Privacy

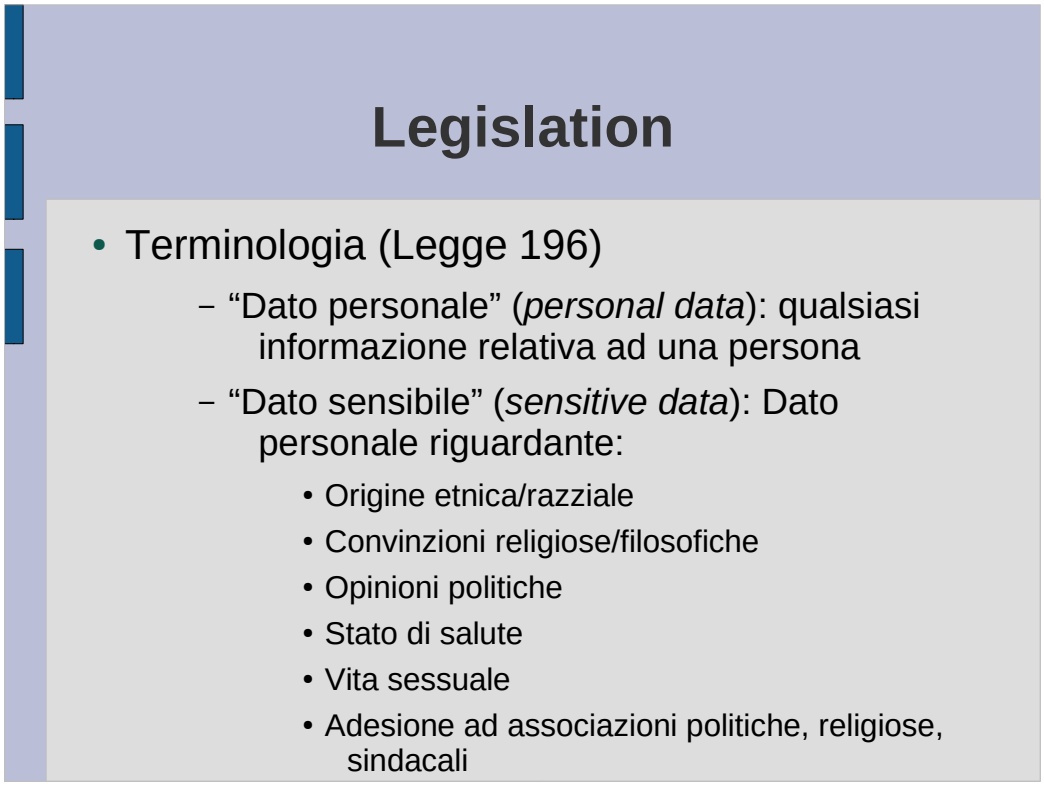
Pericle Perazzo
pericle.perazzo@iet.unipi.it

Legislation

- EU
  - 1995 EU Information Directive (95/46/EC)
  - 2002 EU Directive on Privacy and Electronic Communications (2002/58/EC)
- USA
  - No comprehensive privacy protection law
  - HIPAA (healthcare data)
  - GLB (financial data)
  - COPPA (children data)

Nowadays, there are two references for privacy protection law, the USA laws and the EU directives. The EU directives are the 95/46/EC, concerning information privacy in general, and 2002/58/EC, focusing on digital privacy.

The USA have not a comprehensive privacy protection law. Instead, they have a set of laws specific for different kinds of data: HIPAA for healthcare data, GLP for financial data, COPPA for children's privacy.



La legge italiana n. 196 che regola il trattamento delle informazioni personali è un'applicazione delle direttive UE 95/46/EC e 2002/58/EC. Segue una descrizione della terminologia usata:

Un *dato personale* (personal data) è una qualsiasi informazione relativa ad una persona, che può identificare (direttamente o congiuntamente ad altre informazioni) la persona. La definizione quindi è molto ampia e comprende tutto ciò che può caratterizzare una persona, dal nome al colore dei capelli.

Tra i dati personali, un'attenzione particolare ricevono i *dati sensibili* (sensitive data), che riguardano origine etnica/razziale, convinzioni religiose/filosofiche, opinioni politiche, stato di salute, vita sessuale, adesione ad associazioni politiche, religiose o sindacali.

Legislation

- Terminologia (Legge 196)
  - “Trattamento” (*process*): qualunque operazione sui d.p.
  - “Interessato” (*data subject*): persona a cui si riferiscono i d.p.
  - “Titolare”: persona che decide le finalità e le modalità di trattamento dei d.p.
  - “Responsabile”: persona responsabile dell'attuazione di queste finalità e modalità

Un *trattamento* (process) è una qualunque operazione, anche mediante strumenti automatici, su dati personali. La definizione non comprende soltanto la lettura, ma anche la modifica, l'elaborazione e la cancellazione.

