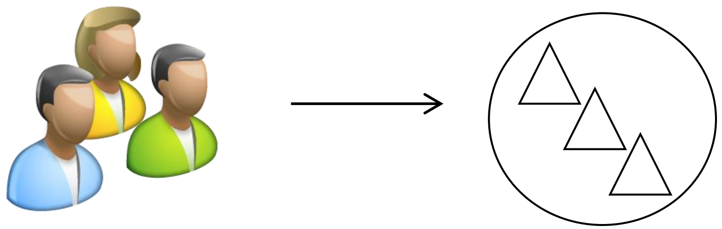


For each tuple :

– IDs must be removed

– The link between QID and SD must be *obfuscated*, but should remain as correct as possible

– This obfuscation is achieved by having each tuple correspond, via its QID to *k* SD values



k-anonymity guarantees

→ A *record linkage* probability of 1/k

I.E. the probability to find exactly which SD value is linked to a given tuple.

k-anonymity algorithms : Bucketization [Xiao, Tao]

Idea : build (random) groups of *k* tuples

Name	*Zip*	Age	BSL
Sue	18000	22	50
Pat	69000	27	70
Bob	18500	21	90
Bill	18510	20	60
Dan	69100	26	70
Sam	69300	28	75

Raw data

k-anonymity algorithms : Bucketization [Xiao, Tao]

Idea : build (random) groups of *k* tuples

Name	*Zip*	Age	BSL
Sue	18000	22	50
Pat	69000	27	70
Bob	18500	21	90
Bill	18510	20	60
Dan	69100	26	70
Sam	69300	28	75

Raw data

k-anonymity algorithms : Bucketization [Xiao, Tao]

Idea : build (random) groups of *k* tuples then divide this information into two tables : QID and SD.

Name	*Zip*	Age	BSL
Sue	18000	22	50
Pat	69000	27	70
Bob	18500	21	90
Bill	18510	20	60
Dan	69100	26	70
Sam	69300	28	75

Raw data

Zip	Age	BSL
18000	22	*G1*
18500	21	*G1*
69000	27	*G1*
18510	20	*G2*
69100	26	*G2*
69300	28	*G2*

QID Table

Group	BSL
G1	50, 70, 90
G2	60, 70, 75

SD Table

3-anonymous Bucketized data

k-anonymity algorithms : Bucketization [Xiao, Tao]

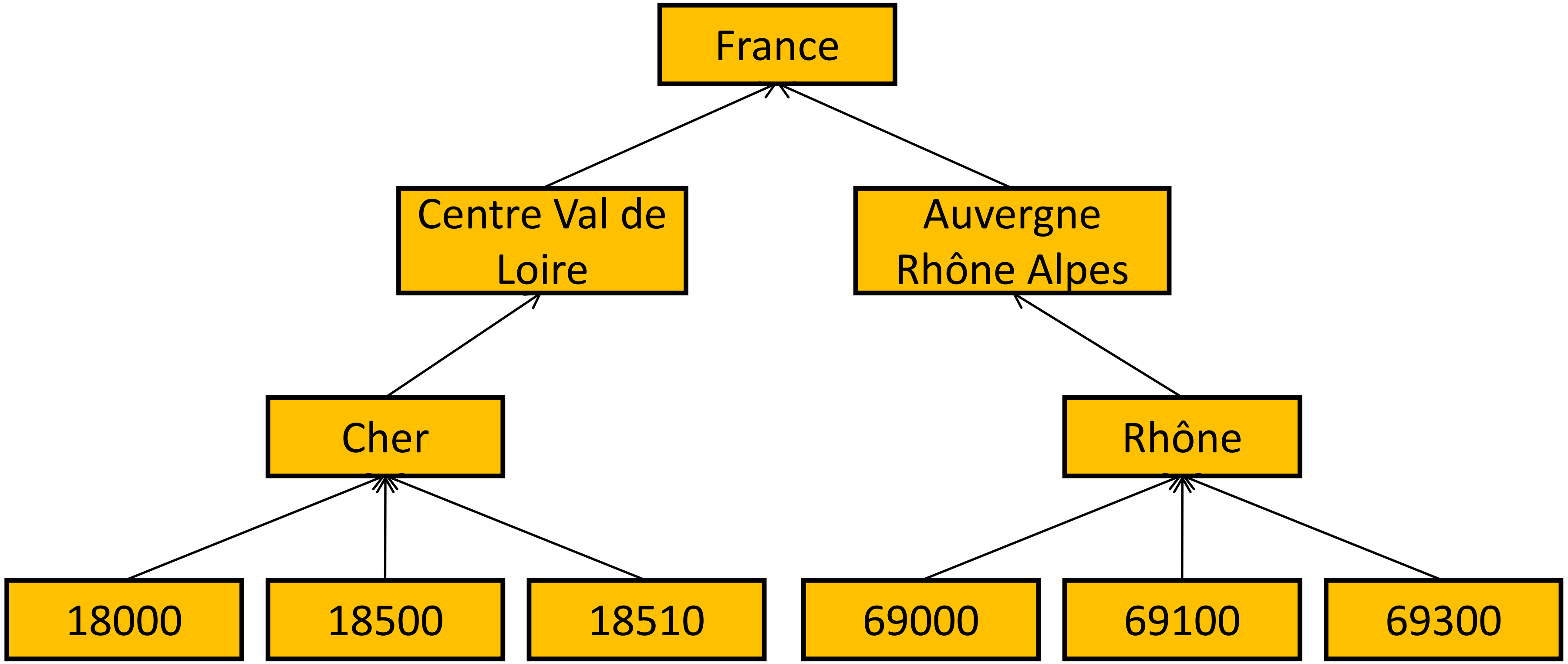
- **Good points :** easy to implement
- **Bad points :** utility of the data may not be preserved

