IMMUTA

WHITE PAPER

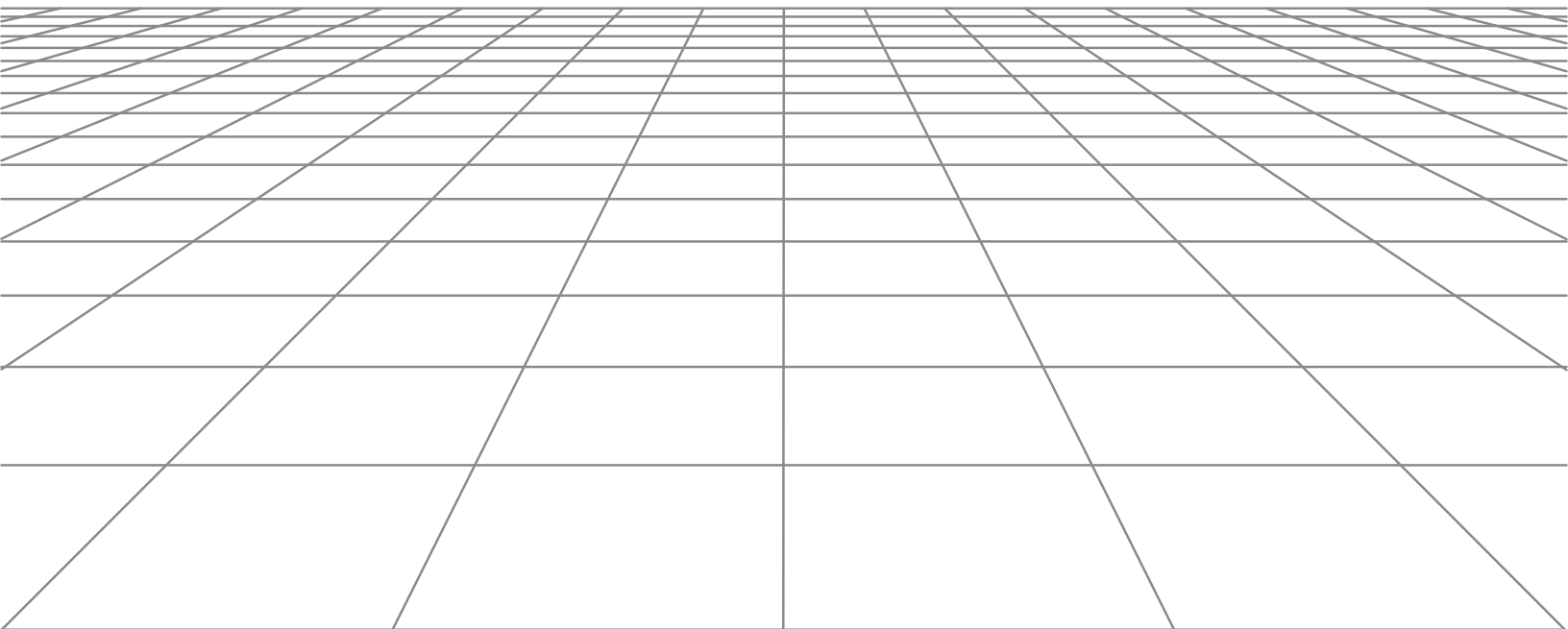
# Data Classification 101

## How to kick off the data protection journey


**SOPHIE STALLA-BOURDILLON**
Principal Legal & Privacy Engineer, Immuta

**CAROLINE BAILEY**
Immuta Scholar

**CLAY GOODE**
Immuta Scholar

# Table of Contents

# Introduction

Security and privacy are converging. In the world of data science, this means two things. First, security and privacy threats largely overlap.[1] Data security is now much more than preventing and mitigating unauthorized access to data – it is also about preventing and mitigating against unintended inferences,[2] which is at the heart of informational privacy.[3] Second, data security controls now mirror data privacy controls. The new cybersecurity paradigm, i.e., Zero Trust Architecture,[4] when operationalized within data science environments, leads to data segmentation, data transformation, and data minimization, three fundamental data protection principles.

