Introduction

M2 SIF - SED

Tristan Allard
Univ. Rennes 1 / Irisa lab.
tristan.allard@irisa.fr

Autumn 2023

Organisation of the class

- **Teachings :** 7 CMs (lessons), 7 TPs (practical works).
- **Teachers :** Tristan Allard (main teacher, resp. class), Gaëtan Le Guelvouit (head of the Trust & Security lab at the IRT BCOM).
- **Evaluation :** presentation (), final exam ().

Themes of the class

Privacy: focus on privacy-preserving data publishing techniques

1. Introduction
   Privacy is not dead (especially anonymization).
2. Lessons from the past: partition-based approaches
   Not completely past, not completely learnt neither...
3. Modern approaches: differential privacy
   Today's "gold standard"?
4. Attacks on privacy-preserving data publishing schemes
   From re-identification attacks to membership inference attacks

Intellectual Property: securing multimedia content delivery

1. Digital watermarking
   Some tattoo sheeps, others tattoo data.
2. Secure distribution
   "Keep an eye on me."

Privacy

Focus: privacy-preserving data publishing

Progress of the Talk

Massive Data Collection

Which dimensions about us generate data today ?

- Geolocations (GPS-enabled devices, cellphone access points, online maps trajectories, public transport cards, . . .)
- Electrical consumption (smart meters, smart plugs, . . .)
- Health and physiological status (social security, wearables)
- Citizen's rights and duties (IDs, taxes, . . .)[1]
- Social networks and messaging applications (friends and family, work, . . .)
- Online activities (search queries, browsing history, . . .)
- . . .
- Students' connected beds at the CROUS[2] ?

Easier to answer to what does **not** generate data about us today ?

[1] See the Digital ID Wallet product: `https://youtu.be/YSb0nLRte_A`.

[2] `https://www.ouest-france.fr/bretagne/rennes-35000/etudiants-rennes-dormez-vous-etes-surveilles-5227294/amp`

The New Oil

"Personal data is the new oil of the internet and the new currency of the digital world."
M. Kouneva, European Commissioner for Consumer Protection, March 2009

Targeted Ads I

For ex : on social networks.

Targeted Ads II

From [2] (CACM '21).