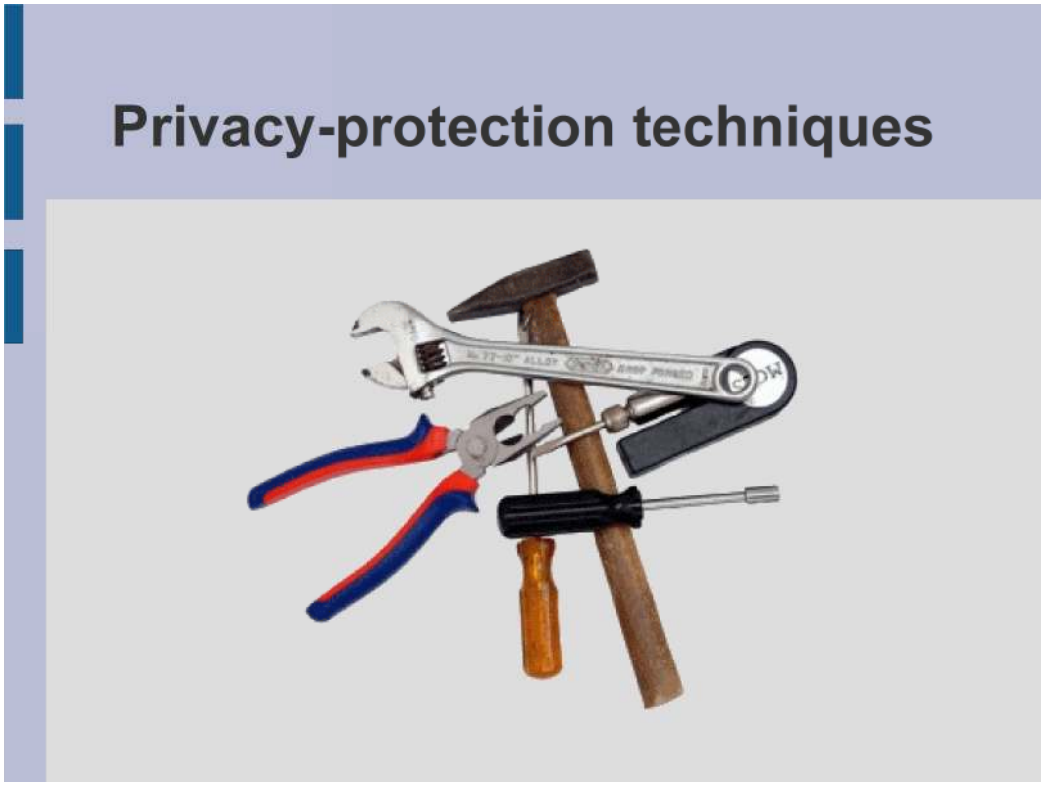
L'*interessato* (data subject) è la persona, fisica o giuridica, a cui si riferiscono i dati personali.

Il *titolare*, è colui (o coloro) che decide le finalità e le modalità del trattamento.

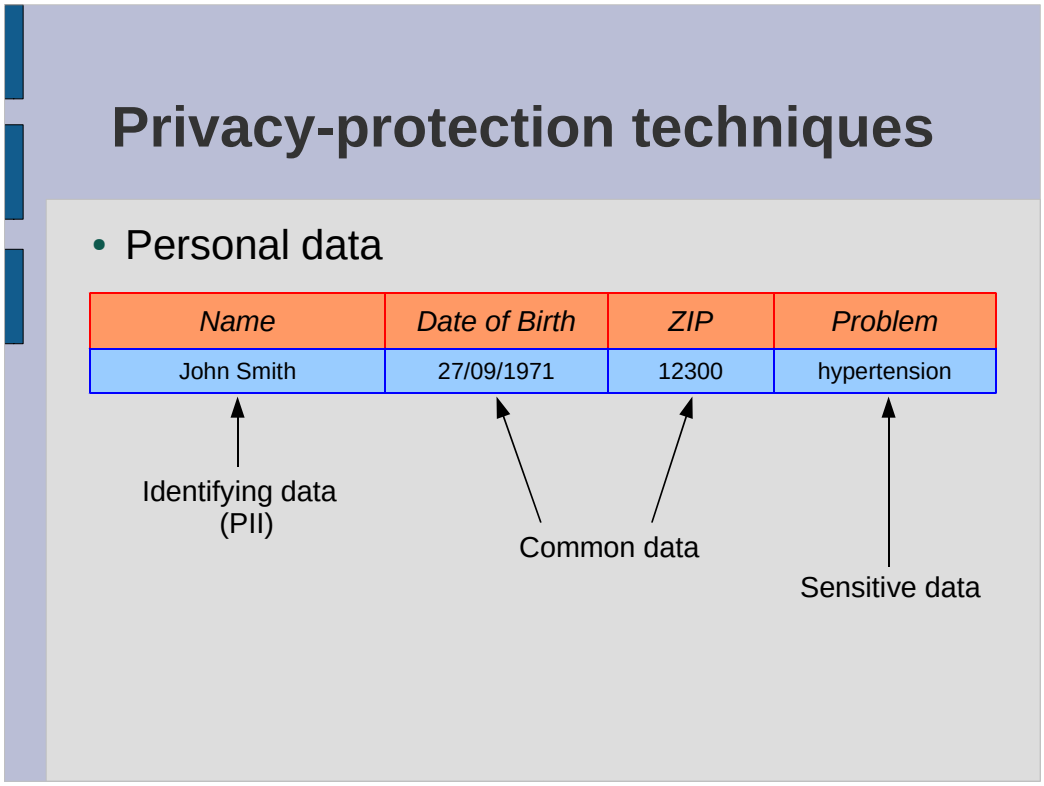
Il *responsabile* è colui (o coloro) che mettono in pratica tali finalità e modalità.