At its core, the Zero Trust Architecture paradigm aims to eliminate trust within IT systems. Users, devices and networks are all deemed unreliable and therefore are not trusted. When combined with a data-centric security management approach,[5] this paradigm usually translates into four high-level requirements. First, exposure to regulated data must be reduced to a minimum; in other words, data consumers must only get access to the data they need to perform their tasks under the least privilege principle.[6] This can entail implementing data transformation techniques to mask unnecessary and/or confidential attributes. Second, data access must be time-based, to force regular verification of users, even when self-executing rules block unauthorized access to sensitive information. Third, data must be segmented through compartmentalization techniques, with a view toward reducing the attack surface and ensuring that if one component of the data architecture is compromised, it will not impact the other components without jeopardizing data sharing. Fourth, data access and usage must be constantly monitored to set base practices and detect any deviations.

The convergence between data security and data privacy controls couldn't be clearer. The least privilege principle aligns with the data minimization principle found in most privacy or data protection frameworks, which mandates that the data

1 Andrew Burt, *Privacy and Cybersecurity Are Converging. Here's Why That Matters for People and for Companies*, Harvard Business Review (Jan. 3, 2019), https://hbr.org/2019/01/privacy-and-cybersecurity-are-converging-heres-why-that-matters-for-people-and-for-companies.

2 *Id*. The traditional information security model relies upon the confidentiality, integrity and availability triad to derive relevant threats for IT systems. With this said, unauthorized access is often the major concern. This model has been operationalized in industry, e.g., through Microsoft's STRIDE methodology, which aims to ensure that an application meets CIA requirements (confidentiality, integrity and availability) in addition to authorization, authentication and non-repudiation. Microsoft, *Microsoft Threat Modeling Tool threats*, Azure Product Documentation (Aug. 25, 2022), https://learn.microsoft.com/en-us/azure/security/develop/threat-modeling-tool-threats).

3 *See, e.g.*, Clarke, *Internet Privacy Concerns Confirm the Case for Intervention*, 42 Communications of the ACM 2 (1999).

4 NIST (National Institute of Standards and Technology), *Zero Trust Architecture*, Special Publication 800-207 (Aug. 10, 2020), https://www.nist.gov/publications/zero-trust-architecture; Cybersecurity & Infrastructure Security Agency, *Zero Trust Security Model* (Apr. 20, 2022), https://www.cisa.gov/zero-trust-maturity-model; Executive Order on Improving the Nation's Cybersecurity (12 May 2021), https://www.whitehouse.gov/briefing-room/presidential-actions/2021/05/12/executive-order-on-improving-the-nations-cybersecurity/; Moving the U.S. Government Toward Zero Trust Cybersecurity Principles M-22-09 (Jan. 26, 2022), https://www.whitehouse.gov/wp-content/uploads/2022/01/M-22-09.pdf; United States Department of Defense (DoD) Zero Trust Reference Architecture, https://dodcio.defense.gov/Portals/0/Documents/Library/(U)ZT_RA_v1.1(U)_Mar21.pdf; United-Kingdom National Cybersecurity Center, Zero Trust Architecture Design Principles (July 23, 2021), https://www.ncsc.gov.uk/collection/zero-trust-architecture.

5 NIST explains that conventional network-centric security measures are inapt to protect current communications and information systems and a shift towards a data-centric security management approach is needed. NIST, *Data Classification Practices: Facilitating Data-Centric Security Management*, NIST's National Cybersecurity Center of Excellence (May 2021), https://www.nccoe.nist.gov/sites/default/files/legacy-files/data-classification-project-description-draft.pdf. Importantly however, all assets (data sources and computing devices) are resources, and therefore should be protected. This is tenet number 1 of the NIST zero trust architecture's approach. NIST, supra note 4.

