WHITE PAPER

# A Data Protection Grammar

## Navigating Regulatory Language for Compliance in a Multi-Compute Environment

2021

**Sophie Stalla-Bourdillon**
Senior Privacy Counsel and
Legal Engineer, Immuta

**Spurthi Jonnalagadda**
Immuta Scholar

IMMUTA™

# Table of Contents

# Introduction

Data protection law emerged in the 1970's in Europe as a means to protect individuals and societies from the risks posed by automated data processing or computer-based processing. Data protection law as a concept thus goes far beyond protecting individuals against the disclosure of nonpublic information, a concern that is still very much at the center of modern US privacy laws such as the California Consumer Privacy Act (CCPA) or its second iteration, the California Privacy Rights Act (CPRA).[1]

Fifty years after the first resolutions of the Council of Europe[2], which have been leading the harmonizing effort at the global level[3], the risks posed by automated data processing or algorithmic data processing are more acute than ever. Suffice it to look at what happened in the last 10 years to find a long list of alarming stories of surveillance, manipulation, and discrimination or malfunction already.[4] There is thus the need to make data protection as robust as possible and give organizations of all sizes the means to effectively implement it on the ground. Notably, privacy laws around the world are progressively being informed by the data protection approach and are evolving. Yet, as time passes, the patchwork of data protection and privacy laws is becoming more complex. There are dozens of new and existing regulations across the world – and each regulation uses different terminology.

Analyzing the structure of European data protection law, with the European Union General Data Protection Regulation (GDPR)[5] as its most illustrative example, and comparing it with the structure underlying modern data protection or privacy laws adopted in other jurisdictions such as CPRA, the Canadian

1 CAL. CIV. CODE § 1798.100 (f)(f) (West 2021.)

2 Council of Europe, Committee of Ministers, Resolution (73) 22 on the Protection of the Privacy of Individuals vis-a-vis Electronic Data Banks in the Private Sector, https://rm.coe.int/1680502830; Council of Europe, Committee of Ministers, Resolution (74) 29 on the Protection of the Privacy of Individuals vis-a-vis Electronic Data banks in the Public Sector, https://rm.coe.int/16804d1c51.

3 Convention for the Protection of Individuals with regard to Automatic Processing of Personal Data, ETS No.108; Modernised Convention for the Protection of Individuals with Regard to the Processing of Personal Data CM/Inf(2018)15-final.

4 *See* e.g., Richard Adams & Heather Stewart, Boris Johnson Urged to Intervene as Exam Results Anger Escalates, GUARDIAN, (Aug. 16, 2020), https://www.theguardian.com/education/2020/aug/16/boris-johnson-urged-to-intervene-as-exam-results-crisis-grow; Julia Angwin et al., Machine Bias, PROPUBLICA (May 23, 2016), https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing; Bamzi Banchiri, Is Amazon Same-Day Delivery Service Racist?, CHRISTIAN SCIENCE MONITOR (Apr. 23, 2016), https://www.csmonitor.com/Business/2016/0423/Is-Amazon-same-day-delivery-service-racist; Carole Cadwalladr & Emma Graham-Harrison, Revealed: 50 million Facebook Profiles Harvested for Cambridge Analytica in Major Data Breach, GUARDIAN, (Mar. 17, 2018), https://www.theguardian.com/news/2018/mar/17/cambridge-analytica-facebook-influence-us-election; Silkie Carlo, Jennifer Krueckeberg & Griff Ferris, Face Off: The Lawless Growth of Facial Recognition in UK Policing. BIG BROTHER WATCH (May 2018), https://bigbrotherwatch.org.uk/wp-content/uploads/2018/05/Face-Off-final-digital-1.pdf; Jayson DeMers, How Much Do We Really Know About Google's Ranking Algorithm?, MEDIUM (May 28, 2020), https://medium.com/swlh/how-much-do-we-really-know-about-googles-ranking-algorithm-ef031586681b; Glenn Greenwald, Ewen MacAskill & Laura Poitras, Edward Snowden: The Whistleblower Behind the NSA Surveillance Revelations, GUARDIAN (June 11, 2013), https://www.theguardian.com/world/2013/jun/09/edward-snowden-nsa-whistleblower-surveillance; Colin Lecher, What Happens When an Algorithm Cuts Your Health Care, VERGE (Mar. 21, 2018), https://www.theverge.com/2018/3/21/17144260/healthcare-medicaid-algorithm-arkansasce-rebral-palsy; Timothy Revell, Face-Recognition Software Is Perfect- If You're a White Man, NEWSCIENTIST (Feb. 13, 2018), https://www.newscientist.com/article/2161028-face-recognition-software-is-perfect-if-youre-a-white-man/. For an overview of automated decision making initiatives and their impact in the EU, see Automating Society Report, ALGORITHM WATCH (2020), https://automatingsociety.algorithmwatch.org/report2020/policy-recommendations/.

5 Council Regulation 2016/679 of the European Parliament and of the Council of 27 April 2016 Relating to the Protection of Natural Persons with Regard to the Processing of Personal Data and on the Free Movement of Such Data, and Repealing Directive 95/46/EC (General Data Protection Regulation), 2016 O.J. (L 119) 1–88.

Personal Information Protection and Electronic Documents Act (PIPEDA),[6] the Brazilian General Personal Data Protection Law (GPDPL)[7] and others, this white paper suggests that it is possible to extract a data protection grammar, which both assesses the similarities and dissimilarities among these different frameworks, and reveals a common denominator at a meta-level around which data teams should start to organize themselves.[8]

The basic postulate of such a grammar is that both core structural rules (i.e., a data protection syntax) and a concise set of lexical items are embedded within this new generation of data protection and privacy laws. This data protection grammar proves particularly useful in comparing the scope and effects of these frameworks, and assessing the capabilities of emerging policy layers built to govern multi-cloud data analytics platforms and meet geographic demands. Layers within data analytics environments make it possible to organize components and separate functions, such as computation and storage, to solve scalability challenges. Policy layers offer a way to author and enforce access and usage-related rules upon data.

This white paper aims to begin defining a data protection grammar, focusing upon key syntactic rules in section 2 and recapping the main takeaways for data analytics platform owners in section 3. The first version of a data protection lexicon is included in Appendix I. The ultimate goal is to help organizations navigate the complex data protection and privacy landscape, and identify the core building blocks of a data analytics platform, thereby contributing to safer data operations.

# The data protection syntax

As globalization accelerates, legal rules, institutions, and concepts move across jurisdictions or help shape, influence, or simply inform the rules, institutions, and concepts of other jurisdictions.[9] In the data protection and privacy world, these interactions are getting more complex day by day, without necessarily impacting the overall functioning of the receiving or observing jurisdiction's legal system. Analyzing a handful of these data protection and privacy laws, it is possible to extract a meta-structure comprising five core rules, i.e. a data protection syntax. To be clear, while this meta-structure is useful for comparison purposes, it does not imply that similar problems get similar answers across jurisdictions. One key structural difference is that in some jurisdictions, such as the European Union, the acceptability of a data processing practice is dependent upon its processing impact, as defined under Rule 4.

Let's take a closer look at these five core rules.

6 Personal Information Protection and Electronic Documents Act, S.C. 2000, c.5 (Can.). PIPEDA fully came into effect in 2004. Draft legislation is in the pipeline: the Digital Charter Implementation Act, 2020 is expected to transform PIPEDA into the Consumer Privacy Protection Act. See Digital Charter Implementation Act, 43d Parliament, 2d Sess. (2020), https://www.parl.ca/LegisInfo/BillDetails.aspx?Language=e&Mode=1&billId=10950130.

