What is fair data processing ?

Pr. Benjamin NGUYEN
INSA Centre Val de Loire
benjamin.nguyen@insa-cvl.fr

ISI2017
MARRAKECH
61st ISI WORLD STATISTICS CONGRESS

Two ways to process personal data in France (Europe)

General Data Protection Regulation -- Regulation (EU) 2016/679

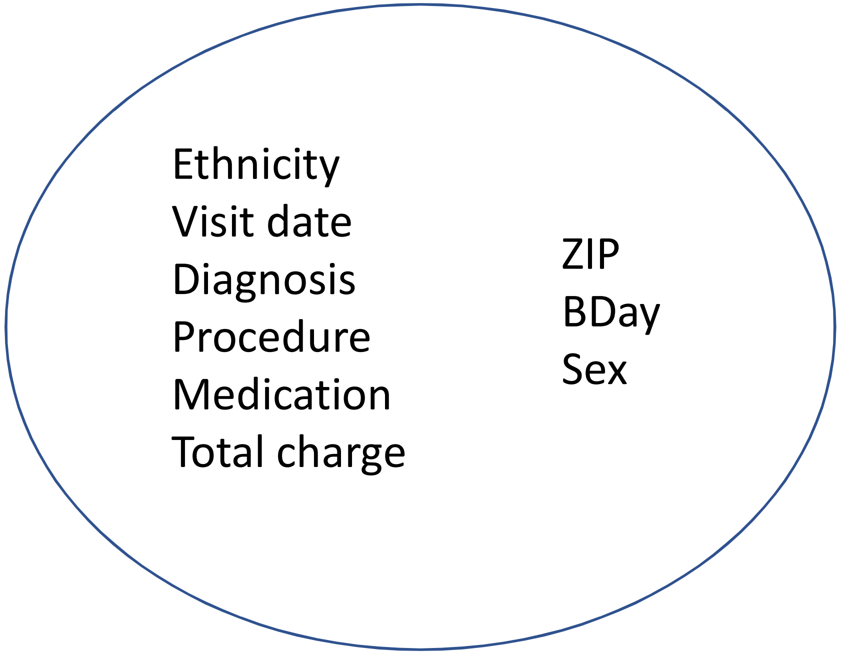
- Get the approval of the CNIL (French National Data Protection Authority)
  - Give all details of data processing : objective / intent, retention period, consent, data collected, right to explenation, right to be forgotten, etc.
  - Get process approved

- Process anonymous data, because anonymous data is no longer personal data
  - Pro : Do what you want with the data
  - Con : How to anonymize ?

What is anonymous data ?

“The principles of data protection should apply to any information concerning an identified or identifiable natural person. 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. 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. 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. 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.”

The problem with pseudonymous data [Swe02]

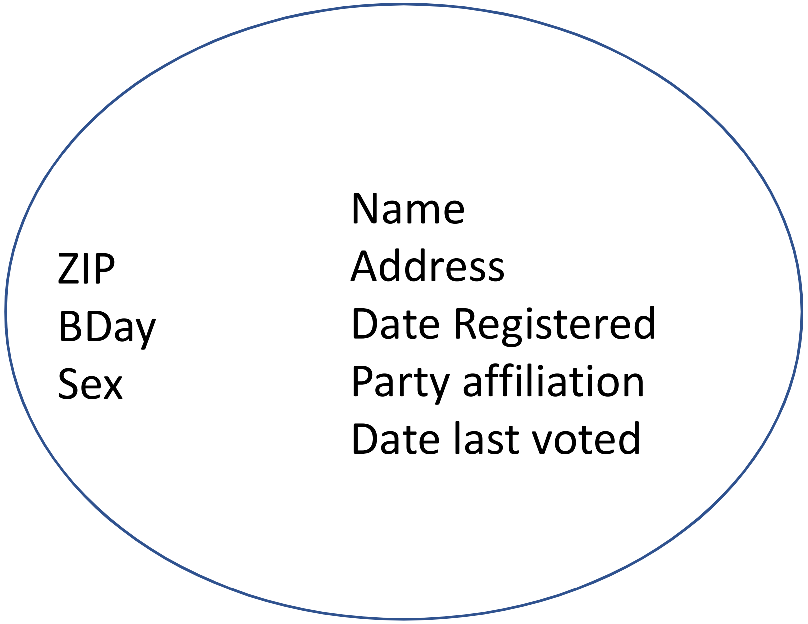


Ethnicity
Visit date
Diagnosis
Procedure
Medication
Total charge

ZIP
BDay
Sex

Medical DB

The problem with pseudonymous data [Swe02]

