Introduction to Data Privacy

104369: Privacitat de Dades i Seguretat

G. Navarro-Arribas

UAB -- DEIC

20261014

©G. Navarro-Arribas – (2026)

Example: sharing a database

Name	Passport	Profession	City	Age	Salary
Alfredo Pareja	AS3682	Teacher	Guayaquil	23	600
Eugenio Espejo	RX3453	Teacher	Quito	48	500
Raul Perez	TS5645	Police	Loja	20	734
Francisco Proaño	XB3456	Writer	Guayaquil	32	543
Cesar Davila	AE4324	Teacher	Cuenca	38	890
Ignacio Laso	FF3455	Writer	Quito	41	678
Raul Serrano	LO0903	Police	Portoviejo	21	399
Luz Argentina	KO8264	Police	Quevedo	44	943
Abdon Ubidia	RC3766	Teacher	Machala	31	632

How will you share or publish this? Is all the information needed?

©G. Navarro-Arribas – (2026)

Example: sharing a database with naive anonymization

Name	Passport	Profession	City	Age	Salary
XXX	XXX	Teacher	Guayaquil	23	600
XXX	XXX	Teacher	Quito	48	500
XXX	XXX	Police	Loja	20	734
XXX	XXX	Writer	Guayaquil	32	543
XXX	XXX	Teacher	Cuenca	38	890
XXX	XXX	Writer	Quito	41	678
XXX	XXX	Police	Portoviejo	21	399
XXX	XXX	Police	Quevedo	44	943
XXX	XXX	Teacher	Machala	31	632

Is the privacy of the citizens protected?

©G. Navarro-Arribas – (2026)

Example 2: sharing a computation[1]

- We want to share the **mean income** of a psychiatric unit of the hospital in a small town in Majorca.
  - **Private** List of incomes:
    1000, 2000, 3000, 2000, 1000, 6000, 2000, 10000, 2000, 4000.
  - **Public** mean: **3300**

Can we infer private information?

[1] From [**?**]

Example 2: sharing a computation[1]

- We want to share the **mean income** of a psychiatric unit of the hospital in a small town in Majorca.
  - **Private** List of incomes:
    1000, 2000, 3000, 2000, 1000, 6000, 2000, 10000, 2000, 4000.
  - **Public** mean: **3300**
- Supose, the landlord of Bearn: Dona Obdúlia, attends the phsychiatric unit, then:
  - **Private** List of incomes:
    1000, 2000, 3000, 2000, 1000, 6000, 2000, 10000, 2000, 4000, **10**
  - **Public** mean: **12090.90**

Can we infer private information?

[1] From [**?**]

?

- What is private data/information?
- What about anonymous data/information?
- Can privacy/anonymity be possible?

©G. Navarro-Arribas – (2026)

Defining privacy

Framework:

- Senders
- Recipients
- Messages
- Communication network

→ Malicious actors, act againts some protection objective: attacker, adversary, intruder, ...

→ Adversary goal: increase **knowledge** on **items of interest** (IOI).

Definitions: Anonymity

Anonymity

Anonymity of a subject means that the subject is not identifiable within a set of subjects, the anonymity set. [**?**]

→ **Anonymity set**: size determines degree of privacy. But not only size!

Anonymity

From an adversary's perspective, anonymity of a subject s means that the adversary cannot achieve a certain level of identification for the subject s within the anonymity set. [**?**].

More Definitions

Unlinkability

Unlinkability of two or more items of interest (IOI) from an attacker's perspective means that within the system (comprising these and possibly other items), the attacker cannot sufficiently distinguish whether these IOIs are related or not. [**?**]

©G. Navarro-Arribas – (2026)

Disclosure

Disclosure

Disclosure takes place when attackers take advantage of the observation of available data to improve their knowledge on some confidential information about an item of interest. [**?**]

- Identity disclosure
- Attribute disclosure

More Definitions

Undetectability

Undetectability of an item of interest (IOI) from an attacker's perspective means that the attacker cannot sufficintly distinguish whether it exists or not. [**?**]

Unobservability

Unobservability of an item of interest means:

- Undetectability of the IOI against all subjects uninvolved in it, and
- Anonymity of the subject(s) involved in the IOI even against the other subject(s) involved in that IOI. [**?**]

Identity

Pseudonym

A **pseudonym** is an identifier of a subject other than one of the subject's real names. [**?**]

Identity

An **identity** is any subset of attribute values of an individual person which sufficiently identifies this individual person within any set of persons. So usually there is no such thing as "the identity", but several of them. [**?**]

Plausible Deniability

- I state that I have nothing to do with the information being released. [**?**]. E.g.
  - Can I (credibily!) state that my data has not been used to train a given model?
  - Can I state that a given record in a database is not mine?
  - Can I state that I am not in a given database?