Tutte le leggi riguardanti la privacy ruotano attorno al concetto di *finalità* con cui si raccolgono i dati personali. Tali finalità devono essere determinate, esplicite e legittime. I dati raccolti devono essere pertinenti e non eccedenti le finalità (principio di necessità). Bisogna quindi poter dimostrare che i dati raccolti ed i trattamenti su di essi siano necessari al conseguimento della finalità.

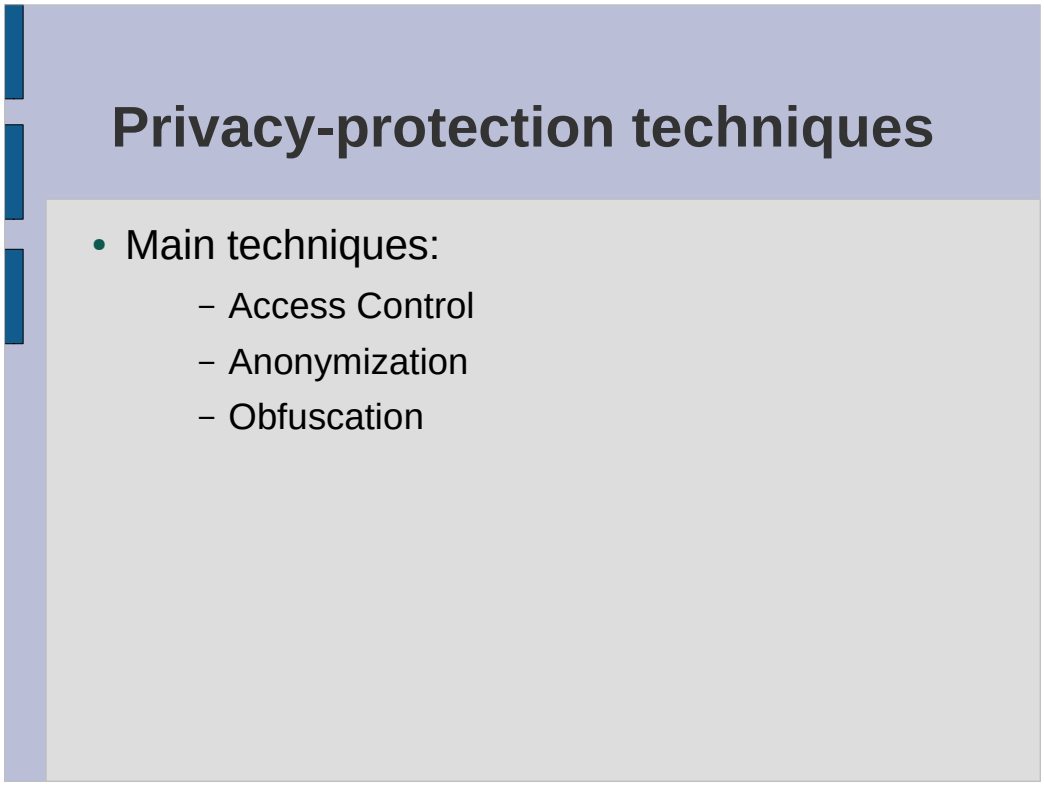
Inoltre i dati devono essere distrutti o resi anonimi una volta cessate le finalità. Infine, l'interessato deve essere precedentemente informato e deve dare il suo consenso alla raccolta e al trattamento dei dati personali. Quest'ultimo punto può essere tralasciato in caso di finalità giudiziarie o di indagine da parte delle istituzioni autorizzate (polizia, magistratura), nel caso in cui informare l'interessato potrebbe inquinare le finalità stesse.



All the techniques for privacy protection aim at reducing the quantity or the quality of the disclosed data, in order to both satisfy the privacy and offer the service.



From the technical point of view, personal data is a tuple in a database. The image above shows an example from a medical scenario. The database contains the symptoms of all the patients of a hospital. All the four fields are personal data. The name is also “identifying data” because it permits a direct identification of the person. The date of birth and the postal code are common data. The symptom is sensitive data, because it concerns healthcare information.