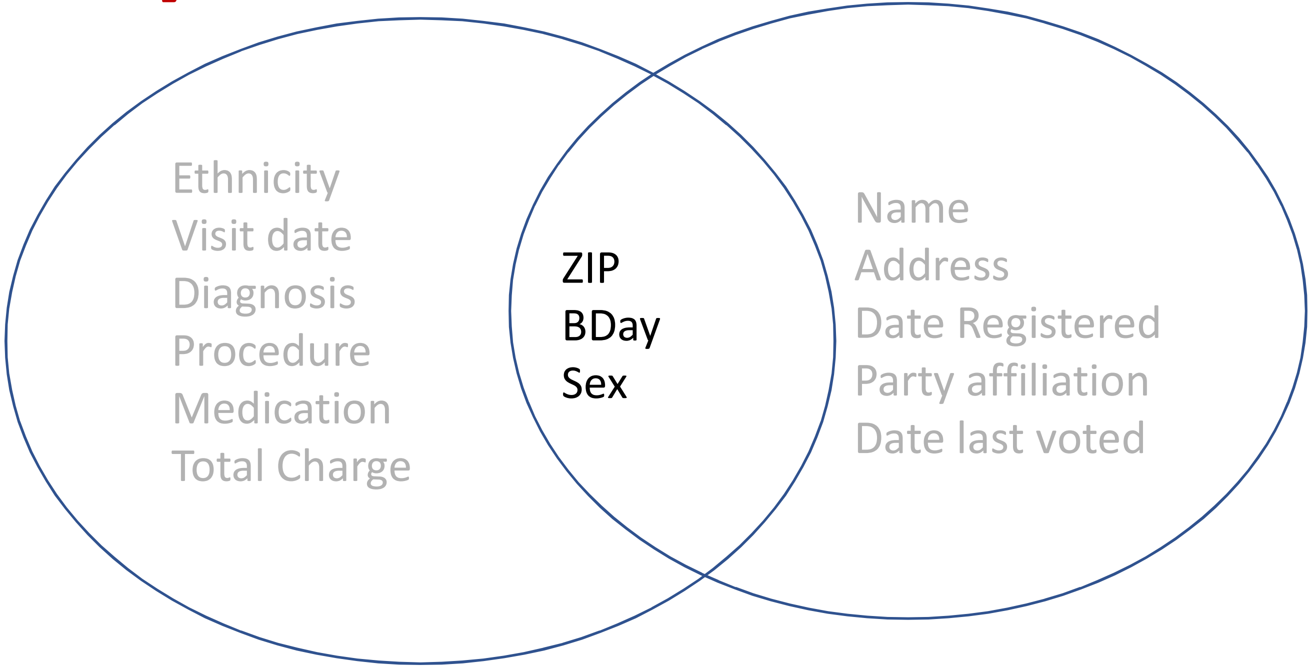


ZIP
BDay
Sex

Name
Address
Date Registered
Party affiliation
Date last voted

Voter List

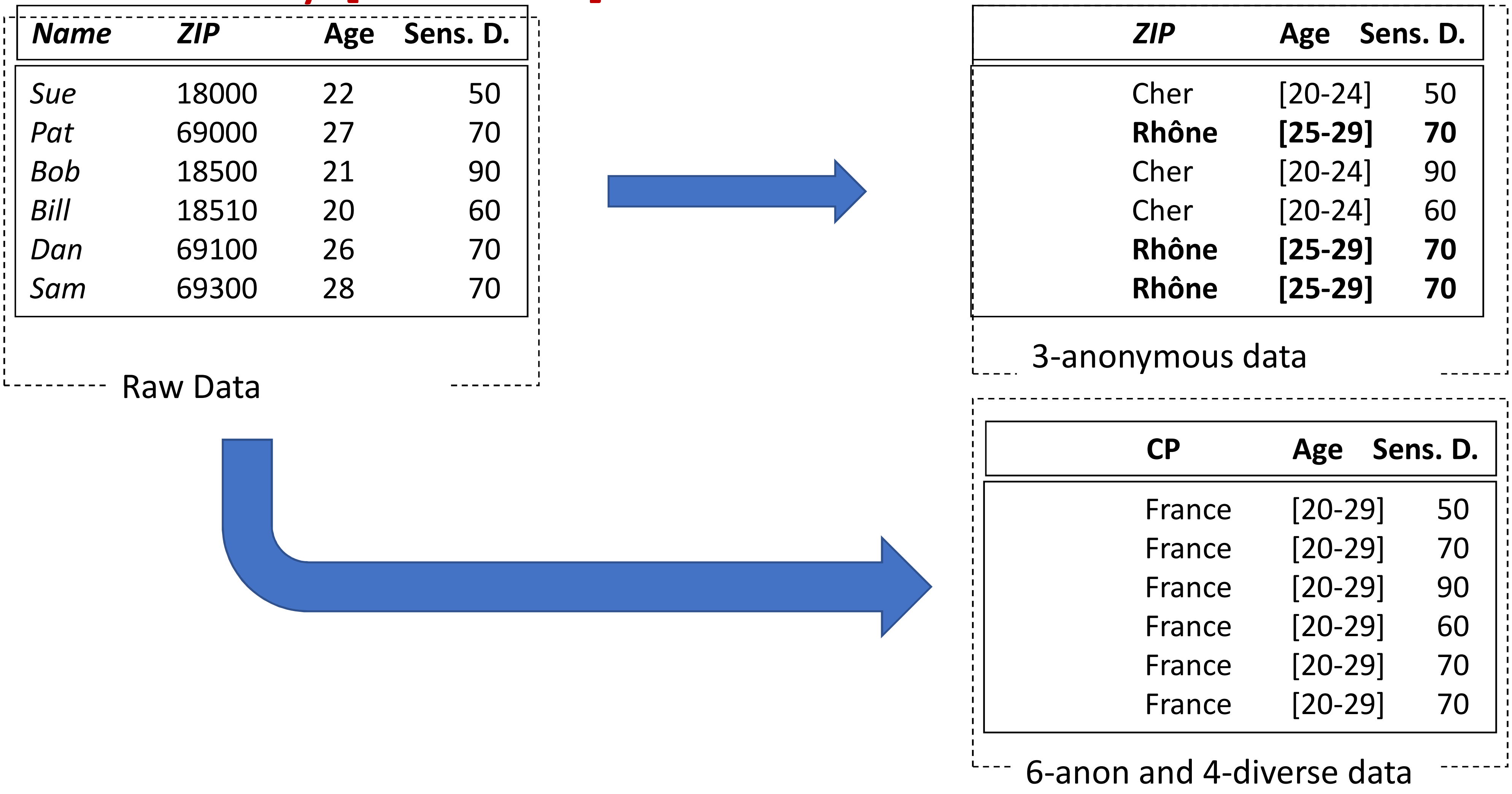
The problem with pseudonymous data [Swe02]



Medical DB

Voter List

Some proposals : K-anonymity [Swe02] and l-diversity [MKG+06]



Name	*ZIP*	Age	Sens. D.
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	Sens. D.
Cher	[20-24]	50
Rhône	**[25-29]**	**70**
Cher	[20-24]	90
Cher	[20-24]	60
Rhône	**[25-29]**	**70**
Rhône	**[25-29]**	**70**

3-anonymous data

CP	Age	Sens. D.
France	[20-29]	50
France	[20-29]	70
France	[20-29]	90
France	[20-29]	60
France	[20-29]	70
France	[20-29]	70

6-anon and 4-diverse data

Some proposals : Differential Privacy [Dwo06]