Could we not group data together better ?

k-anonymity algorithms : Generalization [Sweeney]

Idea :

1. Define a hierarchy for each attribute of the QID



k-anonymity algorithms : Generalization [Sweeney]

Idea :

1. Define a hierarchy for each attribute of the QID

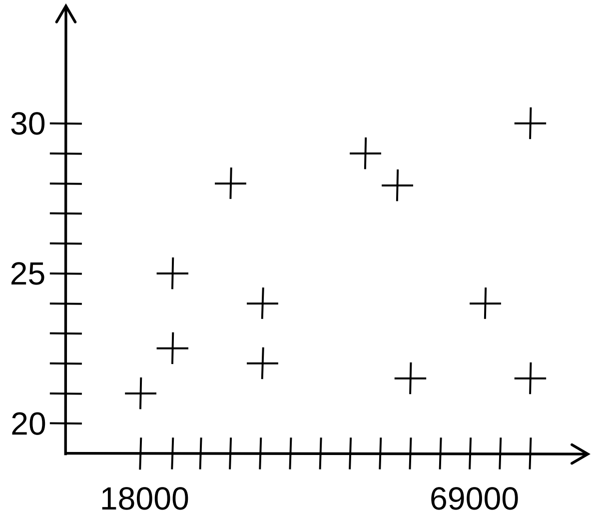
2.Generalize the value of some attributed until each tuple has the same generalized QID as *k-1* others

Name	***Zip***	**Age**	**BSL**
Sue	18000	22	50
Pat	69000	27	70
Bob	18500	21	90
Bill	18510	20	60
Dan	69100	26	70
Sam	69300	28	75

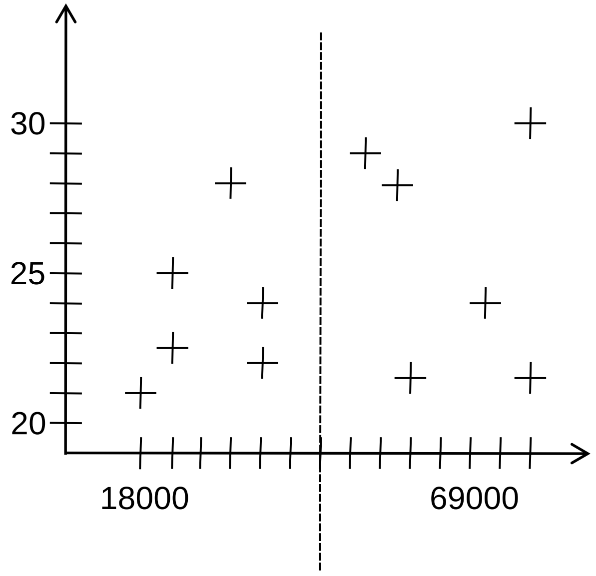
Raw Data

Optimal implementation : The *Mondrian* algorithm

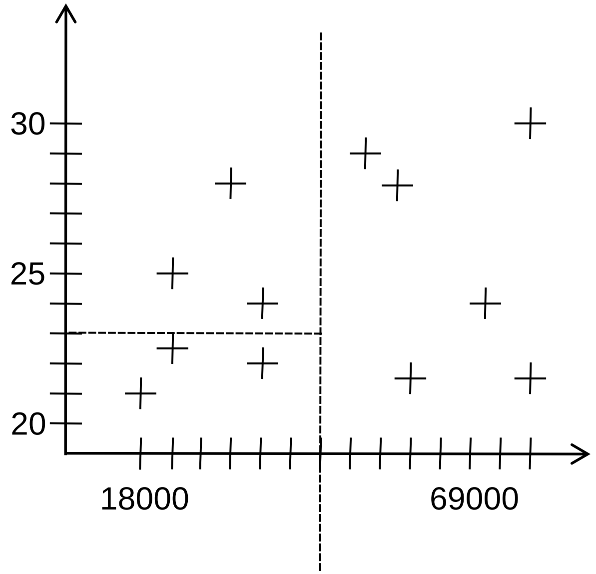
[LeFevre *et al.*]



