Creating GDPR compliant interpretable models

Pedro Strecht[0000−0002−1077−0346]

INESC TEC/Faculdade de Engenharia, Universidade do Porto, Rua Dr. Roberto Frias, 4200-465 Porto, Portugal, pstrecht@fe.up.pt

Abstract. The enforcement of the General Data Protection Regulation in the European Union as of May 25, 2018 pressured organizations to take action regarding the way they process personal data, under threat of the application of very heavy penalties. The GDPR includes recommendations to be adopted in order for organizations to comply. Scientific research often makes use of interpretable models to describe or predict some phenomenon of interest. The datasets used to create them may contain information in the scope of personal data. This paper frames the creation of interpretable models with the privacy principles described in the GDPR and points out the specific safeguards to be deployed in the operations of data extraction and further processing to foster the conformity with it.

Keywords: GDPR compliance · Interpretable models · Safeguards

1 Introduction

The *General Data Protection Regulation* (hereinafter referred to as GDPR) [1] is a large and dense document, consisting of 173 recitals (guidances) and 99 articles (requirements), was approved by the European Commission (EC) on 27 April 2016 and is law from 25 May 2018 (replacing the previous EC Data Protection Directive of 1995). The GDPR deals with the protection of personal data of European Union (EU) citizens and also applies to organisations doing business with the EU.

A major change to previous legislation is the requirement to notify a supervisory authority of a personal data breach whenever there is a risk to the rights and freedoms of people. Organizations have to this in a period no longer than 72 hours. Fines for non-compliance are increased to a maximum of 4% of global turnover. Actual penalties will depend on a number of factors including the cause and size of the breach, the controls in place and the degree of co-operation with the supervisory authority.

The goal of this paper is to frame the creation of interpretable models with the privacy principles described in the GDPR and identifies the measures to promote the compliance with it. The remainder of this paper is structured as follows. Section 2 presents the fundamental concepts described by the GDPR. Section 3 introduces interpretable models and discusses the problem of exposing personal data in them. Section 4 recommends measures to adopt in order to promote the compliance with the GDPR. Section 5 concludes with a few remarks.

2 Fundamentals concepts of the GDPR

2.1 Key terms

The pursuance of the GDPR reinforced the need to clarify a number of key concepts, listed in Article 4, of which the following stand out:

– *Personal data* is any information relating to an identified or identifiable person, the *data subject*. One that can be identified, directly or indirectly, in particular by reference to an identifier. Examples are name, an ID number, location data, physical, physiological, genetic, mental, economic, cultural or social identity of that natural person.
– *Processing* is any operation which is performed on personal data, whether or not by automated means, such as collection, organisation, storage, adaptation, retrieval, use, disclosure by transmission, or destruction.
– *Controller* is the natural or legal person, public authority, agency or other body which, alone or jointly with others, determines the purposes and means of the processing of personal data.
– *Processor* is a natural or legal person, public authority, agency or other body which processes personal data on behalf of the controller.

2.2 Principles relating to processing of personal data

In Article 5, the GDPR establishes seven *privacy principles* that underpin the legislation, laid out in Fig. 1. The principle of *lawfulness, fairness and transparency* states that the controller must have a lawful reason for collecting the personal data and must do it in a fair and transparent way (stating in clear terms to the data subject). The principle of *purpose limitation* declares that personal data can only by used for the reason it was collected. The principle of *data minimisation* enacts that the amount of collected personal data should be kept to the minimum necessary to perform the processing activities. The principle of *accuracy* determines that personal data must be kept up to date and any inaccuracies should be dealt with as soon as possible. The principle of *storage limitation* establishes that personal data shall be kept for no longer than necessary for the purposes for which it is being processed. The principle of *integrity and confidentiality* states that personal data must be protected from loss or tampering (safety) or unauthorized access (security). A further principle of *accountability* implies that the controller should be able to prove that is complying with the six previous principles.