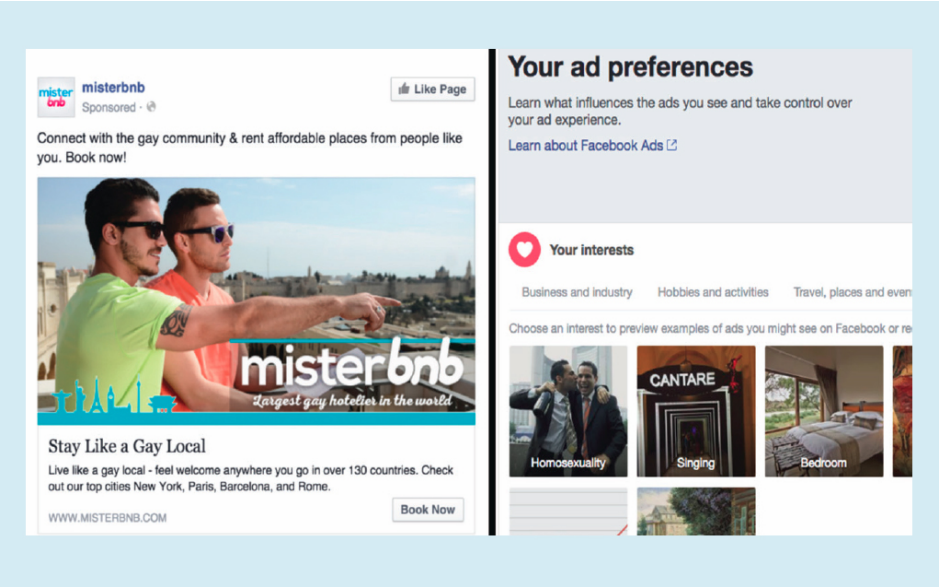


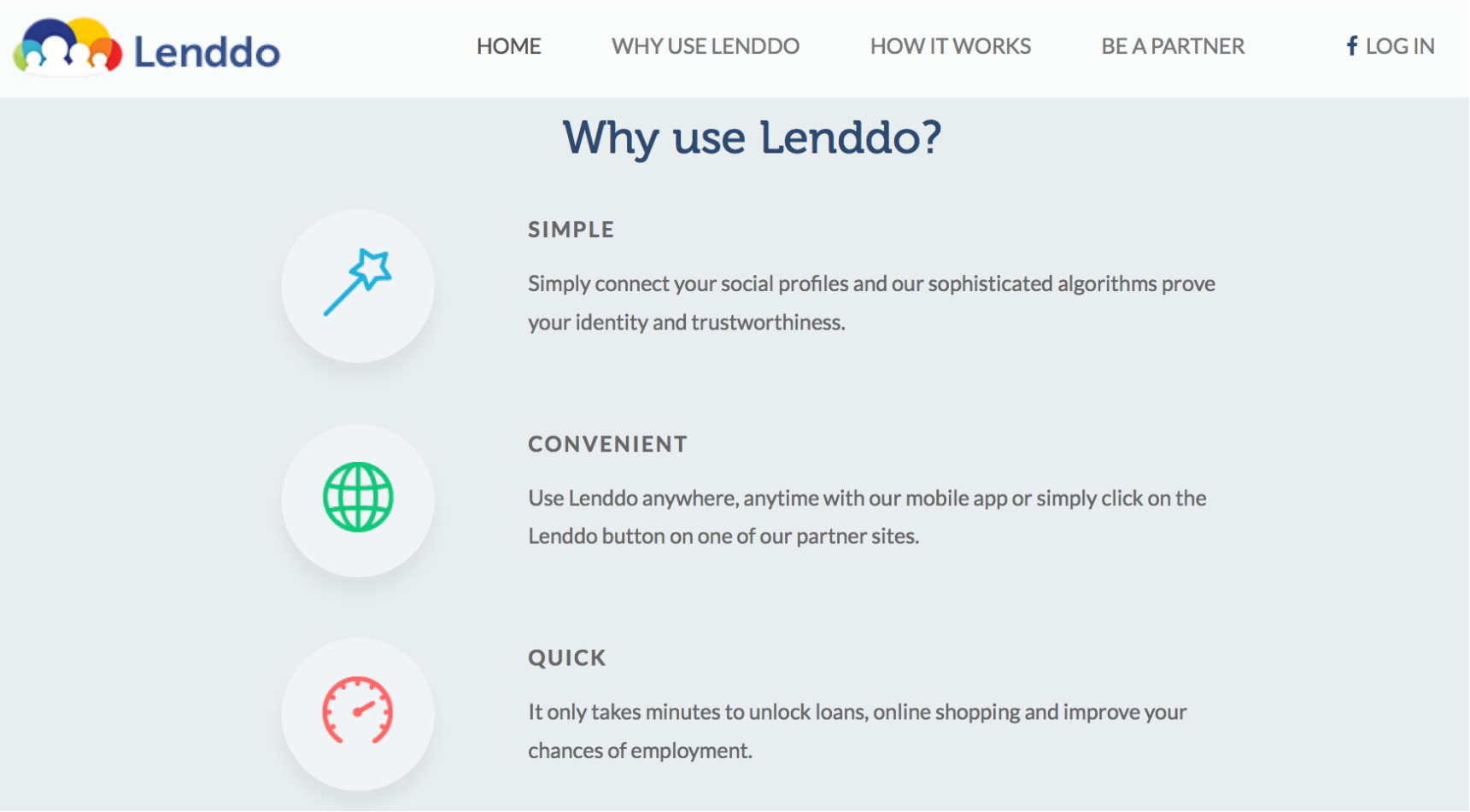
Figure: An example of a sensitive ad related to homosexuality

Code	Country	Homosexuality
AF	AFGHANISTAN	12.31
MR	MAURITANIA	0.99
QA	QATAR	2.35
SO	SOMALIA	1.44
PK	PAKISTAN	1.54
AE	UNITED ARAB EMIRATES	3.00
BN	BRUNEI	5.24
NG	NIGERIA	2.35
SA	SAUDI ARABIA	2.08
YE	YEMEN	1.08
IQ	IRAQ	3.20

