What is personal data under the GDPR?

A fundamental concept of the European General Data Protection Regulation (GDPR), which came into force in May last year, is personal data.

Every data controller and processor (i.e., data holder), who collects or processes the personal data of European citizens (i.e., data owners) should be aware of the exact meaning of these concepts in order to be compliant with the GDPR and avoid hefty fines posed by the regulation. Aside from legal compliance, there are at least two other good reasons for any data holder to protect personal data. First, privacy protection is a competitive differentiator for companies. Second, privacy scandals can cause serious reputation loss as shown by the notorious cases of Equifax, Yahoo, Ashley Madison, and many more.

Personal data in the GDPR

The GDPR identifies personal data as any information related to an identified or identifiable natural person. For example, if a medical dataset contains the patients' name, hometown, and medical diagnosis, then a record (or "row") within this dataset is personal data if the patient who this record is about can be re-identified, meaning that anybody who has access to this dataset is able to associate the record with the patient.

Rec#	Name	Hometown	Diagnosis
1	Jerry Bilbo	LA	Meningitis
2	Tom Sawyer	NO	Prostate cancer
3	Carl Schwartz	LA	Bronchitis
4	John Smith	NYC	Alzheimer
...	...	...	...

Table 1. Perhaps non-personal data

Rec#	Hometown	Diagnosis
1	LA	Meningitis
2	LA	Prostate cancer
3	NO	Bronchitis
4	NYC	Alzheimer
...	...	...

Table 2. Perhaps personal data

At first sight, Table 1 contains personal data due to the names stored in every record. However, this is not always the case as there can be several natural persons with the same name in a population (e.g., in a city) but not all of them are included in the dataset. For example, if there are several persons named John Smith in NYC, then Record #4 may not be personal data, if the attackers accessing the dataset cannot single out the flesh-and-bone individual from the population who Record #4 is about.

On the other hand, a dataset which does not contain personal names can still be personal. For example, consider an attacker who has access to Table 2 as well as Table 3, where the latter contains some demographic data of people including the patients in Table 2. The attacker can see that there are three

individuals named John Smith in NYC, who are aged 26, 35, and 65, respectively. Therefore, Record #4 in Table 2 belongs to Record #3 in Table 3, since Alzheimer is very rare in age 26 and 35.

This means that the attacker probably re-identified a person named John Smith in Table 2 assuming that Table 3 contains all individuals named John Smith from NYC. The hometown and birth date together is the identifier of John which allows his unambiguous re-identification in Table 2.

Rec#	Name	SSN	Hometown	Date of birth
1	Susan Smith	2346758913	NO	12/12/1965
2	John Smith	4545454323	NYC	01/03/1991
3	John Smith	8375835937	NYC	08/05/1952
4	John Smith	3548469234	NYC	28/11/1982
5	Ursula Mayden	3484756773	LA	30/11/1954
...	...	...	...	...

Table 3. All people in a city

The following question immediately arises: if the second attacker can re-identify the person behind Record #4 of Table 2, but the first attacker cannot do it without Table 3, then is this record personal data according to the GDPR? The answer depends on the plausibility of the second attack. If it is plausible that the second attacker can access Table 1 and Table 2 also, then Record #4, and hence Table 1 are regarded as personal data of John Smith, who is identified by his name and hometown.

Complex data

The above toy example contains only a few attributes of individuals like name, hometown, and diagnosis. However, this is rarely the case in practice. In the era of “Big Data”, most datasets contain many different attributes of individuals, such as the list of visited locations, purchased items in a store, credit card transactions, or the watched movies per person, just to name a few. The multitude of stored attributes about a person is detrimental to their privacy because their record is likely to be unique in the dataset. This potentially makes them easily re-identifiable.

For example, consider a dataset containing the list of credit card transactions per user, including John (a transaction is the day and location of a credit card payment). If an attacker knows the approximate location and day of John’s 4 transactions, then it is very likely that only John’s record has these four transactions. Therefore, even if John’s name, address, and account/card number are removed from the dataset (i.e., it is “pseudonymized”), an attacker who knows four of John’s transactions will find John’s record, as, most likely, only this record contains these 4 transactions in the dataset.