©G. Navarro-Arribas – (2026)

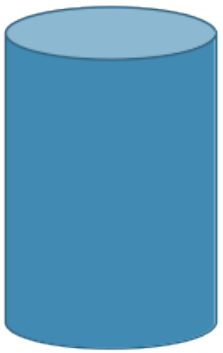
Privacy by design

Privacy “must ideally become an organization’s default mode of operation” [**?**]

1. Proactive not reactive; Preventative not remedial.
2. Privacy as the default setting.
3. Privacy embedded into design.
4. Full functionality—positive-sum, not zero-sum.
5. End-to-end security—full lifecycle protection.
6. Visibility and transparency—keep it open.
7. Respect for user privacy—keep it user-centric.

©G. Navarro-Arribas – (2026)

Motivating example



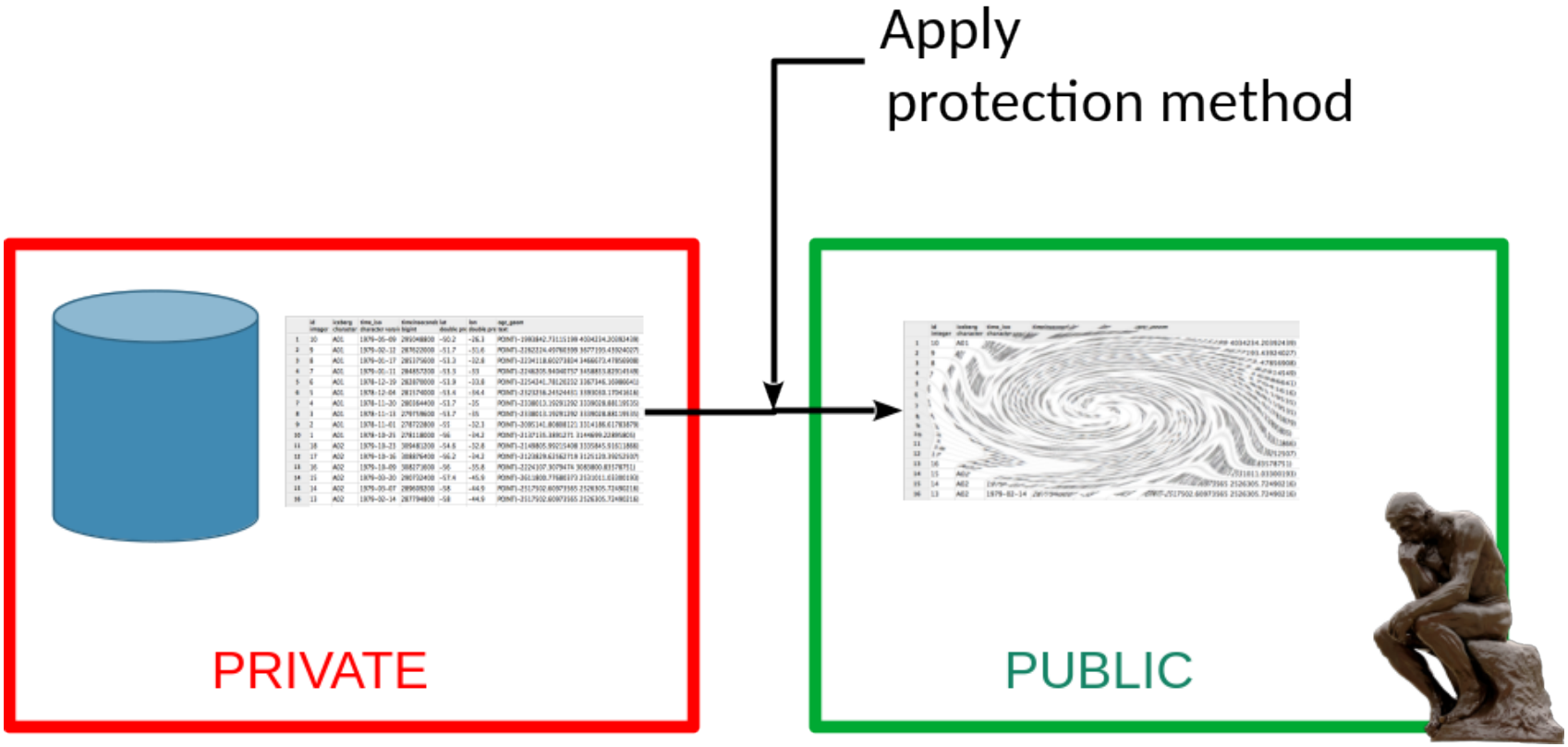
Database

Data publication

	id integer	iceberg character	time_iso character varyin	timeinseconds bigint	lat double pre	lon double pre	wgs_geom text
1	10	A01	1979-05-09	295048800	-50.2	-26.3	POINT(-1993842.73115189 4034234.203924389)
2	9	A01	1979-02-12	287622000	-51.7	-31.6	POINT(-2262224.49760389 3677183.43824827)
3	8	A01	1979-01-17	285375600	-53.3	-32.8	POINT(-2234118.80273834 3466673.47856808)
4	7	A01	1979-01-11	284857200	-53.3	-33	POINT(-2246205.94040737 3458653.82914549)