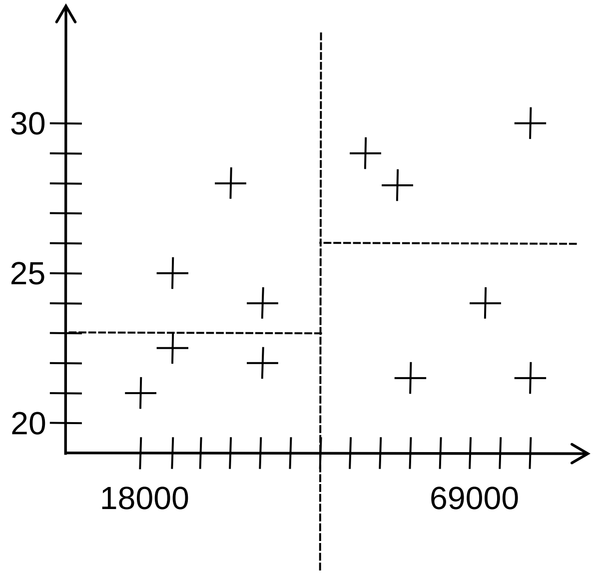
Optimal implementation : The *Mondrian* algorithm [LeFevre *et al.*]



Optimal implementation : The *Mondrian* algorithm [LeFevre *et al.*]

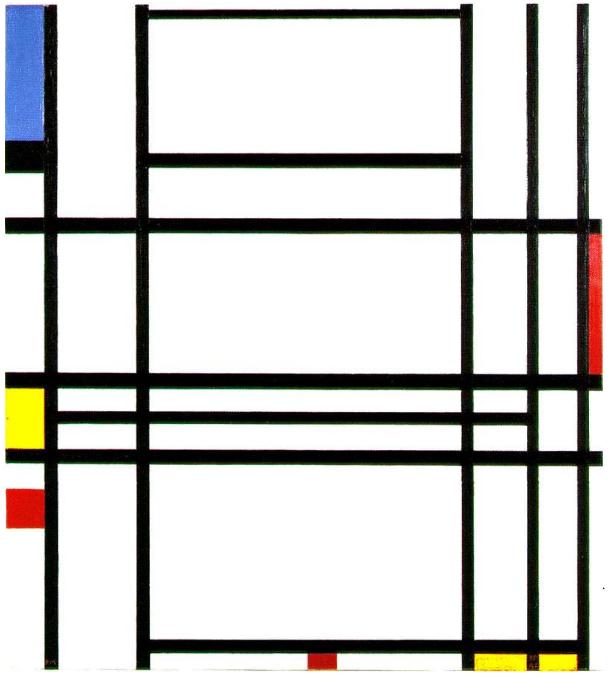


Optimal implementation : The *Mondrian* algorithm [LeFevre *et al.*]



Optimal implementation : The *Mondrian* algorithm

[LeFevre *et al.*]



Composition nr 10
Piet Mondrian

k-anonymity algorithms : Generalization [Sweeney]

This technique allows running SQL aggregate queries :

```
SELECT Zip, AVG(BSL)
FROM T
GROUP BY Zip
```

k-anonymity algorithms : Generalization [Sweeney]

This technique allows running SQL aggregate queries :

SELECT Zip, AVG(BSL)

FROM T

GROUP BY Zip

Zip	BSL
18000	50
69000	70
18500	90
18510	60
69100	70
69300	75

Raw data

k-anonymity algorithms : Generalization [Sweeney]

This technique allows running SQL aggregate queries :

```
SELECT Zip, AVG(BSL)
FROM T
GROUP BY Zip
```

Zip	**BSL**
Cher	66.67
Rhône	71.67

Raw Data

Privacy / Utility tradeoff !

/!\ How to measure utility ? /!\

4- Re-identification risk evaluation

Attacks based on QID characteristics

QID based attacks :

Who is the adversary and what does she know ?