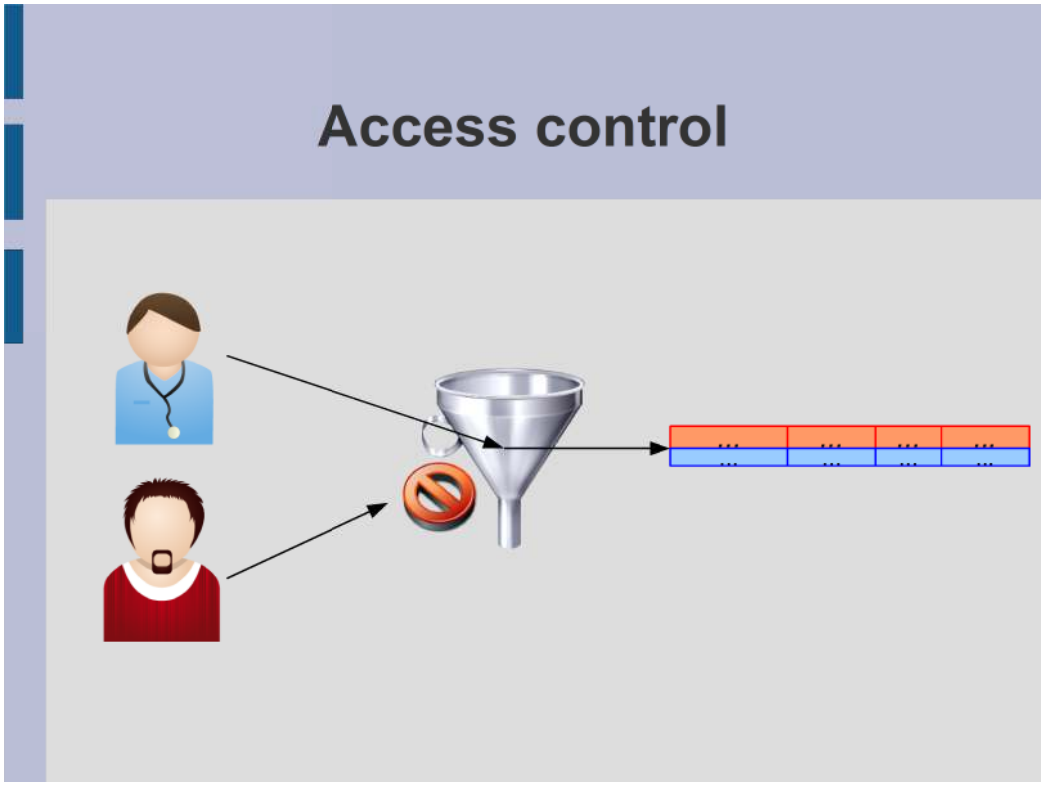
The main techniques for information privacy are the following:

1) *Access control* decides who can access what information. This is mandatory for every privacy-concerned system. In fact, without access control, the data is public, thus no real privacy policy can be applied.

2) *Anonymization*. This techniques disjoints the personal data from the identity of the subject.

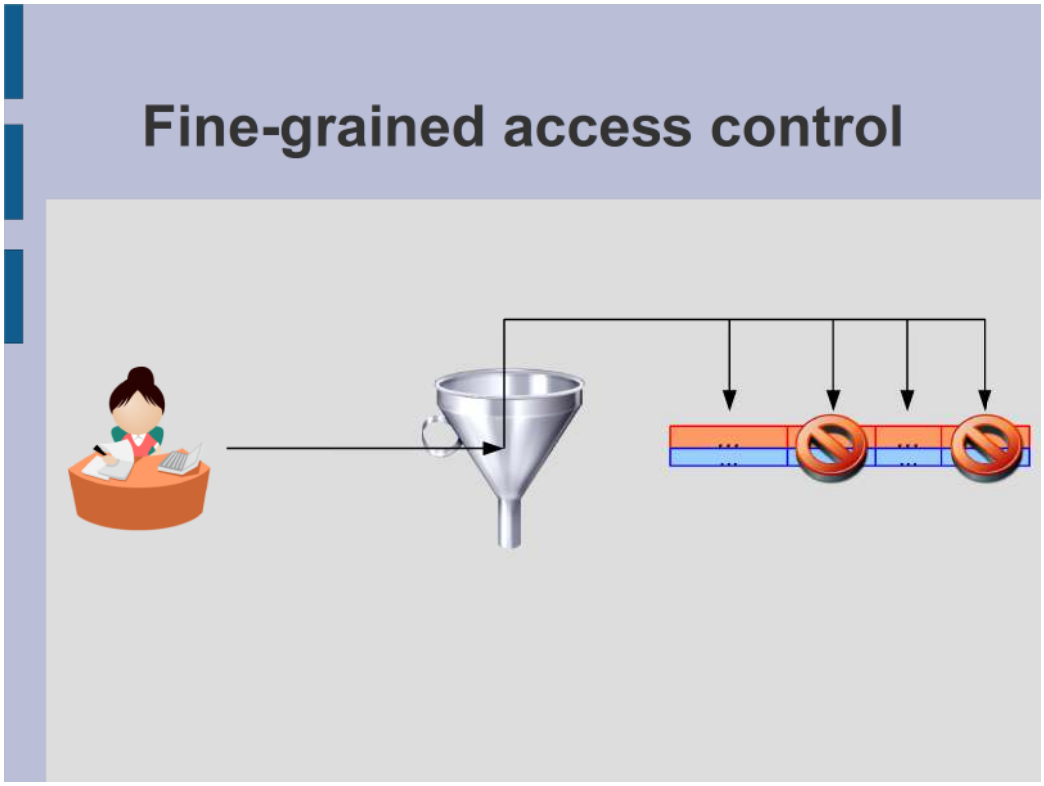
3) *Obfuscation*. This technique degrades the precision of the personal data.