7 Decreto No. 13,709, de 14 de agosto de 2018, DIARIO OFICIAL DA UNIAO [D.O.U] de 3.9.2020 (Braz.). GPDPL entered into force in September 2020 in Brazil.

8 We built upon the work of Sylvain Auroux in linguistics who defines grammatisation as the process by which the language progressively equipped itself through syntax and lexicon, which become external linguistic instruments and contribute to the standardization of the language itself. See Auroux Sylvain. Grammatisation, 11 Histoire Épistémologie Langage 5, (1995), https://www.persee.fr/doc/hel_0247-8897_1995_num_11_1_3396. Our ambition is therefore to contribute to the standardization of the data protection terminology and bridge the gap between the terminology used by lawyers and the terminology used by privacy and data protection technologists or experts in de-identification statistical methods.While the ISO 25237:2017 Health informatics – Pseudonymization standard must certainly be welcome, it does not account for differences of approach between jurisdictions.

9 *See* generally, ALAN WATSON, LEGAL TRANSPLANTS: AN APPROACH TO COMPARATIVE LAW (Edinburgh, Scottish Academic Press 1974).

# Rule 1: Identifiability attracts protection

Data is legally protected if a link can be established between the data and an individual. This is not exactly the same thing as saying that only identifying items of data are protected, as we will explain below.

Identifiability is therefore the capability of linking data to an individual. Most frameworks adopt a relativistic approach to identifiability and inject a standard of reasonableness into its definition, so that data is legally protected if a link can reasonably be established between the data and an individual. As a result, the link between the data and the individual can potentially be broken through a process of de-identification.

Automated data processing implies that the data is organized in a certain format. In practice within data analytics environments, this often means that data with a fixed set of attributes is organized into tables where the rows correspond to records and attributes are organized along columns. Let's take one typical table containing consumer data:

| CUSTOMER ID | NAME | CREDIT CARD NUMBER | ADDRESS | AGE | TEMP | BLOOD PRESSURE | DATE |
|---|---|---|---|---|---|---|---|
| t.t.c.99@gmail.com | Thomas T C. | 5159162191795281 | 3462 Oakwood Avenue | 22 | 98.6°F | 120.5/78.5 | August 3, 2019 |
| c.123@aol.com | Carmen E H. | 5565898335470533 | 706 Union Street | 46 | 97.9°F | 124/78.5. | October 16, 2020 |

Table 1. Health data

These attributes can be classified as personal identifiers (direct identifiers, indirect identifiers) and other individual attributes.

*Personal identifiers* are attribute values that can be used to discriminate among individuals (i.e., can be used to help locate an individual's records or single them out) and are considered to be available to an attacker. An attribute is "available to an attacker" when it is publicly available, observable, or attainable. The two main characteristics of personal identifiers are thus distinguishability and availability.

*Direct personal identifiers* are attribute values that are unique to an individual and are considered to be available to an attacker (such as Social Security number, passport or ID number, or credit card number). In our example, direct personal identifiers are: customer ID, credit card number, and name.

*Indirect personal identifiers* are attribute values (such as height, ethnicity, hair color, etc.) that are not unique to an individual, but can be used in combination with other attributes to distinguish an individual and are available to an attacker. In our example, indirect personal identifiers are: address, gender, age, and date.

*Personal information* covers both personal identifiers and other attribute values that are associated with personal identifiers. These other attributes should be considered conditional personal information. Conditional personal information refers to personal attribute values that are not distinguishable and/or available to an attacker. They are conditional in the sense that if personal identifiers are transformed so that the link between the data and the individual is considered to be broken, the attribute values cease to be personal information. In our example, temperature and blood pressure can be considered

conditional personal information when it is safe to assume that they are not available to an attacker. As we explain under Rule 2, 'safe' usually means 'reasonable.'

Another category of personal information that must be mentioned is sensitive information. Almost all regulations carve out classes of information as being particularly sensitive and thereby warranting extra protection. *Sensitive personal information* can be understood as personal information, the disclosure or misuse of which is considered particularly harmful to individuals.

The line between the categories of indirect personal identifiers and other attribute values linked to an individual (i.e. conditional personal information) is a fluid one which continues to evolve over time. What distinguishes indirect personal identifiers from other attribute values associated with an individual are their inherent distinguishability power and/or availability.[10] Based on experience and reasoning, it is possible to make reasonable assumptions about what kinds of information can be expected to be available to an attacker, and therefore which attributes may be considered identifiers.

Assume now that we have a table with the following columns:

| ADDRESS | WATER CONSUMPTION | PERIOD | PAYMENT AMOUNT |
|---|---|---|---|
| 3462 Oakwood Avenue | 2,550 gallons | March 1 – March 31, 2020 | $45 |
| 706 Union Street | 10,200 gallons | March 1 – March 31, 2020 | $171.50 |

Table 2. Water consumption data

The columns, or attributes, are linked to a household singled out through an address, which is a group identifier. Group identifiers are attribute values that can be used to discriminate among named groups, such as households. As such, they are also a type of indirect personal identifiers.

The period attribute could also be considered a group identifier and an indirect personal identifier, while water consumption and payment amount can be considered conditional personal information if it is safe to assume that they are not available to an attacker.

*Inferences* are attribute values which can be confidently guessed or estimated through analysis when considering attribute values alone within a data source or in combination with information outside the data source. These can also be considered personal information (personal identifiers or conditional personal information), and as such must be legally protected. This includes inferences for attributes not represented in the data.

10 Some experts add a third characteristic to identifiers, that of replicability. They consider that attributes that are not replicable, i.e., that are not consistent with individuals, are not identifiers. We consider however that non-replicability does not necessarily mean that the attribute is not an identifier. Non-replicable attributes could be made available and act as indirect identifiers. However, as a matter of fact, attributes that are replicable are more likely to be available than attributes that aren't.

## GDPR

GDPR (Article 4) defines personal data as

> *"Any information relating to an identified or identifiable natural person ('data subject') meaning someone who can be identified, directly or indirectly, in particular by reference to an identifier."*

While GDPR does not expressly target inferences, inferences that amount to personal data fall within the remit of the framework.[11]

Under GDPR Article 4, the entire row of Table 1 would be considered to be personal data, including both personal identifiers and conditional personal information. The same is true with Table 2, as the address attribute is an indirect personal identifier.

## CPRA

CPRA (section 1798.140(v)(1)) defines personal information as

> *"Information that identifies, relates to, describes, is reasonably capable of being associated with, or could reasonably be linked, directly or indirectly, with a particular consumer or household."*

CPRA expressly includes inferences within the meaning of personal information, as per section 1798.140(v)(1)(K). While the CPRA definition of inferences is essentially concerned with inferred personal indirect identifiers and conditional personal information, inferred personal direct identifiers should obviously also be considered personal information.[12]

Under CPRA section 1798.140(v)(1), the entire row of Table 1 would be considered personal information including both personal identifiers and conditional personal information. The same is true with Table 2, as the attribute 'address' is a group identifier usually associated with a household.