5	6	A01	1978-12-19	282878000	-53.9	-33.8	POINT(-2254241.78120232 3367146.16886643)
6	5	A01	1978-12-04	281574000	-53.4	-34.4	POINT(-2323256.24524431 3393030.17941816)
7	4	A01	1978-11-20	280364400	-53.7	-35	POINT(-2338013.19291292 3339028.881195355)
8	3	A01	1978-11-13	279759600	-53.7	-35	POINT(-2338013.19291292 3339028.881195355)
9	2	A01	1978-11-01	278722800	-55	-32.3	POINT(-2095141.80868121 3334186.617831879)
10	1	A01	1978-10-25	278118000	-56	-34.2	POINT(-2137135.3892271 3144698.228858855)
11	18	A02	1979-10-23	309481200	-54.6	-32.8	POINT(-2149805.99215488 3335845.916118667)
12	17	A02	1979-10-16	308876400	-56.2	-34.2	POINT(-2123829.62562719 3125120.392525507)
13	16	A02	1979-10-09	308271600	-56	-35.8	POINT(-2224107.3079474 3083808.835787511)
14	15	A02	1979-03-20	290732400	-57.4	-45.9	POINT(-2611808.77680373 2631011.033001993)
15	14	A02	1979-03-07	289608200	-58	-44.9	POINT(-2517502.80973585 2526385.724902161)
16	13	A02	1979-02-14	287794800	-58	-44.9	POINT(-2517502.80973585 2526385.724902161)

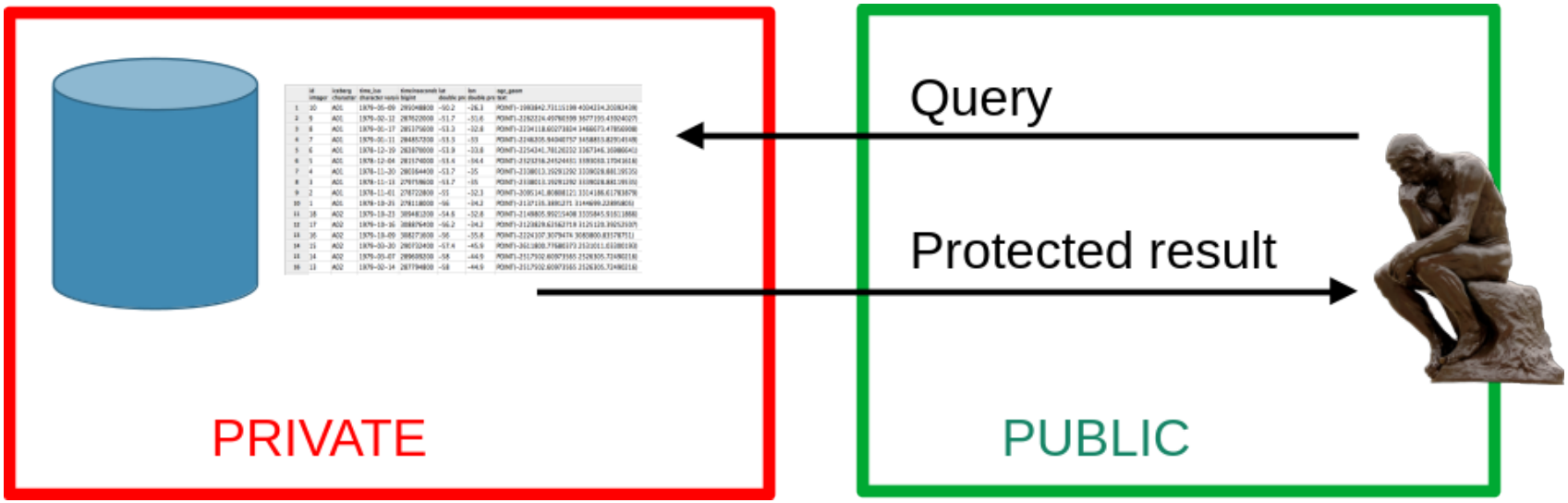
- Public release
- Research
- Long term storage