These techniques do not exclude each other. A system can apply a mix of access control, anonymization and obfuscation.

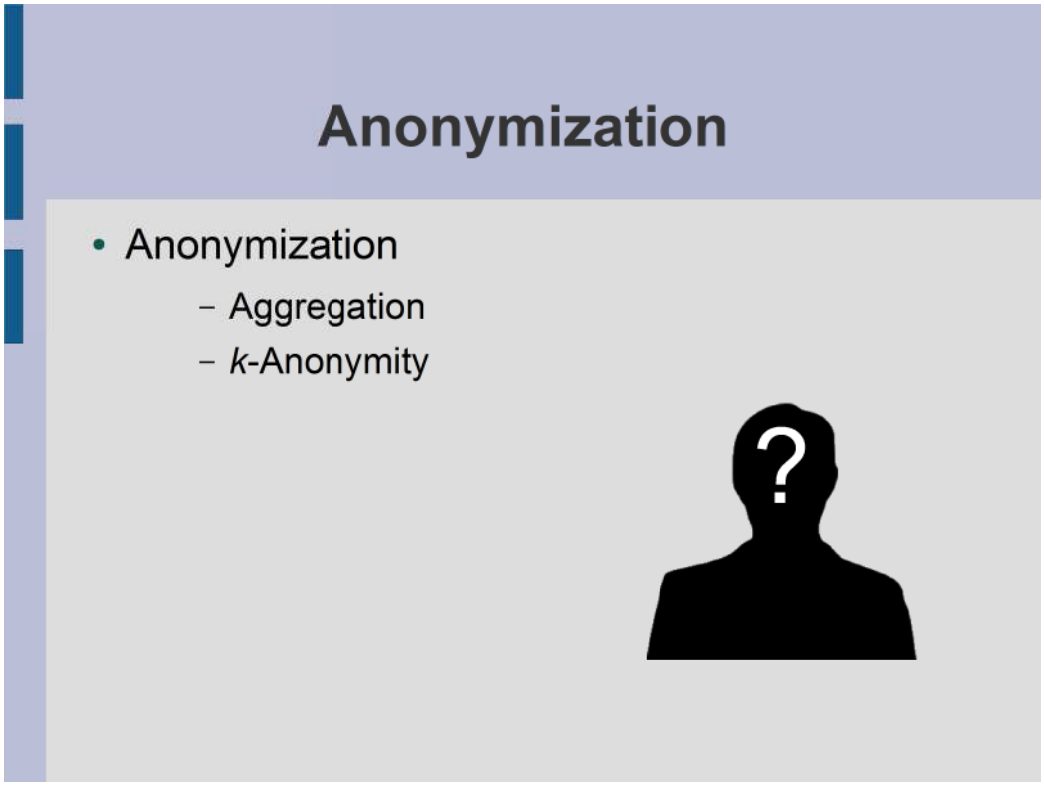


The access control mechanism decides who is authorized to access data and who is not. This is usually implemented with a role-based policy, in which each user has a role (e.g. doctor, secretary, technician, etc.) and assumes the authorizations of his role. In some systems, a user can have a set of roles, and his authorizations are the union of his roles' authorizations.

In the example above, a doctor can access patients' information, whereas a generic user cannot.