Of note, although these examples show how the meaning of "personal information" under CPRA overlaps with the meaning of "personal data" under GDPR and brings the two frameworks closer, there are also obvious differences that keep them apart. In particular, linguistic differences matter when regulations adopt exclusions or exceptions, or when regulations adopt more restrictive rules for sensitive personal information. CPRA, for example, excludes publicly available information;[13] GDPR does not. Both CPRA and GDPR have special rules for sensitive personal information, although these categories only partially overlap. Driver's licenses and state identification cards or passport numbers, for example, are categorized as sensitive information under CPRA, but not under GDPR.

11 Commentators have discussed the extent to which data subjects are able to exercise their rights vis-à-vis inferences. See e.g., Sandra Wachter & Brent Mittelstadt, A Right to Reasonable Inferences: Re-Thinking Data Protection Law in the Age of Big Data and AI, COLUM. BUS. L. REV. 494 (2019). However, a distinction is not always drawn between the reasoning or analysis used to derive the attribute value and the attribute value itself. Article 29 Data Protection Working Party opines that inferences are less protected than data provided by the data subject. See Article 29 Data Protection Working Party, Guidelines on the Right to Data Portability, WP242rev.01 (Apr. 5, 2017), https://ec.europa.eu/newsroom/article29/item-detail.cfm?item_id=611233 ("Even though such data may be part of a profile kept by a data controller and are inferred or derived from the analysis of data provided by the data subject (through his actions for example), these data will typically not be considered as 'provided by the data subject' and thus will not be within scope of this new right. . . Nevertheless, the data subject can still use his or her 'right to obtain from the controller confirmation as to whether or not personal data concerning him or her are being processed, and, where that is the case, access to the personal data' as well as information about 'the existence of automated decision-making, including profiling, referred to in Article 22(1) and (4) and, at least in those cases, meaningful information about the logic involved, as well as the significance and the envisaged consequences of such processing for the data subject,' according to Article 15 of the GDPR (which refers to the right of access).").

12 Inferences are the object of consumer rights to the extent they are collected, sold or shared.

13 This does not mean that publicly available information should not be considered to determine whether the information at hand should be characterised personal information within the meaning of section 1798.140(v).

## PIPEDA

PIPEDA defines personal information as information about an identifiable individual.[14] Under this very broad definition, personal information would consist of all personal attributes, including conditional personal information. There is no reason to exclude inferences from that definition. The entire row of Table 1 would thus be considered personal data under PIPEDA.

# Rule 2: De-identification weakens the legal protection

The strength of the link between the data and the individual can have a significant impact upon the material scope of the framework.

On the left hand side of the identifiability spectrum, as illustrated in Diagram 1, is the absolutist approach: All data that can be linked to an individual, either directly or indirectly, is characterized as personal information. On the right hand side is the relativist approach: All data that can reasonably be linked to an individual, either directly or indirectly, is characterized as personal information. De-identification is the process by which the link between the data and the individual is considered to be broken.

In practice, this normally means implementing a process by which personal identifiers are made undistinguishable and/or unavailable to an attacker. De-identification thus requires taking into account the means available to any situationally-relevant potential attacker, to determine whether the links between the data and the individual can be considered broken. When the data has undergone a successful process of de-identification, its use and disclosure are made easier primarily because individuals can no longer intervene in the processing (see Rule 5).


Diagram 1: The identifiability spectrum

Importantly, when the data is considered de-identified, data custodians and/or data recipients are not necessarily relieved of all obligations. In many cases, if the data remains within a closed environment (i.e, the data is not made publicly available) data recipients will be subject to a series of process firewalls and obligations, including an obligation to not re-identify individuals and to comply with a breach mitigation plan. Hence, de-identification does not necessarily mean that the data user has now complete freedom. Thus, de-identification only weakens the intensity of the legal protection rather than eliminating it: individuals lose the ability to exercise their rights over their data, but entities handling the data can still be subject to some legal obligations.

14 PIPEDA, S.C. 2000, c.5 2(1) (Can.).

In practice, attack models are used to assess the reasonableness of the link established between the data and an individual.[15] An *attack model* is a collection of assumptions and constraints on the data environment, and/or the behavior and background knowledge of a potential attacker.[16] The definition of a specific attack model directly impacts the availability of the data to the hypothetical attacker it posits, and, ultimately, the characterization of personal identifiers. If no personal identifiers are considered to be present within the data source, then the link between the data and the individual is broken.

It is important to note that detecting identifiers within a data source requires acknowledging information that is not included within the data set, i.e., publicly available or observable information, or more generally, any information that would be available to an attacker in a given model, including possible prior knowledge of the attacker. *Pseudonymization* is less demanding than de-identification, as it does not require acknowledging information that is not included within the data set to determine whether the data can be attributed to an individual. As a consequence, pseudonymization is not concerned with the treatment of indirect identifiers, as indirect identifiers are only identifying to the extent there is access to information that is not included within the data set. Of note, the ISO 25237:2017 standard acknowledges at paragraph 5.3.4 that "pseudonymization generally [is] used against direct identifiers."

One key difference between data protection frameworks lies in the way the reasonableness standard affecting the strength of the link between the data and the individual is interpreted. Anonymizing data (or *anonymization*) implies looking beyond the means considered to be reasonably available to the anticipated data recipient and considering the means of all situationally-relevant potential attackers. Of note, anonymization is sometimes described as an irreversible process, as opposed to pseudonymization, which can be reversible.[17] The absolute nature of the term "irreversible" without further qualification of the attacker, however, implies that the data must be held safe from all possible attackers, regardless of their ability or means, including their computational ability. Yet, it is often unreasonable to assume that attackers are in a position to readily brute force state-of-the-art randomizers or encryption. Taking into account the ability of attackers brings us back to assessing the means of all situationally-relevant potential attackers. Diagram 2 offers a visual representation of pseudonymization, de-identification, and anonymization.

The toolset to de-identify data is relatively well established within data analytics environments and relies upon a variety of controls. *Controls* are organizational, legal, or technical measures put in place to reduce re-identification risks.

15 *See* e.g., Catherine Marsh et al., The Case for Samples of Anonymized Records from the 1991 Census, 154 J. ROYAL STAT. SOC'Y 305 (1991); Fida Kamal Dankar & Khaled El Emam, A Method for Evaluating Marketer Re-identification Risk, 2010 EDBT Proceedings of the 1st International Workshop on Data Semantics 1, https://dl.acm.org/doi/10.1145/1754239.1754271.

16 Under the Prosecutor Attack Model for example, it is assumed that a third party, known as the attacker, targeting a specific individual, wants to locate this individual's record within a data set using publicly or otherwise reasonably attainable information. This model assumes the attacker knows the complete set of publicly or otherwise reasonably attainable information about their target, including that which is realistically attainable but may or may not be plausibly readily available. For instance, information that may only be reasonably obtained by surveillance of the target. Under the Journalist Attack Model, the attacker does not know whether or not her target has a record in the data, i.e., she does not know the original data. See Dankar & Emam, supra note 15; Marsh, supra note 15.

17 *See* e.g., definitions 3.2 (anonymization) and 3.30 (irreversibility) in ISO 25237:2017 Health informatics – Pseudonymization.

Data controls affect the visibility of the data and include the familiar techniques of tokenization,[18] k-anonymization,[19] and local and global differential privacy.[20] Context controls, on the other hand, affect the data's environment and include access controls and user segmentation, contracts, training, monitoring, and auditing.

Combining data and context controls is an effective way to significantly reduce re-identification risks while preserving some level of utility. In practice, this requires defining de-identification domains, a set of implementation rules that define the conditions under which the data can be processed in a de-identified state. This set of rules is usually purpose-specific, particularly when the data is intended to remain within a closed environment.