2.3 Rights of the data subject

The GDPR foresees a number of rights of the data subject over his/her personal data. An organisation has to ensure that provides the mechanisms to allow him/her to exercise these rights, described in Articles 12 to 22 and summarized in Fig. 2.

The *right to be informed* means that the data subject has to be clearly told of what is the data to be collected and what we will be done with it while the *right of access* means that the data subject may inquire about details about the data that has been collected about him/her.

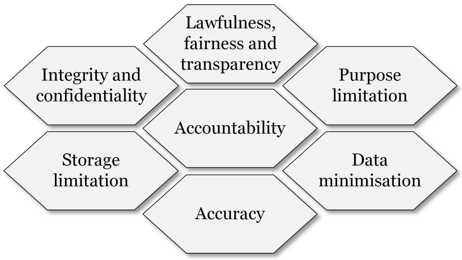

Fig. 1. Privacy principles

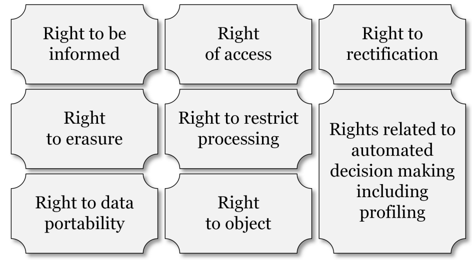

Fig. 2. Rights of the data subject

There are a couple of rights about data quality and maintenance, namely, the *right to rectification* meaning that the data must corrected by the controller as soon as any inaccuracy is discovered and even the *right to erasure* (or right to be forgotten) implying that the data subject may request for the data to be erased (if there are no longer lawful rights to hold it).

The data subject also has rights to control or even stop his/her personal from being processed. The *right to restriction of processing* means that the data subject may restrain the scope of the processing of his/her personal data by a number of reasons (although the controller may still hold it) while the *right to object* means that the data subject may invoke objections to stop the controller from processing his/her personal data altogether.

The *right to data portability* means that the data subject may receive the personal data concerning him/her in a structured, commonly used and machine-readable format. Is also implies that the data subject has the right to transmit it to another controller without hindrance from the previous controller. The *rights related to automated decision making including profiling* means that the data subject must be able to choose if decisions concerning his/her matters are made by human intervention instead of an algorithm.

2.4 Lawful basis of processing personal data (lawfulness)

A pivotal topic to comply with the GDPR is for the controller to ensure that it is processing data rightfully. According to Article 6, for the processing of personal data to be lawful, it must meet at least one of the criteria in Fig. 3. Therefore, it is the responsibility of the controller to establish which of the criteria applies in any given situation.

Consent of the data subject is a lawful basis if he/she has given permission to the processing of personal data for one or more specific purposes. According to Article 4, a consent has to be a freely given, specific, informed and unambiguous indication of the data subject's wishes by a statement or by a clear affirmative action, meaning agreement to the processing of personal data relating to him/her. *Vital interests* is a lawful basis clarified by Recital 46 as processing necessary to protect an interest essential for the life of the data subject or that of another natural person. *Contract* is a lawful basis when processing is necessary for the performance of a contract or pre-contractual arrangements between the data subject and the controller. *Compliance with a legal obligation* is a lawful basis if processing is done to fulfill an obligation under the law of the country in which the controller carries on business. *Public interest* is a lawful basis if processing is necessary for the performance of public functions and powers that are set out in law or to perform a specific task in the public interest that is set out in law. *Legitimate interests of the controller* is a lawful basis if processing is necessary for the purposes of justifiable interest of the controller (as long as it does not affect the data subjects rights and freedoms).