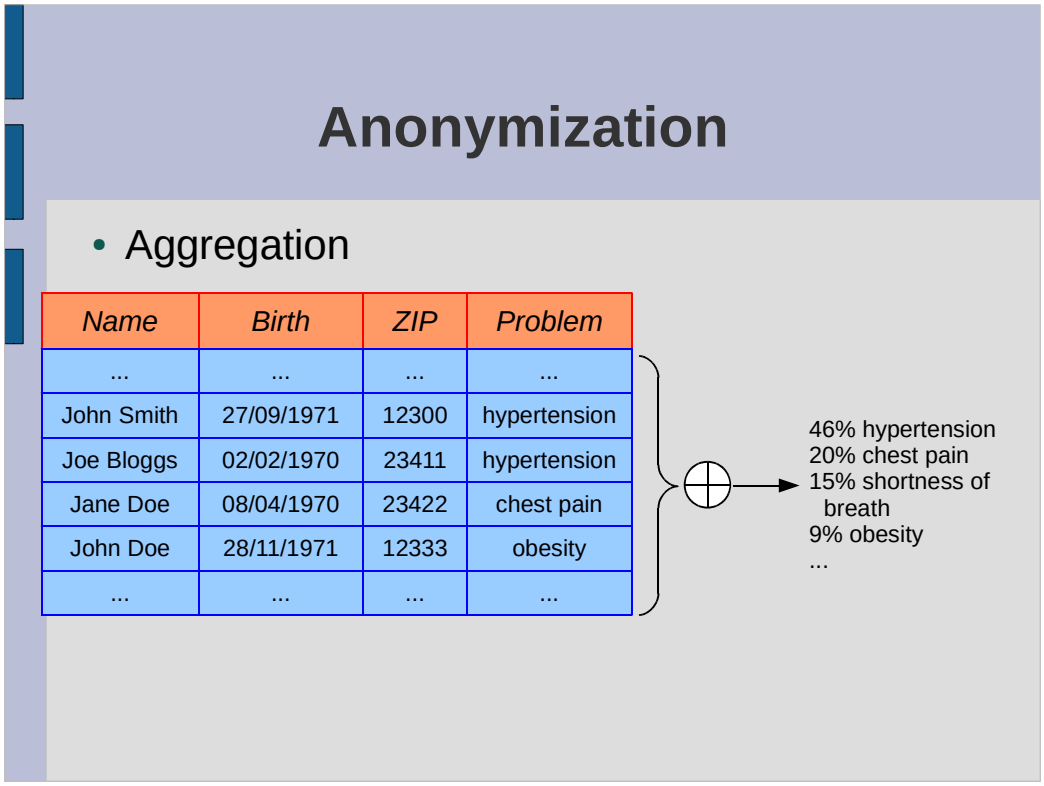


The access control can also be fine-grained. In the example above, a secretary has to send a letter to a patient, so she has to access only the name and the postal code.



Let us suppose now that an external research organization needs to know the information of the database in order to extract some statistics. The organization is interested in symptoms and medical information, but it is unconcerned about the identity of the patients.

An idea is to anonymize the tuples. This anonimization process can be performed by means of two techniques: aggregation and *k*-anonymity.



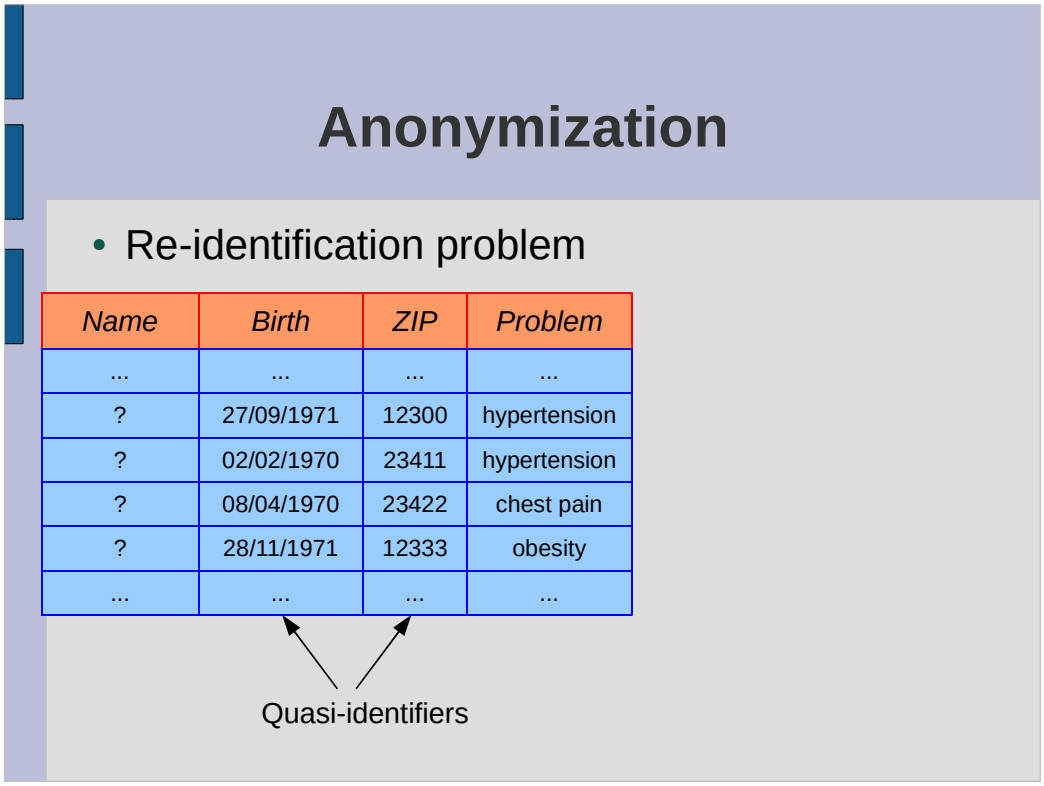
The anonymization through *aggregation* works this way: the percentages of the different problems are computed and disclosed, without disclosing the entire database. This offers a perfect anonymization to the patients, but could be unsatisfactory for the research organization. In fact, more complex statistics could be useful, for example the percentage of occurrence of a specific problem in relation to the age, or in relation to the place of residence.