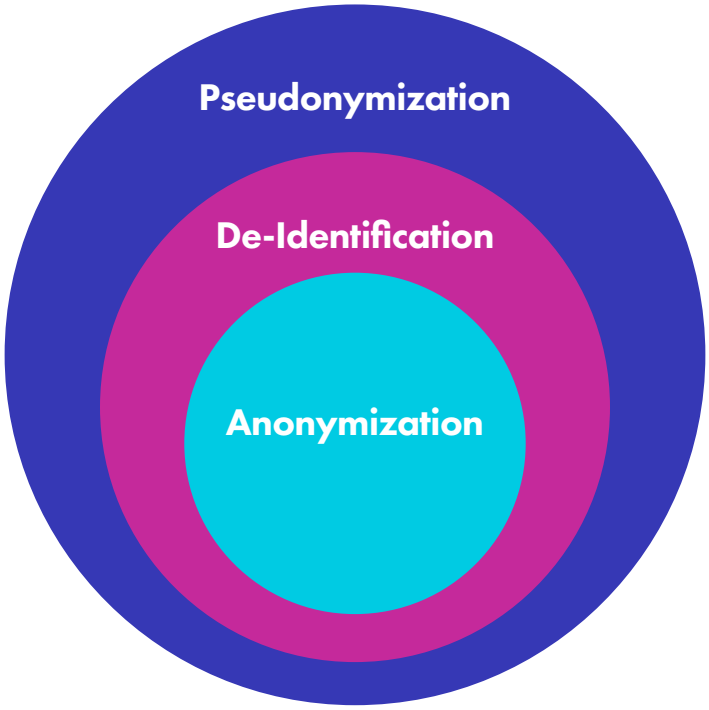


Diagram 2: Comparing pseudonymization, de-identification, and anonymization

18 Tokenization is a specific form of masking where the replacement value, also called "token," has no extrinsic meaning to an attacker. The token is a new value that is meaningless in other contexts. Further, it is not feasible for an attacker to make inferences about the original data from analysis of the token value.

19 k-Anonymization, which is a data generalization technique that ensures indirect identifiers match a specific number of other records, making it difficult to identify individuals within a data set (the total number of matching records is referred to as "k," and hence the name). For example, in data that's been k-anonymized, if k is set to 10 and where indirect identifiers include race and age, we would only see at least 10 records for each combination of race and age. The higher we set k, the harder it will be to use indirect identifiers to find the record of any specific individual.

20 Differential privacy, which is a family of mathematical techniques that formally limit the amount of private information that can be inferred about each data subject. There are two main flavors of differential privacy, offering slightly different privacy guarantees: "global," which offers data subjects deniability of participation, and "local," which offers deniability of record content. Despite being slightly different, both operate by introducing randomization into computations on data to prevent an attacker from reasoning about its subjects with certainty. Ultimately, these techniques afford data subjects deniability while still allowing analysts to learn from the data.

## GDPR

Reading GDPR Recital 26, there is a strong argument that GDPR supports a relativist approach to identifiability and thereby de-identification. More precisely, Recital 26 specifies that "all the means reasonably likely to be used, such as singling out, either by the controller or by another person to identify the natural person directly or indirectly" should be taken into account to determine whether an individual remains identifiable. The GDPR standard is therefore that of anonymization, as defined above.[21] In practice, the Prosecutor Attack Model should be useful in this context, as it assumes the attacker has access to the data set and knows the complete set of publicly or otherwise reasonably attainable information about their target, including that which is realistically attainable but may or may not be plausibly readily available. Of note, the introduction in Recital 26 of the expression 'singling out,' i.e., locating an individual's record within a data set, should imply distinguishability of a record within a data set on the available attributes. That this, both distinguishability and availability should be necessary to single out.

GDPR introduces the notion of pseudonymization in Article 4(5).[22] Recital 26 confirms that pseudonymization is not enough to achieve anonymization. To make sense of this definition and distinguish pseudonymization from anonymization, one must assume that pseudonymization does not require acknowledging information that is not included within the data set to determine whether the data can be attributed to an individual. As a consequence, pseudonymization within the meaning of GDPR is not concerned with the treatment of indirect identifiers. What pseudonymization could thus cover in practice is the tokenization[23] of direct identifiers combined with key segregation. *Key segregation* means that the key used to generate the token is separated from the pseudonymized data through process firewalls.

## CPRA

CPRA clearly adopts a relativist approach to identifiability. Unlike the Health Insurance Portability and Accountability Act of 1996 (HIPAA),[24] CPRA appears to consider more than just the means and status of the anticipated data recipient to determine whether the data has been de-identified. This would suggest that CPRA adopts an anonymization standard. CPRA's de-identification test reads as follows:

"Deidentified" means information that cannot reasonably be used to infer information about, or otherwise be linked to, a particular consumer provided that the business that possesses the information

> (1) Takes reasonable measures to ensure that the information cannot be associated with a consumer or household.

21 Even if guidance released by Article 29 Data Protection Working Party prior to GDPR could appear more restrictive in that it seems to suggest that the raw data would need to be destroyed to meet the GDPR anonymization standard, this does not seem to be the standard adopted in practice. By way of example, the European Medicines Agency suggests that a risk-based approach to anonymization is a valid option under GDPR and that it makes sense when anonymizing clinical trial reports. There is no requirement to delete the raw data in this context. Compare Article 29 Data Protection Working Party, Opinion on Anonymisation techniques, WP 216 (Apr. 10, 2014), https://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2014/wp216_en.pdf with European Medicines Agency, External Guidance on the Implementation of the European Medicines Agency Policy on the Publication of Clinical Data for Medicinal Products for Human Use, EMA/90915/2016 (Oct. 15, 2018), https://www.ema.europa.eu/en/documents/regulatory-procedural-guideline/external-guidance-implementation-european-medicines-agency-policy-publication-clinical-data_en-3.pdf.

22 GDPR Art. 4(5) defines pseudonymization as "the processing of personal data in such a manner that the personal data can no longer be attributed to a specific data subject without the use of additional information, provided that such additional information is kept separately and is subject to technical and organisational measures to ensure that the personal data are not attributed to an identified or identifiable natural person."

23 *See* supra note 18.

24 45 C.F.R. § 164 (2021).

(2) Publicly commits to maintain and use the information in deidentified form and not to attempt to reidentify the information, except that the business may attempt to reidentify the information solely for the purpose of determining whether its deidentification processes satisfy the requirements of this subdivision.

(3) Contractually obligates any recipients of the information to comply with all provisions of this subdivision.

What CPRA also suggests by introducing the notion of pseudonymization is that tokenization combined with key segregation is not sufficient to avoid restrictions set upon use and disclosure.

In addition, CPRA adopts a definition of pseudonymization that is almost identical to the GDPR definition, and seems to suggest that it is possible to de-identify data that was once pseudonymized.[25] This makes sense, as de-identification goes beyond pseudonymization and requires the treatment of indirect identifiers. As the definition of de-identified data requires, both data and context controls will have to be combined, particularly organizational measures and legal obligations, to achieve de-identification.