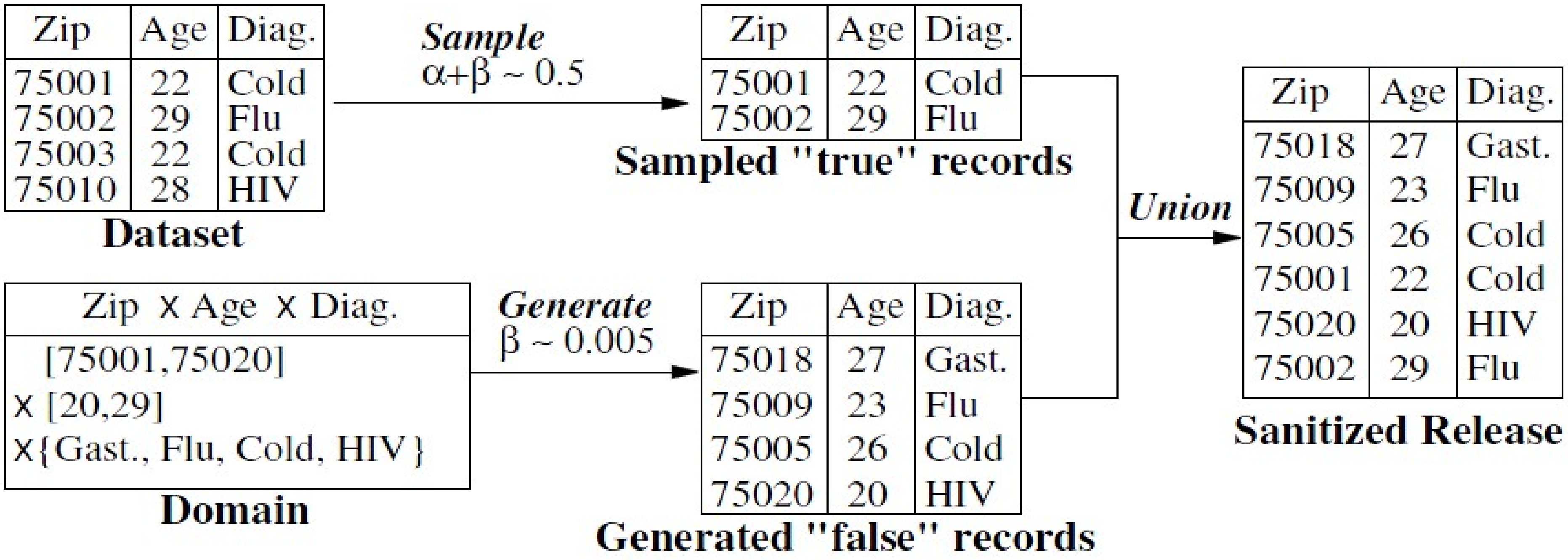
Differential privacy is a *characteristic* of an algorithm, which tries to assess its security. One says a (random) anonymization algorithm satisfies ε-differential privacy if

- For all pairs of tables D_1 et D_2 which vary by only 1 tuple
- For any result Ω of this algorithm

There exists ε such that :

$$\mathbf{Pr[A(D_1) = \Omega] \leq e^{\varepsilon}\, Pr[A(D_2)=\Omega]}$$

α, β – algorithm [Rastogi *et al.*]



Zip	Age	Diag.
75001	22	Cold
75002	29	Flu
75003	22	Cold
75010	28	HIV

Dataset

Zip X Age X Diag.
[75001,75020]
X [20,29]
X{Gast., Flu, Cold, HIV}

Domain

Sample
α+β ~ 0.5

Zip	Age	Diag.
75001	22	Cold
75002	29	Flu

Sampled "true" records

Generate
β ~ 0.005

Zip	Age	Diag.
75018	27	Gast.
75009	23	Flu
75005	26	Cold
75020	20	HIV

Generated "false" records

Union

Zip	Age	Diag.
75018	27	Gast.
75009	23	Flu
75005	26	Cold
75001	22	Cold
75020	20	HIV
75002	29	Flu

Sanitized Release

We can compute aggregate values such as COUNTs based on the estimator :

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

Homomorphic Encryption

Homomorphic Encryption is a characteristic of several crypto-systems such as RSA, Paillier, ElGamal, etc.

Example : Consider RSA. Given the RSA public key (e, m), the encryption of a message x is given by :

$$\mathbf{E(p)=p\text{^}e \bmod m}$$

The homomorphic property is :

$$\mathbf{E(p_1) \times E(p_2) = p_1\text{^}e \times p_2\text{^}e \bmod m = (p_1 \times p_2)\text{^}e \bmod m = E(p_1 \times p_2)}$$

Fully Homomorphic Encrytion means that *all ring* operators are homomorphic (this means + and x).

Fully Homomorphic Encryption [Gen09]

- Any program with bounded input can be transformed into a Boolean circuit
- Any circuit can be transformed into a polynomial modulo 2
- Secure computation of a polynomial equates to securely computing any program
- To securely compute a polynomial, it is necessary and sufficient to securely compute + and x operations.

We say that E is a fully homomorphic encryption from ({0,1}, +, x) to (D, $\oplus$, $\otimes$) if for all c_1, c_2 in D, such that $c_1=E(p_1)$ and $c_2=E(p_2)$

$$E^{-1}(c_1) \oplus E^{-1}(c_2) = p_1+p_2$$

$$E^{-1}(c_1) \otimes E^{-1}(c_2) = p_1 \times p_2$$

Or more generally $E^{-1}(f_D(c_1,\ldots,c_n))=f_{\{0,1\}}(p_1,\ldots,p_n)$

Is anonymization enough ?

- **Do these techniques work for « Big Data » ?**
  - Efficiency ? → hard but research problem
  - Specificity of human generated/related data ? → see [dMon13] « Unique in the crowd»
- **Other options to consider**
  - User control
  - Usage control
  - Auditability
  - Limited collection
  - Limited retention
  - etc