- **Objective :**

Define metrics to evaluate the impact of sets of attributes on reidentification, depending on a given *attack model :*

  - Prosecutor Risk
  - Journalist Risk
  - Marketeer Risk

See [El Emam & Dankar 08] : https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2528029/

- Attacks based on a prior analysis of the uniqueness of the population

Reindentification risk metrics :
Prosecutor risk [El Emam & Dankar 08]

- *Prosecutor risk :*
- *Re-identify a specific individual (known as the prosecutor re-identification scenario). The intruder (e.g., a prosecutor) would know that a particular individual (e.g., a defendant) exists in an anonymized database and wishes to find out which record belongs to that individual.*

Former Governer
Of Massachussetts
Bill Weld

Reindentification risk metrics :

Journalist risk [El Emam & Dankar 08]

- *Journalist risk :*
- *Re-identify an arbitrary individual (known as the journalist re-identification scenario). The intruder does not care which individual is being re-identified, but is only interested in being able to claim that it can be done. In this case the intruder wishes to re-identify a single individual to discredit the organization disclosing the data.*

Thelma Arnold
Aka user #4417749
AoL Search

Reindentification risk metrics :

Marketeer risk [El Emam & Dankar 10]

- *Marketeer risk :*
- *An intruder wishes to re-identify as many records as possible in the disclosed database. We assume that the intruder lacks any additional information apart from the matching quasi-identifiers.*

Cambridge Analytica

Risk Model

- Private Database : U with |U|=n
- Attacked background knowledge (known database) : D with |D|=N
- X the set of all possible equivalence classes
- Z = {z_i} an equivalence class
- J the number of all possible equivalence classes, ~J the number of real equivalence classes
- f_j the number of records of equivalence class j in U
- F_j the number of records of equivalence class j in D

Theorem 1. The expected proportion of U records that can be disclosed in a random mapping from U to D is.

$$\lambda = \sum_{j=1}^{\tilde{J}} \frac{f_j / F_j}{n} \qquad \ldots\ldots\ldots\ldots\ldots(1)$$

Note that if $n = N$ then $\lambda = \frac{\tilde{J}}{N}$.

Source : El Emam & Dankar

$$R_p = 1 \Big/ \min_j \left(f_j \right)$$

Prosecutor Risk

$$R_J = 1 \Big/ \min_j \left(F_j \right)$$

Journalist Risk

Risk Model

- Private Database : U with |U|=n
- Attacked background knowledge (known QID database) : D with |D|=N
- X the set of all possible equivalence classes
- Z = $\{z_i\}$ an equivalence class
- J the number of all possible equivalence classes, ~J the number of real equivalence classes
- f_j the number of records of equivalence class j in U
- F_j the number of records of equivalence class j in D

Theorem 1. The expected proportion of U records that can be disclosed in a random mapping from U to D is.

$$\lambda = \sum_{j=1}^{\tilde{J}} \frac{f_j / F_j}{n} \qquad \text{...............(1)}$$

Note that if $n = N$ then $\lambda = \frac{\tilde{J}}{N}$.

Source : El Emam & Dankar

$$R_p = 1 \Big/ \min_j \left(f_j \right)$$

Maximal Prosecutor Risk

If N=n

R_p=R_J

$$R_J = 1 \Big/ \min_j \left(F_j \right)$$

Maximal Journalist Risk

Risk Model

- Prosecutor and journalist risk can then be averaged over the whole dataset U (or D)

5- Aggregation based anonymization techniques

Many models : L-diversity, T-closeness

Main weakness of *k*-anonymity

What if all sensitive values are the same ?

Name	Zip	Age	BSL
Sue	18000	22	50
Pat	69000	27	70
Bob	18500	21	90
Bill	18510	20	60
Dan	69100	26	70
Sam	69300	28	70

Raw data

Zip	Age	BSL
Cher	[20-24]	50
Rhône	[25-29]	70
Cher	[20-24]	90
Cher	[20-24]	60
Rhône	[25-29]	70
Rhône	[25-29]	70

Anonymous data

→ BSL of all inhabitants of Rhône district is 70 !

L-diversity

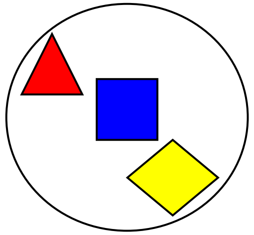
[Machanavajjhala *et al. 06*]

Name	**Zip**	**Age**	**BSL**
Sue	18000	22	50
Pat	69000	27	70
Bob	18500	21	90
Bill	18510	20	60
Dan	69100	26	70
Sam	69300	28	70

Raw data

Zip	**Age**	**BSL**
France	[20-29]	50
France	[20-29]	70
France	[20-29]	90
France	[20-29]	60
France	[20-29]	70
France	[20-29]	70