## GPDPL

GPDPL also adopts a relativist approach to data de-identification. It defines anonymized data as data which is "related to a data subject who cannot be identified, considering the use of reasonable and available technical means at the time of processing."[26] GPDPL Article 5(6) goes on to say that "Anonymization is the use of reasonable and available technical means at the time of the processing, through which data lose the possibility of direct or indirect association with an individual."[27]

It further clarifies that anonymized data shall not be considered personal data unless the anonymization process can be reversed using reasonable efforts. The reasonableness of the technical means will be determined by taking into account objective factors such as "cost and time necessary to reverse the process of anonymization."[28] GPDPL leaves open the possibility for the National Data Protection Authority to further enunciate standards and techniques to be used for anonymization.[29]

What is more, GDPL Article 13(4) defines pseudonymization as "the processing by means of which data can no longer be directly or indirectly associated with an individual, except by using additional information kept separately by the controller in a controlled and secure environment." Again, to make sense of this definition and distinguish pseudonymization from anonymization, one must assume that pseudonymization does not require acknowledging information that is not included within the data set to determine whether the data can be attributed to an individual.

25 *See* CAL. CIV. CODE § 1798.140(s) (dealing with research).

26 GPDPL Art. 5 (I).

27 GPDPL Art. 5 (XI). As translated by Ronaldo Lemos et al, https://iapp.org/resources/article/brazilian-data-protection-law-lgpd-english-translation/.

28 GPDPL Art. 12.

29 GPDPL Art. 12. On December 28, 2018, the Brazilian National Data Protection Authority was created through Executive Order No. 869/18. See Isabel Carvalho & Rafael Loureiro, Brazil Creates a Data Protection Authority, HOGAN LOVELLS ENGAGE (Jan. 11, 2019), https://www.engage.hoganlovells.com/knowledgeservices/news/brazil-creates-a-data-protection-authority. The ANDP (the Portugese abbreviation for the agency) has highlighted 3 goals for regulation for 2021-2023, namely, "(i) to promote the strengthening of the culture of protection of personal data; (ii) establishing the effective regulatory environment for the protection of personal data; and (iii) improve the conditions for compliance with legal powers." See Hunton Andrews Kurth's Privacy and Cybersecurity, Brazilian Data Protection Authority Publishes Regulatory Strategy for 2021 – 2023, NATIONAL LAW REVIEW (Feb. 8, 2021), https://www.natlawreview.com/article/brazilian-data-protection-authority-publishes-regulatory-strategy-2021-2023#:~:text=The%20Brazilian%20National%20Council%20of,to%20privacy%20and%20data%20protection.

# Rule 3: Processing requires justification

Without a valid justification (or a legally valid reason), the processing of personal information shall not take place. The justification must be identified before the processing actually starts.

## GDPR

Under GDPR, justifications are termed legal bases. Depending upon the sensitivity of the data, whether the processing amounts to purely automated decision making or whether the processing includes a restricted transfer, GDPR sets forth four layers of justifications.

The first layer of justifications is found in Article 6: At least one Article 6 legal basis should be used for each processing activity. In other words, any processing of data would require an Article 6 justification. Very importantly, informed consent is not the only Article 6 legal basis available. In fact, GDPR is moving away from informed consent and excludes it as a matter of principle in situations of clear imbalance between the data subject and the controller.[30] Guidance issued by the European Data Protection Board confirms this trend and states that consent is not a valid legal basis in a series of scenarios.[31]

The choice of an Article 6 justification directly impacts the range of rights available to data subjects. By way of example, the right to data portability is only available when the processing is based upon consent or a contract, as per GDPR Article 20.

For special categories of data, i.e. sensitive data, an Article 9 justification must also be established. This is the second layer of justifications.

When the processing amounts to solely automated decision making, one Article 22 justification must be established. This is the third layer of justifications.

In addition, if data is transferred from the European Union to a third country, a chapter 5 legal basis must be established. This is the fourth layer of justifications.

## CPRA

CPRA is organized differently than GDPR and its list of justifications is much more loosely defined.

Justifications are for processing activities that are undertaken by the covered business, a service provider, or a contractor at the request of a covered business to pursue a business purpose, as listed in section 1798.140(e), or a commercial purpose, as defined in section 1798.140(g). These purposes are essentially to support business operations or broadly defined sales activities, such as granting access to data for monetary or other valuable consideration. In other words, implicit within the framework is the assumption that both operational purposes and sales are valid justifications for the processing of data, which is not a given under GDPR.

Again, under CPRA the choice of justification has an impact upon the types of rights available to consumers.

30 GDPR Recital 43.

31 European Data Protection Board, Guidelines 05/2020 on Consent Under Regulation 2016/679 (May 4, 2020), https://edpb.europa.eu/sites/edpb/files/files/file1/edpb_guidelines_202005_consent_en.pdf, p. 18–19.

## PIPEDA

In PIPEDA, consent plays a much bigger role than under either of the other two regimes in which entities are given leeway to collect data without consent. According to 4.3 Principle 3, "[t]he knowledge and consent of the individual are required for the collection, use, or disclosure of personal information, except where inappropriate."

The act does outline exceptions to this rule, including for the safety of the individual, cooperation with a criminal investigation, or when there are medical or mental health barriers to collecting consent.[32]

# Rule 4: Protecting the data means more than securing the data

Data protection is much more than data security, although data security is also a key data protection requirement.

Typical data protection requirements go further and include purpose limitation, data minimization, data accuracy, transparency, accountability, and fairness, in addition to security. Importantly, these data protection requirements are interdependent requirements and imply tradeoffs.

An effective way to build a data protection compliance strategy is to embed these requirements within workflows and systems right from the design stage, when the data platform is being architected, and break down each requirement by failure mode.[33]

Purpose limitation and data minimization are certainly some of the most important data protection requirements.

Let's go back to our previous example, illustrated in Table 1, to fully unpack data minimization. It is important to understand that data minimization is more than data sampling. *Data minimization* ensures that each attribute and attribute value is necessary for the analysis. In practice, this should mean performing data minimization both at the column- and row-levels. Note that the alternative is not necessarily to either remove or keep attribute values in the clear: Format-preserving data masking, for example, makes it possible to mask attribute values, but preserves the length and type of the value.

Data minimization can only be achieved once the goal of the analysis has been established, i.e, the *processing purpose*. In other words, data minimization is interrelated with the purpose limitation requirement. *Purpose limitation* mandates both a specified and limited purpose. Once the purpose has been specified and circumscribed, the data cannot be reused for a secondary purpose, unless a valid justification is established or in some cases, the secondary purpose is considered to be compatible with the primary purpose.

Processing activities must thus be organized by purpose through the creation of unlinkable processing domains. A *processing domain* is a set of implementation rules that is dependent upon the purpose for which the data is being processed and which defines the conditions under which the data can be processed, such as who can access the data and what it looks like.

A corollary to the requirements of data minimization

32 PIPEDA, S.C. 2000, c.5 7(2) (Can.).

33 *See* our previous white paper on data protection by process for a list of failure modes, Sophie Stalla-Bourdillon et al., Data Protection by Process: How to Operationalize Data Protection by Design for Machine Learning, IMMUTA (Dec. 2019), https://www.immuta.com/downloads/data-protection-by-process-fpf-whitepaper.

and purpose limitation is that as soon as the data ceases to be necessary for the processing, access should be terminated. Ultimately, if its retention cannot be justified, the data should be destroyed.

*Data accuracy* is essentially error-free and up-to-date data. Wrong values within a table thus generate data accuracy problems when data accuracy is necessary for the purpose fulfilled.

Transparency refers to being open about one's processing activities and describing them in a way that can be understood by outside individuals.

*Accountability* requires putting oneself in a position to be able to demonstrate compliance. In practice, this means creating a compliance role as well as creating an audit trail, and monitoring and auditing processing activities.