Non-interactive scenario



©G. Navarro-Arribas – (2026)

Interactive scenario



©G. Navarro-Arribas – (2026)

Types of attributes related to an entity: example

Name	Passport	Profession	City	Age	Salary
Alfredo Pareja	AS3682	Teacher	Guayaquil	23	600
Eugenio Espejo	RX3453	Teacher	Quito	48	500
Raul Perez	TS5645	Police	Loja	20	734
Francisco Proaño	XB3456	Writer	Guayaquil	32	543
Cesar Davila	AE4324	Teacher	Cuenca	38	890
Ignacio Laso	FF3455	Writer	Quito	41	678
Raul Serrano	LO0903	Police	Portoviejo	21	399
Luz Argentina	KO8264	Police	Quevedo	44	943
Abdon Ubidia	RC3766	Teacher	Machala	31	632

©G. Navarro-Arribas – (2026)

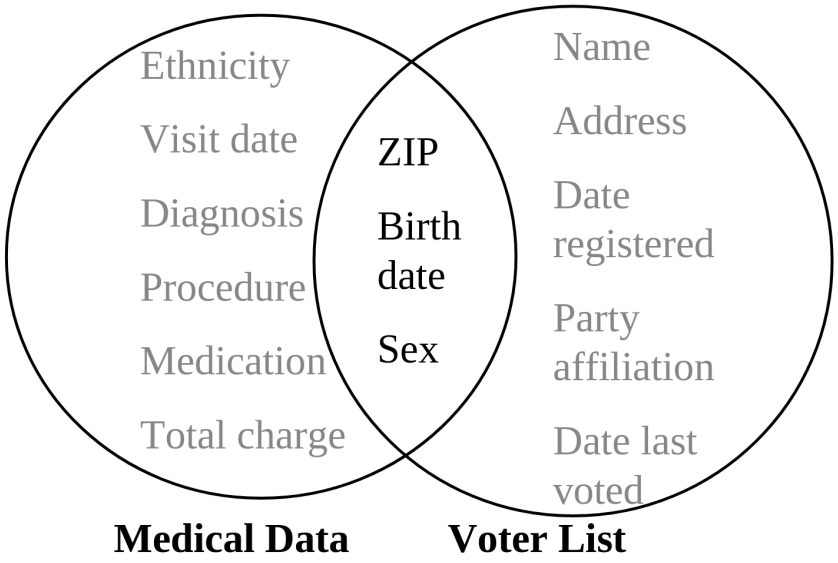
Types of attributes related to an entity

- Identifiers: Unambiguously identify the entity (passport number, full name, ...)
  - Usually removed or encrypted for public release.
- Quasi-identifiers: in combination, can be linked with external information to re-identify an entity (age, zip code, city, ...)
  - Usually a protection mechanism is applied to these attributes.
- Confidential: contain sensitive information about the entity (salary, religion, political affiliation, health condition, ...)
  - Usually not modified (object of study, confidential thus not known by attackers).

©G. Navarro-Arribas – (2026)

Importance of Quasi-identifiers

Example: US Census and Medical data linkability.



- Researchers found that:
  - date of birth
  - gender
  - 5-digit ZIP code
- Uniquely identifies 87.1% of USA population!!! [**?**]

©G. Navarro-Arribas – (2026)

Other studies

- Review of Sweeney results with 2000 Census data (more recent) [**?**]:
  - {gender, ZIP code, full DOB}: **63%** of the U.S. population uniquely identifiable (lower than Sweeney's 87%).
  - {gender, county, full DOB}: **14.8%** unique (1-anonymous); 43.6% are 5-anonymous or less.
- Probability of correctly identification from a incomplete sample of the population population [**?**]:
  - **99.98%** of Americans would be correctly re-identified in *any* dataset using just **15** demographic attributes.

Behavioral quasi-identifiers

- Mobility traces [**?**]:
  - Just **4** randomly chosen spatio-temporal points are enough to uniquely characterize **95%** of individual traces. Two points already uniquely identify more than 50%.
- Credit card metadata [**?**]:
  - **4** spatiotemporal purchase points (approximate place + date) are sufficient to uniquely re-identify **90%** of individuals; knowing the approximate price of a purchase increases uniqueness further.

©G. Navarro-Arribas – (2026)

References i