So, is this transactional data personal according to the GDPR? It is, if any record owner is re-identifiable, i.e., a re-identification attack is (1) likely to happen and (2) also likely to succeed. Specifically, the attack must have reasonable (1) plausibility (e.g., is it plausible that the attacker can access the transactional dataset and know John’s 4 transactions?) and (2) success probability (e.g., does only John’s record have these 4 transactions? That is, can the attacker find John’s record in the dataset?). The last question has been answered in a study showing that 4 transactions make a unique record with a chance of 90% in a dataset of 1.1 million individuals.

The first question is much more difficult to answer as no one can argue what an attacker might already know about John from other sources that may be seemingly unrelated to the transactional dataset.

Intuitively, learning four transactions of an individual should be easy, as people share a lot of information about themselves through social media – perhaps John shared his favorite stores or photos about his recently purchased products. In fact, such data sharing can also happen involuntarily. For example, web-trackers and data brokers constantly collect the list of visited web pages of people using web cookies or other means of tracking. The collected web-history of a person can be easily linked to his offline purchases as many people first check current deals online. Therefore, an attacker can associate a web history with a transactional record if it has access to both. This is not a far-fetched scenario; obtaining web-histories of millions is seemingly easy, and re-identification becomes straightforward when a person's web history contains his public Facebook or Twitter profile.

The above example shows that not only hometown or birth date can be identifiers but also credit card transactions. Besides, researchers have shown that 6-8 movies that a person watched can also serve as an identifier in a dataset of 500k Netflix users. For re-identification, the data records can be linked to IMDB where the same individuals often review the same movies but using their real names. Similarly, it has been demonstrated that 5-7 laboratory tests of a patient can be an identifier with a probability of more than 95% among 61K patients depending on the test. Moreover, the list of visited web pages or simply the list of installed apps on a smartphone can equally be identifiers of anybody in a large population.

Unstructured data

The GDPR's definition of personal data is very general and includes many kinds of information which may seem non-personal at first sight. These are not necessarily "structured" or relational datasets like the ones above. For example, an article saying, "A person sells his Mustang in Innsbruck" can be personal data especially if there is only a single person in Innsbruck who has a Mustang.

The source code of a software can be personal data, even without direct authorship information, as the coding style is often unique to a developer. Likewise, reviews about a product made under a pseudonym can still be attributable to the real author due to his/her unique writing style. Medical data like screening images and drug prescriptions can also be personal if there is only a single person in the population who takes the prescribed combination of drugs or has a unique skull shape on a CT scan. Statistical (aggregate) data can also be personal; for example: publishing only the total number of visitors of some locations in a large city in every half an hour allows to reconstruct the individual trajectories of people.

Of course, in order to be personal, the above "pseudonymous" records must be associated with their real flesh-and-bone owners. Indeed, one must argue that there is a plausible attacker who has access to the data above and also knows enough to link individuals to their data. This can be quite easy as people often share pieces of these data publicly elsewhere (without being aware of the potential privacy implications); the Mustang owner can share a photo of his car on Facebook, the developer may publish open source code on GitHub with his real identity, the head shape of a person might be approximated (perhaps soon) from his/her Instagram photos using sophisticated machine learning techniques. Matching these shared pieces (and the corresponding online profiles) with the "pseudonymous" records above, the attacker can easily reveal the real identity of data owners.

Conclusion

Data are personal if the data owners are re-identifiable no matter what the data are about. Re-identifiability depends on who can access the data and what they may know about the data owners from any external source (e.g., social media). These are hard to assess in practice.

First, people share a lot of personal data about themselves willingly or unwillingly which makes linking their real identity to their „pseudonymous" data easy. Second, rapidly developing AI technologies always reveal new, surprising relations between seemingly unrelated information like facial photos and sexual orientation. Some data which seem non-personal today may turn out to be personal tomorrow.