*Security* means protecting personal information from incidents or unwanted actions, such as unauthorized access. Integrity, availability, and confidentiality are three key security sub-requirements. It is important to understand that by locking some data items too early, such as through encryption at ingest-time,[34] one loses the ability to dynamically apply a variety of data masking techniques and reach a high level of data minimization. Even if encryption does not happen as part of the ingest process, the data can still be encrypted both at rest and in transit within the data analytics environment.

Ensuring a fair processing or *fairness* is one of the most challenging requirements to unpack and define.[35] It is related to discrimination and individual expectations, but arguably goes beyond it. At a high level, what is important to consider and distinguish is:

- The *processing impact*: the anticipated effect of the processing upon individuals' situations and rights, either because their data have been used as input to the analytics process and/or because their data will be used as input to the data product generated through the analytics process. Three types of harm are relevant at this stage: informational harm, behavioral harm, and collective harm.[36]
- The *processing assumptions*: considerations upon which the risk assessment is based to derive the anticipated processing impact.
- The *processing technique*: the method or mode of investigation employed to analyze the data.
- The *independent variables*: the variables upon which the results of the analysis (i.e., the dependent variables) depend. They are not necessarily properties of an individual or record. Independent variables can have a positive or negative impact upon the results of the analysis. Equality laws prohibit the use of protected characteristics when making decisions about individuals, so when the data product is intended to support decision making, independent variables should generally exclude protected characteristics. In some cases, unprotected characteristics can act as a proxy to protected characteristics and should also be carefully scrutinized.

34 Unlike other forms of masking, encryption is a function that can be reversed with what's called a "decryption key." An encryption algorithm, also called a cipher, is what takes a readable chunk of text and turns it into seemingly random values that are not decipherable to others (at least, not without the decryption key). In other words, organizations rely on encryption when they want the value of that data to be discoverable to specific users, but not to the entire world. For that reason, encryption is heavily relied upon for data security. Once data is encrypted, it is generally not useful until it is decrypted by someone who holds the decryption key. When encryption happens as part of the ingest process, the data arrives within the data analytics environment already encrypted.

35 *See* e.g., Damian Clifford & Jef Ausloos, Data Protection and the Role of Fairness, 37 YEARBOOK OF EUROPEAN LAW 130 (Aug. 9, 2018); Gianclaudio Malgieri, The Concept of Fairness in the GDPR: A Linguistic and Contextual Interpretation, 2020 Proceedings of FAT* 27 (Jan. 10, 2020); see also Reuben Binns, On the Apparent Conflict Between Individual and Group Fairness, 2020 Proceedings of FAT* 27 (Jan. 27, 2020), https://arxiv.org/abs/1912.06883; Sorelle A. Friedler, Carlos Scheidegger & Suresh Venkatasubramanian, The (Im)Possibility of Fairness: Different Value Systems Require Different Mechanisms for Fair Decision Making, 64 COMM. ACM 136 (2021). In a COVID-19 world, see Roberts Driggs et al. Common Pitfalls and Recommendations for Using Machine Learning to Detect and Prognosticate for COVID-19 Using Chest Radiographs and CT Scans, 3 NATURE MACHINE INTELLIGENCE 199 (2021).

36 Sophie Stalla-Bourdillon et al., Warning Signs: The Future of Privacy and Security in the Age of Machine Learning, IMMUTA (Sept. 2019), https://www.immuta.com/downloads/warning-signs-the-future-of-privacy-and-security-in-the-age-of-machine-learning/.

- The *risk assessment procedure*: the workflow and persona involved in deriving the processing assumptions, performing the risk assessment, and documenting and reviewing both over time. It is important that data analysts understand that they *"ought to be intentional and explicit about world-views and value assumptions: the systems they design will always encode some belief about the world."*[37]

Organizing processing domains by purpose and processing impact can help with identifying high-risk activities and activities that require more intensive monitoring and auditing than others.

## GDPR

GDPR is probably the most comprehensive data protection framework, based on the list of requirements imposed upon data controllers. GDPR Article 5 lists lawfulness, fairness and transparency, purpose limitation, data minimization, storage limitation, accuracy, confidentiality, integrity, and accountability as requirements. What is more, GDPR Article 25 provides that these requirements be implemented before the processing actually starts, at the time at which the processing means are determined. This is called data protection by design and is extended by a data protection by default approach.

Additionally, GDPR imposes recording obligations to all data controllers with 250 employees or more and risk assessment obligations upon all data controllers in high-risk situations.[38] Characterizing the processing impact is thus key to achieving GDPR compliance. As per GDPR Article 36, if it proves impossible to mitigate a high risk through technical and organizational measures, the controller is obliged to consult the supervisory authority. As a result, it is crucial to be able to detect high-risk activities, tailor technical and organizational measures to mitigate the risk, and document these measures.

## CPRA

CPRA is less explicit than GDPR but appears to comprise purpose limitation,[39] data minimization,[40] storage limitation,[41] transparency,[42] data accuracy[43] and security[44] requirements. It is worth noting that CPRA, contrary to GDPR, expressly identifies processing purposes and thereby implies that they are legitimate justifications for processing personal information.[45] Given the wide range of business and commercial activities these processing purposes cover, it seems easier to justify a processing activity under CPRA than GDPR.

Of note, as provided by section 1798.185(a)(15), "regulations requiring businesses whose processing of consumers' personal information presents significant risk to consumers' privacy or security" to "submit to the California Privacy Protection Agency on a regular basis a risk assessment" weighing the benefits of the processing against the risks posed to consumer rights are expected.[46]

37 *See* Friedler, Scheidegger, Venkatasubramanian, supra note 35.

38 *See* GDPR Arts. 30 and 35.

39 *See* CAL. CIV. CODE §§ 1798.100(a)(1), (c), 1798.140(e).

40 CAL. CIV. CODE § 1798.100(c).

41 CAL. CIV. CODE § 1798.100(c).

42 *See* CAL. CIV. CODE §§ 1798.100(a)(1), 1798. 110.

43 *See* CAL. CIV. CODE § 1798.106.

44 CAL. CIV. CODE §§ 1798.100(e), 1798.81(5), 1798.150.

45 CAL. CIV. CODE §§ 1798.140(e) and 1798.140(g).

46 CAL. CIV. CODE §§ 1798.185(a)(15).

## PIPEDA

An interesting feature of the PIPEDA is that data protection requirements are expressly defined in an interdependent fashion. For example, under the Identifying Purposes principle, paragraph 4.2.1 states that "[t]he organization shall document the purposes for which personal information is collected in order to comply with the Openness principle (Clause 4.8) and the Individual Access principle (Clause 4.9)."

Paragraph 4.2.2 explains that "[i]dentifying the purposes for which personal information is collected at or before the time of collection allows organizations to determine the information they need to collect to fulfill these purposes. The Limiting Collection principle (Clause 4.4) requires an organization to collect only that information necessary for the purposes that have been identified." Paragraph 4.2.4 mentions the consent requirement and paragraph 4.2.6 states that the Identifying Purpose principle is closely tied to the Limited Collection principle and the Limited Use, Disclosure, and Retention principle.

Similarly, paragraph 4.4.3 states that the Limited Collection principle is closely tied to the Identifying Purposes and the Consent principles. It appears that these various principles are built to bolster protections and feed into each other. Without the Identifying Purposes principle, it would be difficult to accurately gauge whether the Limiting Collection principle is being met. Similarly, without the Openness principle, the consent requirement would be quite barebones. In this way, each principle stands on its own as a protection, but they also each support the others.