Anonymous and diverse data



Computed by putting constraints on sensitive data

l-diversity guarantees

- An indivicual whose QID belongs to a class, and who took part in the release can be associated to any of the L values with a given probability
- E.g., Bob can be associated with any value of {Flu, HIV, Cancer} with the same probability

- → Attribute linkage probability = 1/L

Bob

Zip	**Age**	**C.E.**
France	[20, 29]	50
France	[20, 29]	60
France	[20, 29]	90
…		

Intuition

- Each *k*-anonymous group must also be *diverse* enough
- Each equivalence class must be associated to at least L « well represented » sensitive values.
- “Well represented” is a loose definition
- **Consequences :**

– précision loss ☹

– anonymity gain ☺

Intuition

- Each *k*-anonymous group must also be *diverse* enough
- Each equivalence class must be associated to at least L « well represented » sensitive values.
- “Well represented” is a loose definition
- **Consequences :**
  – précision loss ☹
  – anonymity gain ☺

ILLUSTRATION USING ARX

There are many more aggregation models …

- T-closeness
- δ-disclosure
- …

6- Statistical Methods

Local perturbation method, and local differential privacy model

The Randomized Response approach

A.K.A.« Local Differential Privacy »

Context : Yes/No answer

- Set a probablity *p* to tell the truth and (*1-p*) to lie (same *p* for each individual)
- *In general : $p=0.5 + \varepsilon_{RR}$*
- Estimator:
  - Let π represent the propotion of the population for which the true answer is « Yes »
  - The expected propotion of « Yes » is :
    $P(Yes) = (\pi * p) + (1 - \pi)*(1 - p)$
    → $\pi = [P(Yes) - (1 - p)] / (2p - 1)$
  - If m/n individuals have answered « Yes » then, π_{est} is an estimate for π :
    $\pi_{est} = [m/n - (1 - p)] / (2p - 1)$

Local Differential Privacy
Jordan & Wainwright [2013]

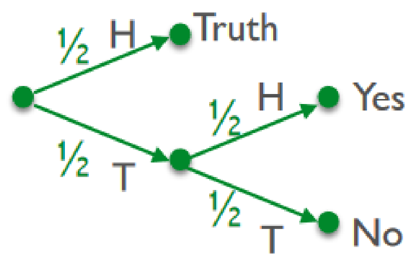
Definition Let $\mathcal{X}$ be a set of possible values and $\mathcal{Y}$ the set of noisy values. A mechanism $\mathcal{K}$ is ε-locally differentially private (ε-LDP) if for all $x_1, x_2 \in \mathcal{X}$ and for all $y \in \mathcal{Y}$

$$P[\mathcal{K}(x) = y] \leq e^{\varepsilon}\ P[\mathcal{K}(x') = y]$$

or equivalently, using the conditional probability notation:

$$p(y \mid x) \leq e^{\varepsilon}\ p(y \mid x')$$

For instance, the Randomized Response protocol is (log 3)-LPD



Shamelessly taken from C. Palamidessi

y

x	yes	no
yes	¾	¼
no	¼	¾

« ***Randomized Response*** » **algorithm with ε_{RR} = 0.25**

Post-Randomization Matrix (PRAM)
Another approach for local differential privacy

- Used in Mu-Argus (Eurostat) software

- Define a probability matrix for each value to transition towards another one, then apply these probabilities

	Flu	**Covid19**
Flu	0.75	0.25
Covid19	0.1	0.9

7- Differential Privacy

Formal guarantees

Differential privacy

Dwork 2006, *Differential Privacy* (ICALP)

C. Dwork

- The main problem of *k*-anonymity is that its security depends on the background knowledge of the attacker
- A *framework* was proposed in 2006 by Dwork. It aims to quantify the fact that an attacker can know whether a specific person participated in a data release or not.
- We say that a (randomized) algorithm A satisfies (ε,δ)-differential privacy if

- For each ***adjacent*** database pair D_1 and D_2 (i.e. which differ by at most one invididual

- For any output Ω of A, There exists ε such that :

$$\boxed{Pr[A(D_1) = \Omega] \leq e^{\varepsilon} Pr[A(D_2)=\Omega]+\delta}$$

Laplace mechanism

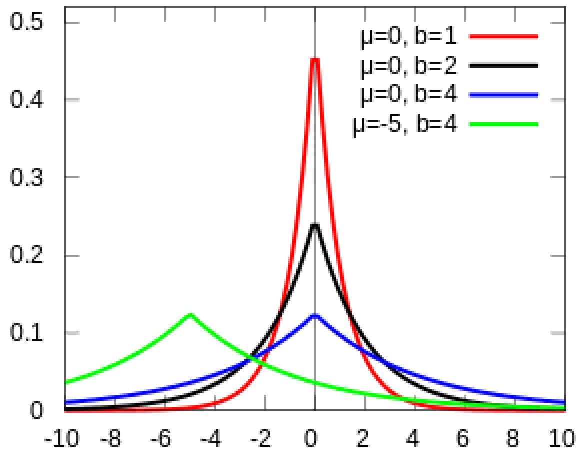
Dwork introduced the *mechanism of Laplace* which shows that if one adds random noise to a function, drawn from the Laplace distribution centered on 0 and of scale $\Delta f/\varepsilon$ then this mechanism is ε-differentially private