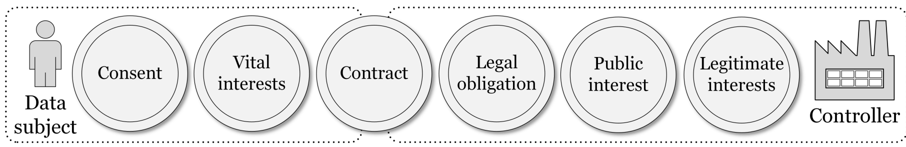

Fig. 3. Lawful basis of processing personal data

Although controllers often request consent from data subjects, most of the time it is unnecessary as a significant proportion of the personal data in organisations processes does not require it. Contractual (such as providing services to customers), legal (such as paying employees or dealing with the tax authority) and legitimate interests (such as collection of fingerprints for access to facilities) are more appropriate lawful basis. In fact, consent should be the last lawful basis to be invoked, because the process of obtaining and maintaining it involves changes to business processes and systems.

2.5 Personal data breaches

The GDPR defines a *personal data breach* in Article 4 as a breach of security leading to the accidental or unlawful destruction, loss, alteration, unauthorised disclosure of, or access to, personal data transmitted, stored or otherwise processed.

According to the *Guidelines on Personal data breach notification under regulation 2016/679* [2] by the Article 29 Working Party (WP29), it should be clear that a breach is a type of security incident. The GDPR only applies where there is a breach of personal data, consequently, while all personal data breaches are security incidents, not all security incidents are necessarily personal data breaches. In its *Opinion 03/2014 report* [3] on breach notification, the WP29 explained that breaches can be categorised as *breach of confidentiality* when there is unauthorized or accidental disclosure or access to personal data; *breach of availability* when there is unauthorized or accidental disclosure loss of access or destruction of personal data; *breach of integrity* when there is unauthorized or accidental disclosure alteration of personal data.

Furthermore, Article 33 states that a personal data breach has to notified to a supervisory authority unless it is unlikely to result in a risk to the rights and freedoms of natural persons. At the other end, when it is likely to result in a high risk to the rights and freedoms of natural persons, the controller must also notify the data subject without undue delay, as pointed out in Article 34.

3 Personal data in interpretable models

Creating models is a scientific activity, that can be carried out to fulfill either research purposes or statistical purposes. The models are created through data analysis which implies the collection of data, assembled in *data sets*. If these contain personal data, then creating models can be considered "processing" as defined by the GDPR.

Usually personal data is laid out in physical documents or stored in operational systems. One example are Enterprise Resource Planning (ERP) software systems which offer integrated management of core business processes, often in real-time. Another are University Information Systems (UIS) in Higher Education Institutions. In both there are databases providing convenient sources of personal data, paving the way to the creation of prediction models, one application of automatic knowledge discovery from databases (or data mining) [4].

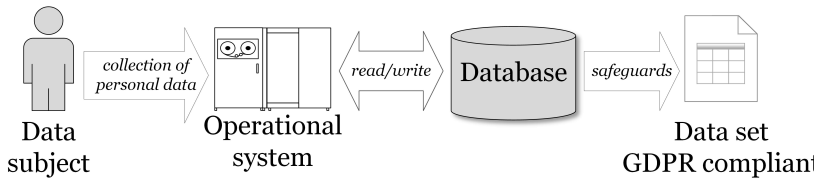

Fig. 4. Collecting personal data to data sets

As depicted in Fig. 4, the process of collecting personal to data sets being by a data subject providing personal data to a controller which, in turn, processes into an operational system and stores it in a database. Managers, however, are able to recognize that a great wealth of information is being accumulated over the years and can be explored for more possibilities than the sheer storage or process support. Therefore, it is possible to create data sets from databases using tailored queries to extract relevant information for specific data analysis. Inevitably, this informations also contains personal data intertwined with data related to the process under analysis.

A challenge is to ensure that the assembled data sets are "GDPR compliant". In others words, the personal data included in them does not reveal who are the data subjects. The GDPR pinpoints that safeguards should be used when collecting data (discussed in subsection 4.1). Firstly, however, it is essential to introduce a few fundamental concepts about interpretable models and how they can expose personal data.

3.1 Interpretable models and related concepts

A *variable* x (also referred to as attribute or feature), takes values from a range of values (either discrete, as a set of values, or continuous, as a set of numbers limited by lower and upper bounds). An *example* (also referred to as an instance or observation), is a fixed ordered set of values of different variables. Examples may refer to entities or occurrences of an event. A *dataset* is an unordered set of examples. The *dimension* of the dataset is the number of variables in contains.