# Rule 5: Individuals are able to intervene into the processing

One key goal of data protection is to inject individual control within data processing activities for at least two reasons: first to empower individuals; and second to make enforcement more effective.

Individuals can thus intervene at various points of the process with different types of prerogatives.

The most common individual rights are the rights to processing information, data access, object or opt-out, deletion, correction, and portability. None of these rights are absolute and, depending upon the framework at stake, the individual must meet a certain set of conditions to be able to exercise their rights. It is worth noting that the choice of the justification used to ground the processing can have an impact upon the range of rights available to individuals. The same is true with the sensitivity of the data and the degree of data transformation.

Generally speaking, it is possible to distinguish between seven types of individual intervention:

- **Data correction request**: a request to correct wrong attribute values or update attribute values that are associated with an individual.
- **Data access request**: a request to access one's personal information. In practice, this usually means requesting access to both the attributes and attribute values associated with an individual. Additional information can also be available through a data access request, particularly for information that relates to the data's environment, such as the purposes for which the personal information is being processed.
- **Data portability request**: a request to move one's personal information from one data environment to another and thereby to make it reusable in another data environment.

- **Data deletion request**: a request to have one's personal information put beyond use and then destroyed.
- **Restriction request**: a request to restrict or limit the purposes for which the personal information is being processed.
- **Opt-out request**: a request not to proceed with or to terminate the processing of one's personal information for a given purpose.
- **Opt-in consent**: consent to proceed with the processing of one's personal information for a given purpose. This type of intervention is sometimes considered to be a valid justification for processing data.

As a result, a data analytics environment should support the following four intervention functions and generate logs to capture metadata when these functions are being performed:

- **Processing termination**: the action of stopping the processing. In practice, this will mean that access to a processing domain must be time-based.
- **Data deletion**: the process by which data is put beyond use and destroyed. Processing termination is thus an essential primary step of a data deletion process.
- **Data export**: the outputting of data for use by other systems. This involves translating the data into a format that can be reused by other systems.
- **Data rewriting**: the process of replacing attribute values. This function does not necessarily imply that all data analysts should have rewrite permissions for all processing purposes.

Of note, while preventing personal direct identifiers from being ingested within a closed and controlled data analytics environment is likely to make individual intervention moot, dynamically generating tokens[47] makes it possible to preserve opt-out requests in a greater number of circumstances.

## GDPR

Under GDPR Chapter 3, these seven types of interventions are possible. Informed consent is the justification that gives individuals the most prerogatives under this framework.[48] It is worth noting that, as a matter of principle, if the controller is capable of demonstrating that they are not in a position to identify the individual, most data subject rights are not applicable.[49]

The four intervention functions described above are thus essential to claim compliance with GDPR. It is also important to identify the justification for which the data is being processed prior to performing such functions.

47 A token (i.e., the output of a masking process using a tokenization method) replaces an attribute with a mathematically unrelated value through a transformation process that is difficult to reverse. A token replacing a personal identifier becomes identifying when the attacker possesses (or has access to) additional information that allows him to reverse the transformation. In addition, even if the token is not reversed, a token that replaces an indirect identifier can still act as a personal indirect identifier if tokenization is not performed by value (and not by record.) In other words, a token can still act as a personal indirect identifier if the tokenization method is homogeneously applied to the entire dataset.

48 *See* GDPR Arts. 7 and 20.

49 GDPR Art. 11(2).

## CPRA

Most of these interventions are provided under CPRA as well. The right for restriction, in which a data subject may restrict the use of their information to that which is necessary to perform the requested services, is provided specifically for sensitive data. The CPRA's opt-in provision only applies to minors, who must be given 12 months after declining a request to share their data before being asked to opt-in again. Notably, these rights were strengthened under CPRA from their CCPA counterparts to provide consumers with more intervention prerogatives. According to the implementing regulations that will be issued by July 1, 2022, there is a chance that consumers will be given the right to also opt-out of automated decision making.[50] As such, it would be important to ensure that processing environments are equipped to handle increased intervention rights for consumers.

## GPDPL

Article 18 of GPDPL seems to provide many of the same interventions provided under GDPR[51] To exercise these rights, GPDPL requires that the consumer provide an express request.[52] Given that GPDPL provides most of the seven interventions outlined above, it would be important that the processing environments have capabilities for the four functions.

50 CAL. CIV. CODE §§ 1798.185(a)(16) and (22)(d).

51 For a comparison between GDPR and GPDPL see Comparing Privacy Laws: GDPR v. LGPD, ONETRUST DATAGUIDANCE, https://www.dataguidance.com/sites/default/files/gdpr_v_lgpd_revised_edition.pdf.

52 GPDPL Article 18(IX) para. 3.

# Key takeaways

Table 4 recaps the main findings for each syntactic rule exposed in section 2.

| RULE | IMPLICATIONS |
|---|---|
| **Rule 1: Identifiability attracts protection** | ▪ Data protection and privacy frameworks protect personal information.<br>▪ Personal information covers personal identifiers (direct identifiers, indirect identifiers) and other attribute values associated with these personal identifiers. All these items of data are protected under data protection and privacy frameworks.<br>▪ Group identifiers are also personal indirect identifiers.<br>▪ Inferences can be either personal identifiers or other attribute values associated with these personal identifiers. |
| **Rule 2: De-identification weakens the legal protection** | ▪ De-identification requires detecting both direct and indirect identifiers.<br>▪ Pseudonymization on its own does not necessarily achieve de-identification, as it is only concerned with direct identifiers.<br>▪ Anonymization is a demanding de-identification method, which requires taking into account the re-identification means of all situationally-relevant potential attackers, not only those of the anticipated recipient.<br>▪ De-identification results in individuals losing their ability to intervene into the processing of de-identified data.<br>▪ However, in many cases both data custodians and data recipients remain subject to a series of obligations, such as not to re-identify and to comply with data breach mitigation plans. |
| **Rule 3: Processing requires justification** | A justification for processing the personal information must be established for each processing activity. |
| **Rule 4: Protecting the data means more than securing the data** | ▪ Securing the data is only one data protection requirement. Beyond security, one finds in particular data minimization, purpose limitation, transparency, accountability, and fairness.<br>▪ Data minimization and purpose limitation require organizing processing activities by purpose.<br>▪ Data minimization should be done both at the column- and row-level.<br>▪ Fairness requires organizing processing activities by impact and setting up an appropriate audit trail. |
| **Rule 5: Individuals are able to intervene into the processing** | Seven types of individual interventions have been recognized within data protection and privacy frameworks. These seven types of individual interventions require the implementation of four key data-related functions: processing termination, data deletion, data export, and data rewriting. |

Table 4. Summary of the 5 rules

It is now possible to draw concrete lessons for the architecturing of a data analytics platform built upon a multi-cloud environment. The owner of such a data analytics platform should make sure its policy layer is able to do the following:

1. Discover and/or catalog personal identifiers, conditional personal information, and sensitive personal information.
2. Distinguish between processing activities in relation to their justification. One effective way to do this is to leverage attribute-based access control, or ABAC.[53]
3. Combine and enforce both data and context controls, including an audit trail supporting monitoring in real time and auditing.
4. Enforce de-identification policies as often as possible.
5. Implement purpose-based access control at the processing domain level.
6. Make processing domains unlinkable.
7. Organize processing domains by risk level.
8. Make it possible to attach risk assessment documentation to processing domains and enable their monitoring and auditing.
9. Make subscription to processing domain time-based.
10. Offer a wide range of data masking techniques that can be implemented both at the column- and row-levels.
11. Support processing termination, data deletion, data export, and data rewriting.
12. Enable different personas including compliance roles to interact and collaborate as closely as possible to the data and processing activities.