Anonymization

- De-identification

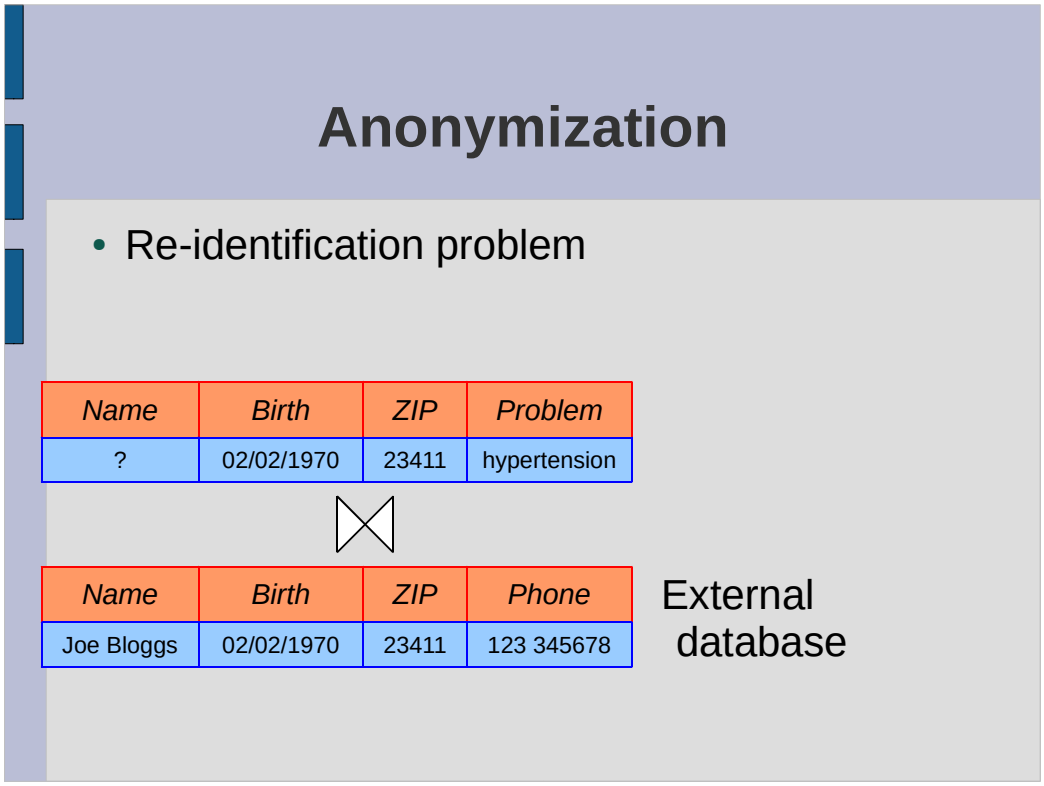
Name	*Birth*	*ZIP*	*Problem*
...	...	...	...
?	27/09/1971	12300	hypertension
?	02/02/1970	23411	hypertension
?	08/04/1970	23422	chest pain
?	28/11/1971	12333	obesity
...	...	...	...

Another idea is to disclose the entire database, except for the identifying data. This is called “de-identification”.



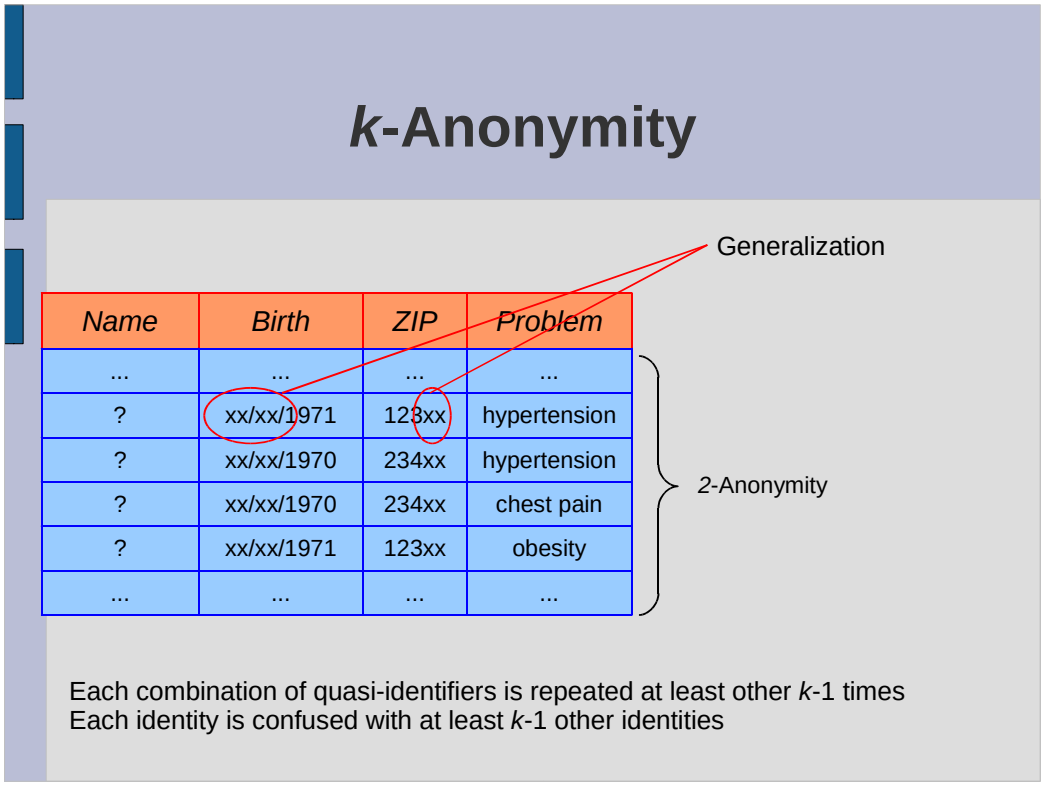
The de-identification suffers from the problem of *re-identification*. In fact, some disclosed personal data, in our example the date of birth and the postal code, are *quasi-identifiers*. A quasi-identifier could reveal the identity of the subject indirectly, with the help of some other external information.