A *training set*, denoted as Tr, is a dataset used to create a prediction model. For this dataset it is useful to distinguish between a set of *independent variables*, denoted as X (eq. 1), and the *target* variable (y) to be predicted. The full set of variables (V) includes both (eq. 2).

$$X = \{x_1, x_2, \ldots\} \quad (1) \qquad V = X \cup \{y\} \quad (2)$$

A *prediction model*, denoted as M, is a function that maps a set of independent variables to a target variable (eq. 3). The prediction made by the model for the target variable when, given an example, is denoted as $\hat{y}$. A learning algorithm, is used to create a prediction model. The operation is referred to as *TrainModel* because it learns a model from a training set. It is necessary to specify the learning algorithm $\mathcal{L}$, the training set *Tr*, the set of independent variables X and the target variable y (eq. 4).

$$\hat{y} = M(X) \quad (3) \qquad M = \mathit{TrainModel}(\mathcal{L}, \mathit{Tr}, X, y) \quad (4)$$

In knowledge discovery from databases, models are created to predict future events or to understand the relationship between variables involved in a phenomena being studied. In the latter case, the characteristic of *interpretability* is essential, i.e., the models must be understood by humans. These are called *white-box models* which means that it is possible to navigate through a model to follow the reasoning behind a prediction, as illustrated in Fig. 5. The opposite are *black-box models* where the details of how the model makes predictions are either hidden or not readable by a human. Although standing very well in prediction quality, they do not have the characteristic of interpretability, i.e., it is not possible to "look inside the box" and understand how a prediction is made, as Fig. 6 suggests.

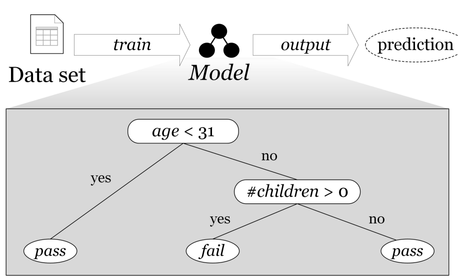

Fig. 5. White-box model

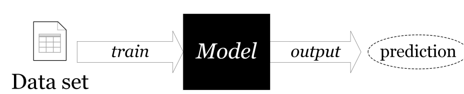

Fig. 6. Black-box model

Decision trees [5] is a common example of an algorithm that induces white-box models by creating tree-like structures where the independent variables are tested in nodes and the leaves hold the values of the target variable. Although there are several algorithms to create decision trees, the most popular are *Classification and Regression Trees* (CART) [6] and *C5.0* [7]. Algorithms for constructing decision trees usually work top-down, by choosing a variable at each step that best splits the data set [9]. Different algorithms use different metrics for measuring what is considered "best". These generally measure the homogeneity of the target variable within the subsets. Some examples are *Gini impurity*, *Information gain* and *Variance reduction*. One of these metrics is applied to each candidate subset, and the resulting values are combined to provide a measure of the quality of the split. In the example of Fig. 5, of all the independent variables in the data set, *age* and *#children* were selected to be splitting variables according to some metric. Correspondingly, 31 and 0 are split points, i.e., the values where a split in the tree navigation path takes place. Each path from the root node to the leaves defines a *decision rule*.

3.2 Exposing personal data in interpretable models

The decision tree in Fig. 5 relates to a course in an university where the goal is to predict if a student is going to pass or fail in it [8]. It is straightforward to grasp that the model first evaluates the age of the student. If it is under 31 then the model predicts that he/she passes, otherwise, it evaluates the number of children. If the student has children (regardless of how many), the model predicts that the student fails, alternatively he/she passes. Nonetheless, a word of caution about models is necessary. The predictions are based on the data used to create it, i.e., they describe the general behavior in a group of students taking a specific course in an academic year. They do not describe what happens with any student taking a course in the university.