Definition 4 (ℓ_1-sensitivity). *The ℓ_1-sensitivity of a function $f : \mathbb{N}^{|\chi|} \rightarrow \mathbb{R}^k$ is :*

$$\Delta f = \max_{\substack{x,y \in \mathbb{N}^{|\chi|} \\ ||x-y||_1 = 1}} ||f(x) - f(y)||_1$$

Definition 7 (Laplace Distribution). *The Laplace Distribution with scale b is the distribution with probability density function*

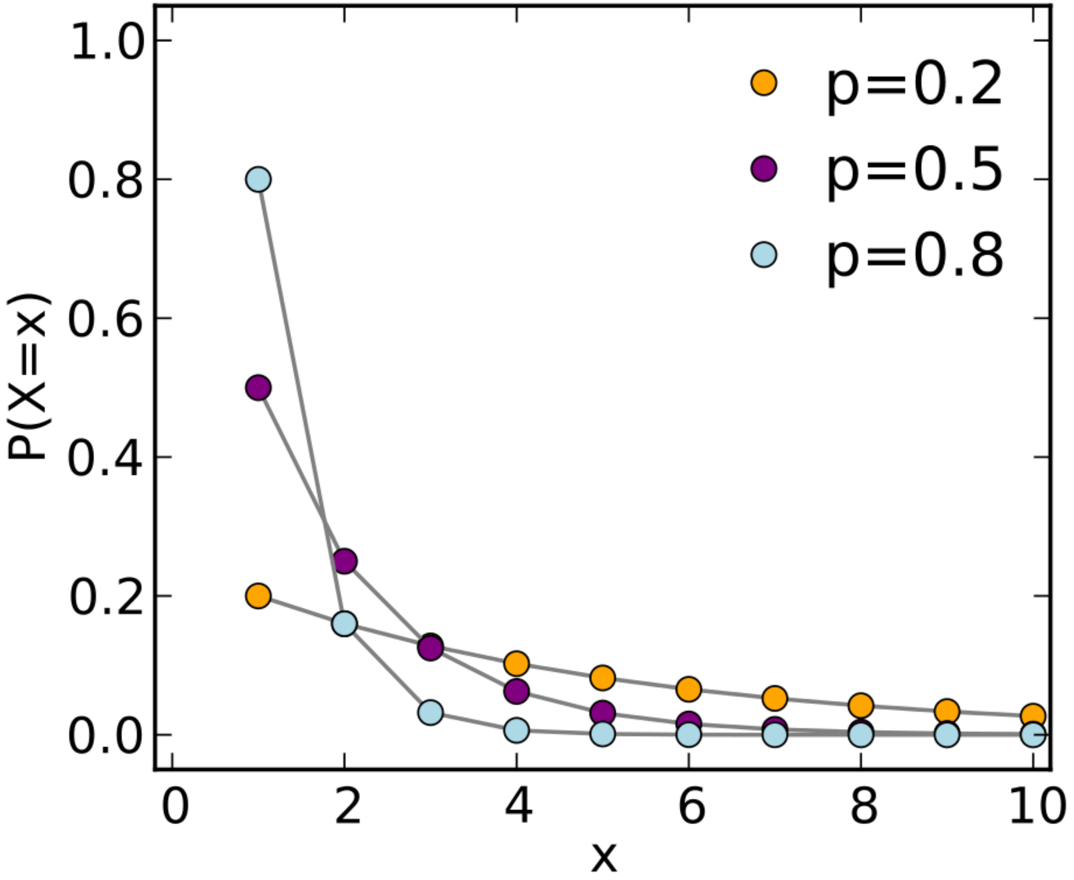
$$Lap(x|b) = \frac{1}{2b} \exp\left(-\frac{|x|}{b}\right)$$



Geometric mechanism

Ghosh *et al.* adapted this mechanism to integers, using a mechanism called the *geometric mechanism*.

This approach can also be used on finite intervals (aka Trunkated Laplace).