Figure: Percentage of FB users (FFB) tagged with the interest "Homosexuality" in countries where being homosexual may lead to death penalty/

Risk Optimization

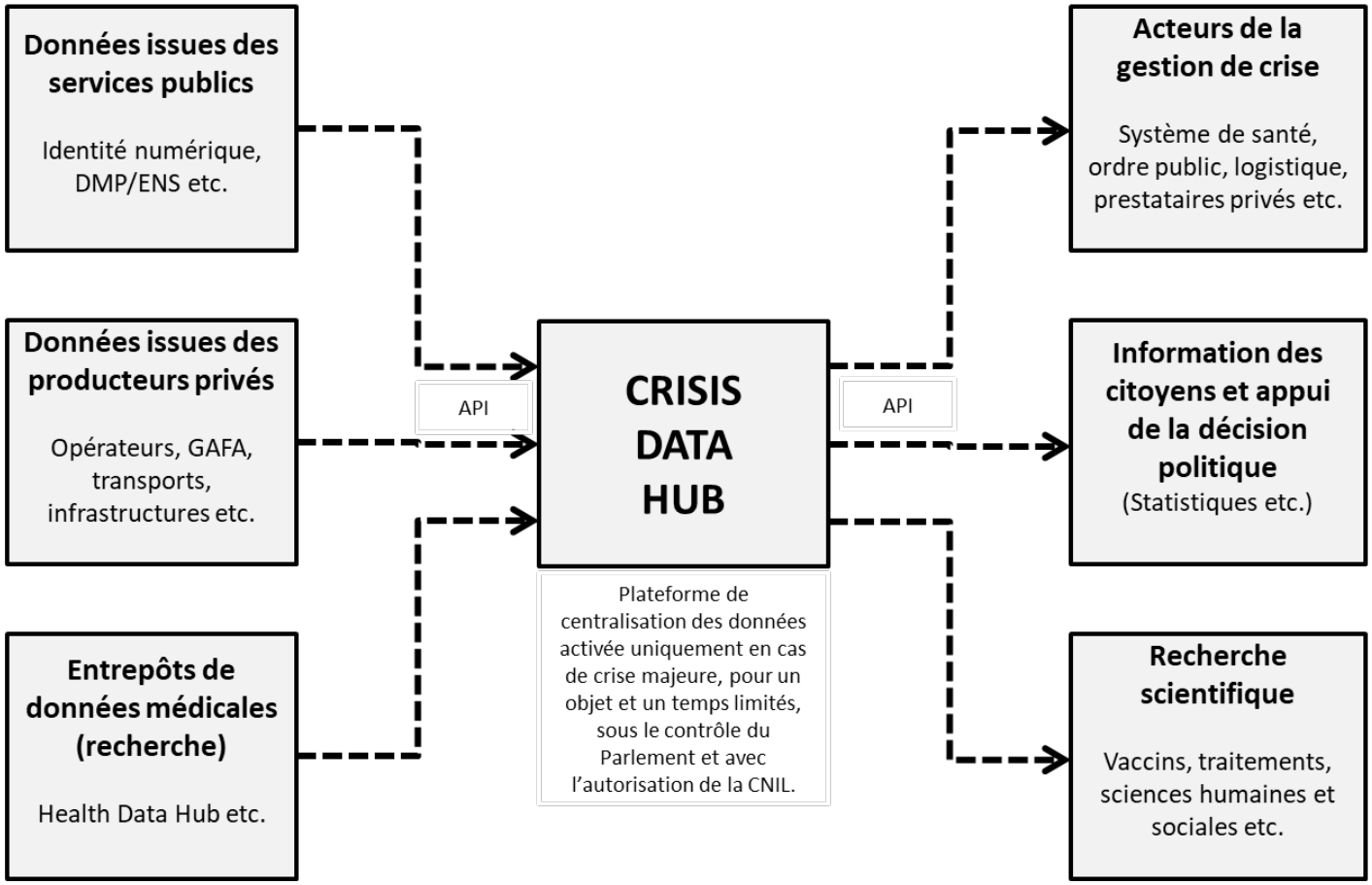
For ex : credit scoring.



https://lenddo.com/

Crisis Mitigation?

For ex : the "crisis data hub"[3].