Decision trees expose personal data in the split points. By selecting a particular value for a variable as a split point, the model reveals that there are examples in the data set with that value. In the example above, the *fail* rule ($age \geq 31 \wedge \#children > 0$) reveals that there is at least one student in these circumstances. This is a merely illustrative example of the problem as it is not foreseeable that disclosing the fact that a person's has children could put him or her in any way at risk. However, there may be other variables with the potential to do so.

There is no straightforward solution to the dilemma. On the one hand, including variables describing personal data are necessary because of their promising capability as explanatory variables relating a phenomenon of interest. On the other, the selection of some split points can unintentionally expose private information about the data subjects. Obviously, in the previous example, one can remove the variable *#children*. Still, that may not be the case in other circumstances where the inclusion of such variable may be instrumental to the quality of the model.

4 Towards creating models GDPR compliant

With the entry into force of the GDPR, it became mandatory to find measures that can reduce the risk of exposing information connected to personal data. In the context of creating interpretable models, these translate into taking particular attention in extracting variables from databases to data sets and also on the data pre-processing tasks.

Article 32, about security of processing, declares that taking into account the state of the art, the costs of implementation and the nature, scope, context and purposes of processing as well as the risk of varying likelihood and severity for the rights and freedoms of natural persons, the controller and the processor shall implement appropriate technical and organisational measures to ensure a level of security appropriate to the risk. These measures are repeatedly referred to in the GDPR as *safeguards* with pseudonymisation and encryption of personal data as common examples.

There are, however, further safeguards that can be addressed in data sets towards GDPR compliance. A few more are presented next and are then discussed in the context of the privacy principles more closely related to them. Ultimately, even with a complete set of measures, a visual inspection by a human being will always be necessary to make sure that, due to unforeseen set of circumstances, there is no risk of a data subject being identified in a model.

4.1 Safeguards in the data sets

The following safeguards should be considered when preparing a data set for the creation of an interpretable model. Certainly not all of them will always be necessary and it depends on each problem at hand to decide which ones make sense to be applied.

Identifiers removal. Although it may look rather clear, it is important to stress out that any variables that uniquely identify a person have to be removed from the data sets. In fact, their inclusion does not even make sense as they have no role in a prediction model. Since the goal is to find patterns in the data and extract rules, a variable in which every value only occur in a single example can never qualify as explanatory. Examples are a *citizen's card number*, *tax ID number*, or *social security number*.

Replacement of variables. There are variables related to personal data that can be replaced by others (derived variables) to make it harder to trace to a data subject. The first kind is by calculating a new variable from existing data, e.g., replacing a *date of birth* by a person's *age*. The second kind is creating a categorical variable from a continuous variable. In this case, instead of the person's *age*, the value would be to an *age range*.

Decrease of granularity. One way to conceal personal data is to diminish the information's level of detail. An example is to replace an *address* by a *parish*, a *county* or a *district*. Another is to replace a *job title* by a *career*. In a university context, replacing the *course* by the *scientific area* it relates to.

Dimensionality reduction. Encompasses all processes for reducing the number of independent variables of a data set to obtain a set of *principal variables*, i.e., those that can actually be used by models. In the context of interpretable models a recommendation is to use *feature selection* techniques which, besides removing redundant variables, have the potential of simplifying the models and improve generalization by reducing overfitting (an analysis that corresponds too closely or exactly to a particular set of data, and may therefore fail to fit additional data or predict future observations reliably).

Data volume reduction. Consists on efforts for restricting the volume of a dataset into less training examples. One method is *data compression* in which an aggregation function or another more sophisticated method to downsize the dataset and train a model, described by Yael [10]. Another is *instance selection* in which examples are selected from a dataset if these are considered sufficient to train a model representative of what would be a model trained with all available data. These have been proposed by Liu [11] and Blum [12] as a strategy to deal with computationally intensive algorithms. The datasets are downsized so that learning is focused on a chosen set of *informative examples*.