Other mechanisms …

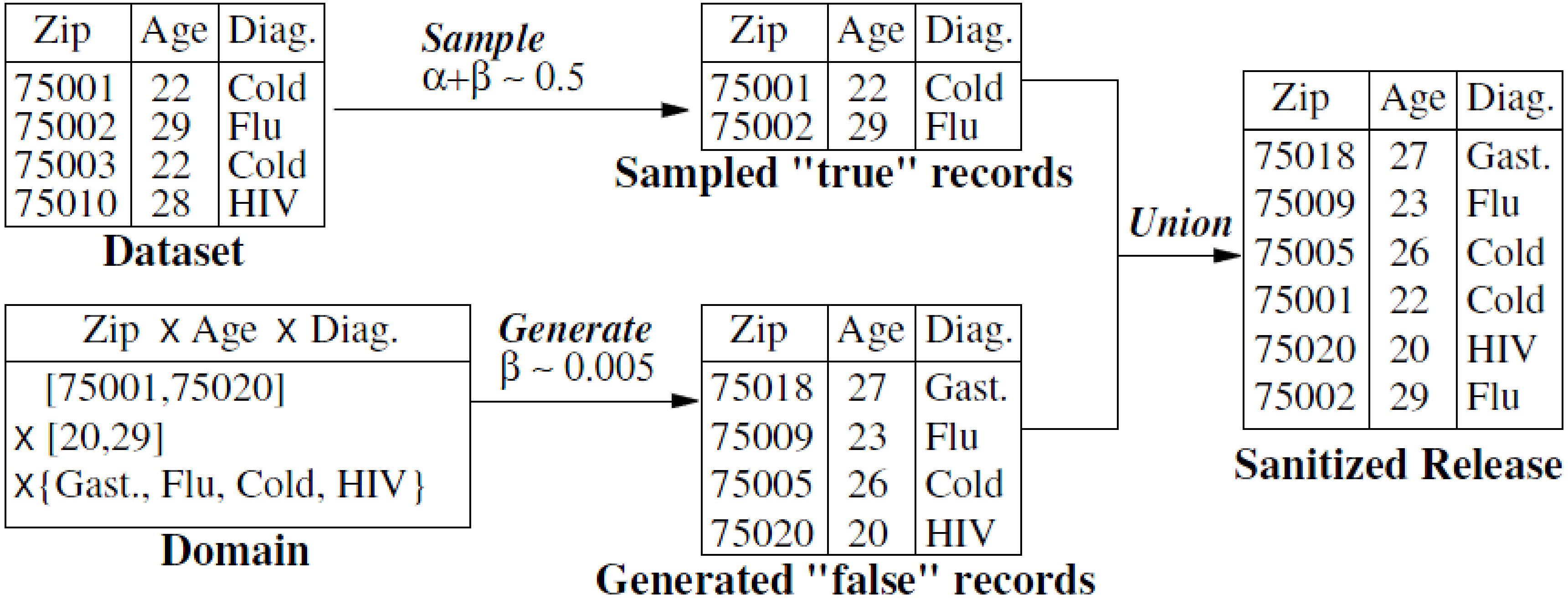
- The exponential mechanism (introduced by Dwork) is able to manage the generation of categorical data (e.g. eye colour).
- The composition theorem explains how to compose differentially private mechanisms, and introduces the privacy budget.

Hands on Differential Privacy

- Algorithms are quite easy to develop (random sampling in known distributions)
- Libraries are available such as the IBM privacy library (in Python)

https://github.com/IBM/differential-privacy-library

An example : The α, β – algorithm [Rastogi *et al.*]



We can compute COUNT agregations using a statistical estimator :

$$Q_{Cold} = (\underbrace{n_{sanitized}}_{=2} - \underbrace{\beta . n_{Domain}}_{=200*0.005=1}) / \underbrace{\alpha}_{=0.5}$$

Linking *k*-anonymat et DP
J.Domingo-Ferrer [2015]

In later years, work has been put into trying to link both approaches

Domingo-Ferrer *et al.* have shown it is possible to achieve *t*-closeness while respecting DP guarantees

The SafePub Algorithm (ARX)

Bild, Kuhn, Passer [2018] : avoiding the optimality attack

Input: Dataset D, Parameters ϵ_{anon}, ϵ_{search}, δ, *steps*
Output: Dataset S

1: Draw a random sample D_s from D $\triangleright$ (ϵ_{anon})
2: Initialize set of transformations G
3: **for** (Int $i \leftarrow 1, ..., steps$) **do**
4: Update G
5: **for** ($g \in G$) **do**
6: Anonymize D_s using g $\triangleright$ (ϵ_{anon}, δ)
7: Assess quality of resulting data
8: **end for**
9: Probabilistically select solution $g \in G$ $\triangleright$ (ϵ_{search})
10: **end for**
11: **return** Dataset D_s anonymized using the best solution selected in Line 9 $\triangleright$ (ϵ_{anon}, δ)

Fig. 4. High-level design of the SafePub mechanism. The search strategy is implemented by the loop in lines 3 to 10.