[3] French Senate report num. 673 (3rd June 2021)
`http://www.senat.fr/rap/r20-673/r20-67322.html`

Etc.

(point/mass surveillance, impersonation, influencing voters, *etc.*)

(health science, public transportation planning, energy transition, *etc.*)

Progress of the Talk

Privacy and Computer Science I

General goal: Solve the eternal tradeoff between security/privacy and utility!

- Many meanings of "privacy"
- Many privacy-preserving techniques

Privacy and Computer Science II

Privacy-Preserving Techniques

- Any security technique applied to personal data can help (e.g., encryption schemes, access control mechanisms).
- Specificities:
  - **Specific security properties** required by the context (e.g., related to vote privacy)
  - **Specific tolerance to non-negligible leaks** because data must be **both used and protected** (e.g., disclosure of useful statistics by "anonymized" data).

Privacy-Preserving Data Publishing

Privacy-Preserving Data Publishing (PPDP) :

- Publish *personal data* for analysis purposes (accurate aggregate queries)...
- ...while preserving individuals' *privacy* (uncertain point queries)
- Also called *sanitization* (please do) and sometimes *anonymization* (please don't)

Privacy-Preserving Data Publishing and the Law

Scope of the EU law :

- GDPR (General Data Protection Regulation) : the new European regulation about the protection of personal data.
- Protects personal data.
- It does not apply to personal data "made anonymous".

$\Rightarrow$ Crucial to define "personal data" and "made anonymous" !

An Encompassive Definition of Personal Data

GDPR Article 4[4] **(1)** : "*'personal data' means*

- **any information relating to an identified or identifiable natural person** *('data subject');*
- *an identifiable natural person is* **one who can be identified, directly or indirectly**,
- *in particular by reference to an identifier such as*
  - **a name**,
  - **an identification number**,
  - **location data**,
  - **an online identifier**
  - *or to one or more factors specific to the* **physical**, **physiological**, **genetic**, **mental**, **economic**, **cultural** *or* **social identity** *of that natural person;*"

[4] https://gdpr-info.eu/art-4-gdpr/

But a Fuzzy Definition of Privacy-Preserving Data Publishing Techniques

GDPR Recital 26 [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 Regulation does not therefore concern the processing of such anonymous information, including for statistical or research purposes.*"

⇒ Crucial to design strong privacy-preserving data publishing techniques !

[5] https://gdpr-info.eu/recitals/no-26/

Progress of the Talk

Pseudonymization and Scandals

Historically, pseudonymization was considered as a valid anonymization method. But:

GIC 2002 [4] : Health data of the GIC `X` list of voters from the state of Massachussets ⇒ Health folder of M. Weld (governor)

AOL 2006 [1] : search queries `X` ground investigation ⇒ Interests of Mrs T. Arnold (USA citizen)

Netflix 2006 [3] : private movie ratings on Netflix `X` a few public movie ratings on IMDB together with user names ⇒ Private movie ratings of two users (with high probability)

NYC 2013 : taxi trips `X` "dé-pseudonymisation" table or blogs of *stars*.

etc ...