Data anonymisation. Defined in Recital 26 as the process of removing personal identifiers, both direct and indirect, that may lead to an individual being identified. An individual may be directly identified from their name, address, postcode, telephone number, photograph or image, or some other unique personal characteristic. An individual may be indirectly identifiable when certain information is linked together with other sources of information, including, their place of work, job title, salary, their postcode or even the fact that they have a particular diagnosis or condition. Once data is truly anonymised and individuals are no longer identifiable, the data will not fall within the scope of the GDPR and it becomes easier to use.

Data pseudonymisation. Defined in Article 4 as the processing of personal data in such a way that the data can no longer be attributed to a specific data subject without the use of additional information, as long as such additional information is kept separately and subject to technical and organizational measures to ensure non-attribution to an identified or identifiable individual. consists on a de-identification procedure by which personally identifiable information fields within a data record are replaced by one or more artificial identifiers, or pseudonyms. A single pseudonym for each replaced field or collection of replaced fields makes the data record less identifiable while remaining suitable for data analysis and data processing.

Data encryption. It is a process of cryptography which consists in encoding a message or information in such a way that only authorized parties can access it and those who are not authorized cannot. Encryption does not itself prevent interference, but denies the intelligible content to a would-be interceptor, therefore, is a security measure in interpretable models.

4.2 Compliance with the privacy principles

One way to prove compliance with the GDPR and its underlying demands is to align the processing operations with the privacy principles described in Article 5 (introduced in subsection 2.2). This section revisits a few of those principles in the light of the creation of interpretable models, in particular by discussing what needs to be done to foster conformity.

Principle of lawfulness, fairness and transparency. Fairness and transparency are directly linked with the right to be informed and the right of access. Therefore, it has to be traced to the time when the data was first collected for a specific purpose.

Lawfulness has to be guaranteed under a valid lawful basis. The controller may resort to legitimate interests as long as it ensures that the model does not allow the identification of any data subject. Consent, although remaining an alternative, may be infeasible due to frequently large number of examples that a data set needs to include in order to create reliable models. This is reinforced by Article 6, stating that the processing for a purpose other than that for which the personal data have been collected is not based on the data subject's consent. The controller shall, in order to ascertain whether processing for another purpose is compatible with the purpose for which the personal data are initially collected take into account, among others the existence of appropriate safeguards, which may include encryption or pseudonymisation.

Additionally, according to Recital 50, the processing of personal data for purposes other than those for which the personal data were initially collected should be allowed only where the processing is compatible with the purposes for which the personal data were initially collected. In such a case, no legal basis separate from that which allowed the collection of the personal data is required.

Principle of purpose of limitation. As creating models can be framed in the context of research or statistical purposes, it fits into the directives of Recital 156 which states that the processing of personal data for archiving purposes in the public interest, scientific or historical research purposes or statistical purposes should be subject to appropriate safeguards for the rights and freedoms of the data subject. Processing may be carried out when the controller has assessed the feasibility to fulfill those purposes by processing data which do not permit or no longer permit the identification of data subjects, provided that appropriate safeguards exist.

Principle of data minimization. Keeping collected data to the strict minimum necessary must be looked upon both in the perspective of the number of variables used (dimension of the data set) and the number of examples (volume of the data set).

Applying dimensionality reduction may cause the variables involving personal data to be rejected. However, there must be avoided any statistical procedures that change the variables themselves (also known as *feature projection*). An example is Principal Component Analysis (PCA) that uses an orthogonal transformation to convert a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables (the principal components). If this is done, the pivotal characteristic of interpretability of the models is lost, because the variables become devoided of their business meaning.

Data reduction techniques, although motivated by the problem of training models from very large datasets, hold the potential that a model can be obtained using less data, which is aligned with the principle of data minimization. Therefore, their deployment may help to decrease the volume of the data set necessary to train a model.