53 *See* Steve Touw, Role-Based Access Control vs. Attribute-Based Access Control — Explained, IMMUTA, https://www.immuta.com/articles/attribute-based-access-control/.

# Appendix I: The data protection lexicon

The formulation of the five syntactic rules exposed in this white paper was made possible by the construction of a data protection lexicon, which should inform the generation of metadata and their management within data analytics environments. The data protection lexicon comprises the following terms:

- **Accountability**: putting oneself in a position to be able to demonstrate compliance. In practice, means creating an audit trail, and monitoring and auditing processing activities.
- **Anonymization**: the process by which the link between the data and the individual is considered broken after having acknowledged information that is not included within the data set to determine whether the data can be attributed to an individual and considered all potential attackers. In practice, this means implementing a process by which personal identifiers are made undistinguishable and/or unavailable to all situationally-relevant potential attackers.
- **Attack model**: a collection of assumptions and constraints on the control environment and/or the behavior and background knowledge of an attacker.
- **Attribute**: a piece of information associated with a record or individual, which can either be unique to one individual or record (i.e. highly distinguishable) or common to many individuals or records (undistinguishable). Attributes are also either available (attainable by an attacker) or unavailable (unattainable by an attacker).
- **Conditional personal information**: personal attribute values that are not distinguishable and/or available to an attacker. They are conditional in the sense that if personal identifiers are transformed and additional controls are in place to satisfy the applicable de-identification test, they are no longer considered personal information.
- **Controls**: organizational, legal, or technical measures put in place to reduce re-identification risks. Data controls affect the visibility of the data, whereas context controls affect the environment of the data.
- **Data access request**: a request to access one's personal information. In practice, this means requesting access to both the attributes and attribute values associated with an individual.
- **Data accuracy**: error-free and up-to-date data.
- **Data correction request**: a request to correct wrong attribute values or update attribute values that are associated with an individual.
- **Data deletion**: the process by which data is put beyond use and destroyed. Processing termination is thus an essential primary step of a data deletion process.
- **Data deletion request**: a request to have one's personal information beyond use and destroyed.
- **Data export**: the outputting of the data for use by other systems. It thus involves translating the data into a format that can be reused by other systems.
- **Data masking**: a function that is applied to raw personal information to hide its true value. Masking is a broad term that can describe a wide range of functions, including hashing, encryption, and a number of other techniques.
- **Data minimization**: the requirement to calibrate the data to the processing purpose, ensuring that only the data necessary to pursue the purpose is being processed. In practice, this means ensuring that each attribute and attribute value is necessary for the analysis.
- **Data portability request**: a request to move one's personal information from one data environment to another and thereby to make it re-usable in another data environment.

- **Data rewriting**: the replacement of attribute values.
- **De-identification**: the process by which the link between the data and the individual is considered broken after having acknowledged information that is not included within the data set to determine whether the data can be attributed to an individual. In practice, this means implementing a process by which personal identifiers are made undistinguishable and/or unavailable to a situationally-relevant potential attacker (e.g., the anticipated recipient).
- **De-identification domain**: a set of rules that define the conditions under which the data can be processed in a de-identified state. In practice, if tokenization is implemented, this will mean that tokenization will be domain specific.
- **Direct personal identifier**: an attribute value that is unique to an individual and available to an attacker.
- **Group identifier**: an attribute value that can be used to discriminate among groups. Note that a group identifier can also be a personal identifier if the group is relatively small, such as a household. In this case, the group identifier is, in effect, an indirect personal identifier.
- **Identifiability**: the capability of linking data to an individual. It depends upon both distinguishability (the ability of an attacker to distinguish the individual from others) and availability (the ability of an attacker to access the information within the data set and information that is not included within the data set but can be matched with the former).
- **Independent variables**: the variables upon which the results of the analysis (i.e., the dependent variables) depend. They are not necessarily properties of an individual or record. Independent variables can have a positive or negative impact upon the results of the analysis.
- **Indirect personal identifier**: an attribute value, such as height, ethnicity, or hair color, that is not unique to an individual but can be used in combination with other attributes to distinguish an individual and is available to an attacker.
- **Inferences**: attribute values which can be confidently guessed or estimated through analysis when considering attribute values within a data source alone or in combination with information outside the data source, can also be personal information (personal identifiers or conditional personal information) and as such legally protected. This includes inferences for attributes not represented in the data.
- **Justification**: a legally valid reason upon which the processing is based.
- **Opt-in consent**: consent to proceed with the processing of one's personal information for a given purpose.
- **Opt-out request**: a request not to proceed with or to terminate the processing of one's personal information for a given purpose.
- **Personal identifier**: an attribute value that distinguishes an individual and is available to an attacker. An attribute is available to an attacker when it is publicly available, observable, attainable. The two main characteristics of attributes are therefore distinguishability (or their ability to single out) and availability.
- **Personal information**: personal identifiers and attribute values that are associated with personal identifiers.
- **Processing assumptions**: considerations upon which the risk assessment performed to derive the anticipated processing impact is based.
- **Processing domain**: a set of rules that define the conditions under which the data can be processed. It is dependent upon the purpose for which the data is being processed.

- **Processing impact**: the anticipated effect of the processing upon individuals' situations and rights, either because their data have been used as input to the analytics process and/or because their data will be used as input to the data product generated through the analytics process.
- **Processing purpose**: the objective the data processing intends to achieve.
- **Processing technique**: the method or mode of investigation employed to analyze the data.
- **Processing termination**: the action of stopping the data processing. In practice, this will mean that access to a processing domain will end.
- **Pseudonymization**: the process by which the link between the data and the individual is considered broken without acknowledging information that is not included within the data set to determine whether the data can be attributed to an individual. As a consequence, pseudonymization is not concerned with the transformation of indirect identifiers as indirect identifiers are only identifying to the extent there is access to information that is not included within the data set. In practice, this means implementing a process by which direct identifiers are made undistinguishable and/or unavailable to a situationally-relevant attacker (e.g., the anticipated recipient).
- **Purpose limitation**: the requirement to specify and limit the purpose for which the data will be processed. In practice, this means defining a processing domain for each processing purpose and making sure the data is not reused within other processing domains without a justification, unless the secondary processing purpose is deemed compatible with the primary processing purpose.
- **Restriction request**: a request to restrict or limit the processing purposes for which personal information is being processed.
- **Risk assessment procedure**: the workflow and persona involved in deriving the processing assumptions, performing the risk assessment, and documenting and reviewing both the processing assumptions and the risk assessment over time.
- **Security**: protecting personal information from incidents or unwanted actions such as unauthorized access. Integrity, availability, and confidentiality are three key security sub-requirements.
- **Sensitive personal information**: attribute values that are associated with personal identifiers, the disclosure or misuse of which are considered particularly harmful to individuals.
- **Transparency**: being open about the processing activities that one undertakes and describing them in a way that can be understood by outside individuals.

IMMUTA™

115 Broad Street, 6th Floor, Boston, MA 02110 | immuta.com | (800) 655-0982