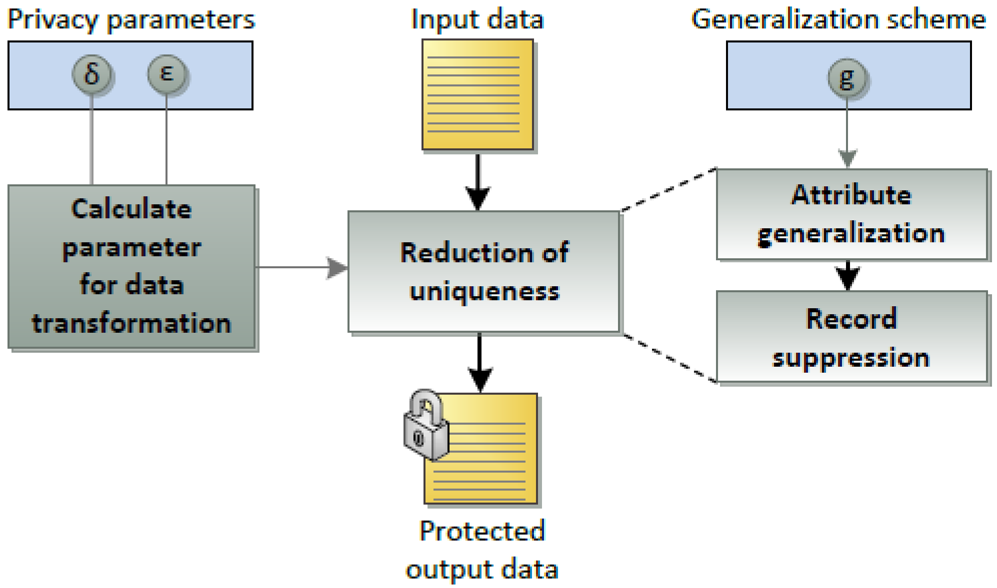


Fig. 5. Overview of the anonymization operator.

Idea : Random choice of *k-anonymity* parameters

8- Conclusion

Anonymization is a tradeoff between security and utility

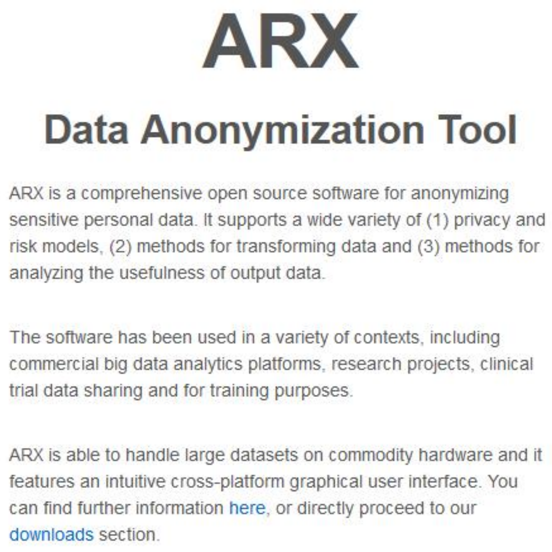
- When thinking about using anonymous data, it is essential to be able to
  - Evaluate the risk
  - Evaluate the utility of anonymized data (not discussed here)
- What model to use ?
  - *Differential Privacy* has been the go-to model in the computer science community for over 10 years
  - Aggregation techniques (*k*-anon et al.) are still very much used as a *pragmatic* solution for risk reduction (just as pseudonymization !)
- It is not possible to give absolute guarantees !
  - The GDPR requires obligation of means
  - Efficiency should be evaluated experimentally (i.e. GDA Score, DARC competition, …)

If data is not anonymous, then GDPR applies

9- Hands on basic anonymization using ARX

ARX Data Anonymisation Tool

- Available in opensource at https://arx.deidentifier.org/anonymization-tool/
- Research project of TU München


ARX

Data Anonymization Tool

ARX is a comprehensive open source software for anonymizing sensitive personal data. It supports a wide variety of (1) privacy and risk models, (2) methods for transforming data and (3) methods for analyzing the usefulness of output data.

The software has been used in a variety of contexts, including commercial big data analytics platforms, research projects, clinical trial data sharing and for training purposes.

ARX is able to handle large datasets on commodity hardware and it features an intuitive cross-platform graphical user interface. You can find further information here, or directly proceed to our downloads section.