Principle of storage limitation. It is important to acknowledge that the GDPR does not set specific time limits for different types of data, but it makes organisations responsible for determining on how long they need to hold the data for their specified purposes. Recital 39 decodes, in a clear and concise manner, that the limits imposed by Article 5, by stating that the period for which the personal data is stored should be limited to a strict minimum and that time limits should be established by the data controller for deletion of the records (referred to as "erasure" in the GDPR) or for a periodic review.

Article 5 identifies the circumstances when personal data may be kept for longer than necessary for the purposes for which it is being processed as "personal data may be stored for longer periods insofar as the personal data will be processed solely for archiving purposes in the public interest, scientific or historical research purposes or statistical purposes in accordance with Article 89 subject to implementation of the appropriate technical and organisational measures required by this Regulation in order to safeguard the rights and freedoms of the data subject ('storage limitation').". These measures could include the deletion of the data sets after creating the models, while keeping the latter.

4.3 Ensuring the rights of the data subjects

A model can be used in a decision support tool for automated decision making. The right to object and the rights related to automated decision making including profiling demand special attention concerning this situation.

It is necessary to ensure that the data subjects who have invoked these rights are no longer included in the data sets used to create models. Thus, firstly it is necessary to scan the existing data sets in order to remove data subjects and then re-create the models. Next, in creating new data sets, it is mandatory to include a new condition in the extraction queries specifically to reject the data subjects who have given this indication.

5 Conclusions

Organizations have been preparing for the requirements of the GPDR, which has led to changes in processes and the introduction of measures as safeguards. The main consequence of non-compliance is the possibility of data breaches being punished with heavy fines. Creating interpretable models is a processing operation which, under certain circumstances, may inadvertently disclose personal data leading to the identification of a data subject. It is necessary for organizations pursuing such activities (e.g., including models in decision support systems) to take steps to prevent this from happening.

It is imperative to act both on the way personal information is extracted from databases and also the one already existing in data sets and physical supports. Also the rights of data subjects have to be fulfilled, by removing them permanently from those very data sets. Certainly, not a simple task, but at this point, organizations have no choice but to revisit all their processing operations involving personal data and making sure lawfulness is guaranteed.

References

1. European Union. Regulation 2016/679. *Official Journal of the European Communities*, 2014(March 2014):1–88, 2016.
2. Article 29 Data Protection Working Party. Guidelines on Personal data breach notification under regulation 2016/679. Technical Report October, 2017.
3. Article 29 Data Protection Working Party. Opinion 03/2014 on Personal Data Breach Notification. Technical Report March, 2014.
4. J. Han, M. Kamber, and J. Pei. *Data Mining: Concepts and Techniques.* Morgan Kaufmann, San Francisco, 2011.
5. J. Quinlan. Induction of Decision Trees. *Machine Learning*, 1(1):81–106, 1986.
6. L. Breiman, J. Friedman, R. Olshen, and C. Stone. *Classification and Regression Trees.* Chapman and Hall/CRC, 1984.
7. M. Kuhn, S. Weston, N. Coulter, and J. Quinlan. C50: C5.0 Decision Trees and Rule-Based Models. R package version 0.1.0-16, 2014.
8. P. Strecht, J. Mendes-Moreira, and C. Soares. Merging Decision Trees: A Case Study in Predicting Student Performance. In X. Luo, J. Yu, and Z. Li, editors, *Advanced Data Mining and Applications*, Lecture Notes in Computer Science, pages 535–548. Springer International Publishing, 2014.
9. Lior Rokach and Oded Maimon. Top-Down Induction of Decision Trees: A Survey. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 35(4):476–487, 2005.
10. B. Yael and T. Elad. A Streaming Parallel Decision Tree Algorithm. *Journal of Machine Learning Research*, 11:849–872, 2010.
11. H. Liu and H. Motoda. On Issues of Instance Selection. *Data Mining and Knowledge Discovery*, 6(2):115–130, 2002.
12. A. Blum and P. Langley. Selection of Relevant Features and Examples in Machine Learning. *Artificial Intelligence*, 97(1-2):245–271, 1997.