For example, what happens if only one person in the world is born on 02/02/1970 and contemporaneously live in the zone having 23411 as postal code?



Let us suppose that a public database exists, which contains our quasi-identifiers (birth date, postal code) together with identifying data (name). Such a database can be a telephone directory, a registry office or a voter list.

One could join the two databases in order to re-identify the person.



The *k*-Anonymity assures that each identity is confused at least with *k*-1 other identities in the same database. This is done by obfuscating (through generalization) the quasi-identifiers.

In the image above, the day of birth, the month of birth, and the first two digits of the postal code are suppressed. In this way, every combination of quasi-identifiers is repeated at least *k*-1 other times, thus assuring a 2-Anonymity.

k-Anonymity

- Generalization of the quasi-identifiers
- Least generalization necessary to provide a given *k*-anonymity

The aim is to find the minimal generalization necessary to provide a given *k*-Anonymity to the subjects. This usually involves complex computations.

k-Anonymity

- High level of privacy
- Complex operations to compute minimal generalization
- Not suitable if the identity is required

k-Anonymity assures a high level of privacy to the users. However, it involves complex algorithms to find the least necessary generalization, thus it does not fit well on performance-centric applications.

In addition, *k*-Anonymity is not suitable, along with all the anonymization methods, wherever the identity of the subjects is required.

Beyond *k*-Anonymity

How to protect privacy
if identity is required?

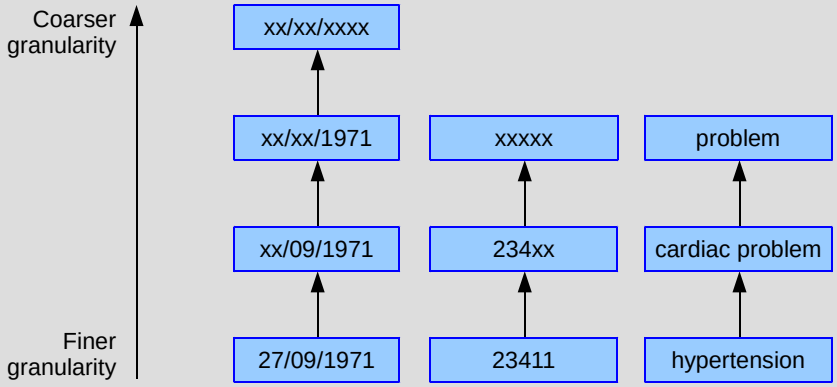
Data obfuscation

- Protect the data, instead of the identity
- Obfuscation: reduction of the data precision:
  - Generalization
  - Perturbation

When the identity of the subjects is required, anonymization techniques are not applicable. A better approach is *obfuscation*, which aims at reducing the precision of the data. This can be done by means of two methods: *generalization* and *perturbation*.

Data obfuscation

- Generalization
  - Make the granularity of information coarser

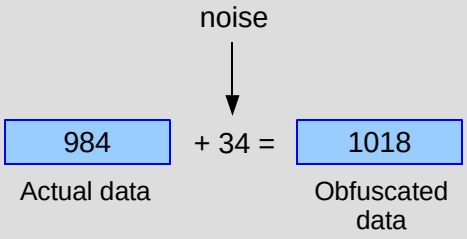


Generalization makes the granularity of information coarser. This means, for example, suppressing the day and the month of a date, or the least significant digits of a number, or substituting a specific health problem with a family of health problems.

This is the same method used to provide *k*-Anonymity. However, a predefined and constant quantity of generalization is applied here, so that we have not to find a “least necessary generalization”. This method is much more efficient that *k*-Anonymity.

Data obfuscation

- Perturbation
  - Add a random noise to information
  - Suitable only for numerical data



  - To avoid multiple-query deobfuscation, noise value must be memorized and re-used

Perturbation adds a random zero-mean noise to data before disclosing it. Obviously, this method is suitable only for numerical data.

If the obfuscated data can be queried several times, the noise must be equal from query to query. Otherwise, one could gather many samples of the obfuscated data and compute the mean value, eliminating in such a way the noise (multiple-query deobfuscation). A way to avoid this is to choose once the random noise, memorize it together with the data, and then re-use it every time such data is asked.

References

- P. Samarati, L. Sweeney "*Protecting Privacy when Disclosing Information: k-Anonymity and Its Enforcement through Generalization and Suppression*"
- EU directive 95/46/EC
  - http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=CELEX:31995L0046:en:HTML
- EU directive 2002/58/EC
  - http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=CELEX:32002L0058:EN:HTML