Pseudonymized Data: A Join is Enough

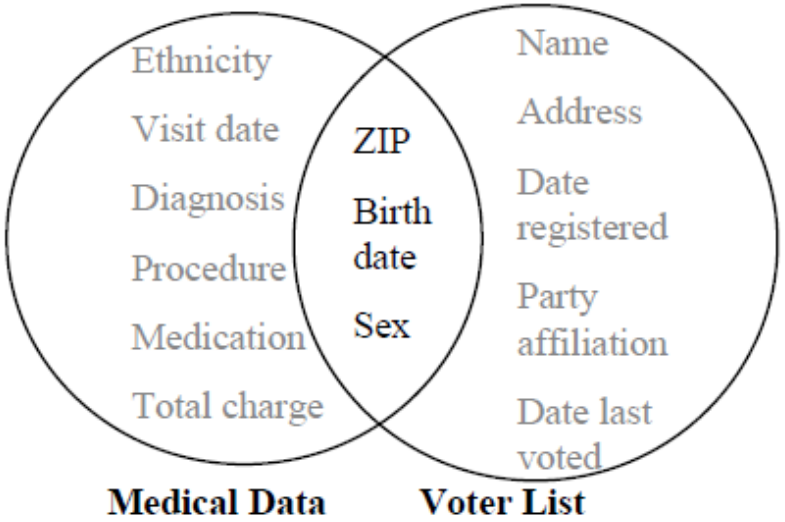


Figure: Gov. Weld's Case : `Medical JOIN Voter ON (zip, DoB, sex)`

A straightforward disclosure

- Governor Weld's case lived in Cambridge and was part of the GIC dataset.
- In the voter list: 6 individuals had his birthdate, 3 of them were men, only one had Weld's zipcode.
- `(zip, DoB, sex)` : a *quasi-identifier*.

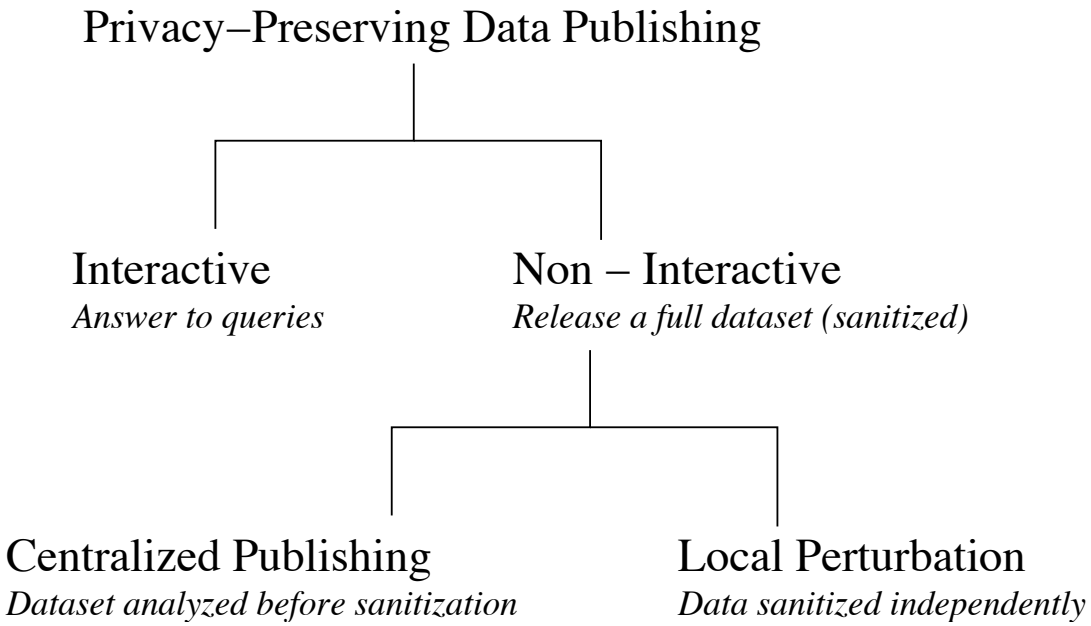
Components of a Privacy-preserving Data Publishing Solution

Three components:

1. **Privacy model**: What does it mean for the data released to be privacy-preserving?
2. **Privacy mechanism**: How to produce the privacy-preserving data to be released?
3. **Utility metric**: How much useful is the released data?

Pseudonymity does not work. . . Which component(s) does it miss ?

Variations on a Theme



- Privacy mechanism: in one of the following families: interactive, centralized publishing, local perturbation.
- Privacy model: essentially agnostic to these families.

Let's go?

Progress of the Talk

[1] M. Arrington.
AOL Proudly Releases Massive Amounts of Private Data.
TechCrunch, 6th of August 2006.

[2] J. G. Cabañas, A. Cuevas, A. Arrate, and R. Cuevas.
Does facebook use sensitive data for advertising purposes?
Commun. ACM, 64(1):62–69, Dec. 2020.

[3] A. Narayanan and V. Shmatikov.
Robust de-anonymization of large sparse datasets.
In *2008 IEEE Symposium on Security and Privacy (S&P 2008), 18-21 May 2008, Oakland, California, USA*, pages 111–125, 2008.

[4] L. Sweeney.
k-anonymity: a model for protecting privacy.
Int. J. Uncertain. Fuzziness Knowl.-Based Syst., 10(5):557–570, 2002.