6 *See* CNSSI 4009, https://www.cnss.gov/CNSS/issuances/Instructions.cfm; NIST, *An Introduction to Information Security*, Special Publication 800-12 Revision 1 (June 2017), https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.800-12r1.pdf; NIST, *Protecting Controlled Unclassified Information in Nonfederal Systems and Organizations*, Special Publication 800-171 Revision 2 (February 2020)), https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.800-171r2.pdf

be adequate, relevant and, above all, limited to what is necessary in relation to the purpose for which it is processed.[7] Time-based data access can be seen as a variation of the storage limitation principle, which requires that personal data be kept or accessed for no longer than is necessary for the data's processing. Data segmentation is a technique used to guarantee unlinkability across processing operations or data domains and enforce the principle of purpose limitation, which imposes the processing of personal data for limited purposes.[8] The monitoring and auditing of data processing is critical to detect deviations from recording of processing activities,[9] data protection risk assessments,[10] and more generally, to abide by the principle of accountability, which forces data controllers to put themselves in a position to demonstrate compliance with the prescriptions of the framework.[11]

To prevent data breaches, security and privacy teams must work together to build end-to-end controlled data environments. This process starts with data classification. Data classification is commonly understood as the identification of data types an organization holds and processes. Each data type is usually associated with a confidentiality and criticality level, with a view to drive the formulation of access control and usage rules. Despite the apparent simplicity of the exercise, which is often executed through the creation of three to four data classes (e.g., public, internal, confidential, and eventually, restricted)[12], data classifications often lack the granularity required to produce fine-grained data access and usage rules. These rules make access or usage dependent upon the data consumer's precise needs. They are a must-have for implementing best-in-class data security and privacy controls, such as least privilege and data minimization.[13]

The goal of this white paper is to help security and privacy teams kick off the data protection journey[14] and build an effective and scalable data classification strategy. To this end, we will break down the data classification process into three steps:

1. Identifying data classes and class properties
2. Mapping entity tags to classification tags
3. Building policy tables

The white paper thus suggests that data classification must be multi-dimensional to effectively support the production of fine-grained access control and usage rules. At a minimum, data classes should be associated with three innate or independent dimensions: class properties (e.g., confidentiality, criticality), processing purposes (e.g., payment processing, auditing), and user attributes (e.g., revenue operations, IT security, or marketing analytics), and one consequential dimension related to data treatment (e.g., data masking), which is dependent upon the three innate dimensions mentioned above. Additional consequential dimensions, such as data actions and restrictions set upon the range of tools used to consume the data, can also be added to the data classification primitive laid out in this white paper.

Throughout the white paper, you'll also find a set of ten recommendations for data security and privacy teams to help streamline the data classification process and ensure that all security and privacy requirements are satisfied.

7 *See, e.g., Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on 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)*, OJ L 119, 4.5.2016, p. 1–88, (GDPR) art. 5(1)©.

8 *See, e.g.*, as an expression of the unlinkability data protection goal, the German Standard Data Protection Model, https://www.datenschutzzentrum.de/uploads/sdm/SDM-Methodology_V2.0b.pdf.

9 *See, e.g.*, GDPR art. 30.

10 *See, e.g.*, GDPR art. 35.

11 *See, e.g.*, GDPR art. 5(2).

12 *See* in particular data classifications used by academic institutions, e.g., https://www.umsystem.edu/ums/is/infosec/classification-definitions; https://www.cmu.edu/data/guidelines/data-classification.html; https://www.bristol.ac.uk/secretary/data-protection/policy/classification/; https://www.dcu.ie/sites/default/files/2020-11/24-data_classification_policy_iss_v1.1.pdf.

13 As explained by NIST, "existing NIST standards and guidance regarding data classification and labeling, such as Federal Information Processing Standard (FIPS) 199 and NIST Special Publication (SP) 800-60, address federal government-specific requirements, but not the many other requirements to which federal agencies and other organizations are subject." NIST, *supra* note 5.

14 In this white paper, the term data protection is understood broadly and covers both security- and privacy-related concerns. These terms often overlap, which is a reason to favor the expression 'data protection' over 'data privacy.' For example, NIST makes it clear that, "a critical factor for achieving success in any business is the ability to share information and collaborate effectively and efficiently *while satisfying the security and privacy requirements* for protecting that information" (emphasis added). *Id.*

# Formulating Data Classes and Class Properties

A data protection journey usually comprises three steps: 1) mitigating risks *to* the data (information leakage)[15], 2) mitigating risks *from* the data (information flaws such as bias[16] or error), and 3) mitigating risks *with* the data (information misuse which could lead to discrimination or other unfavorable consequences for the organization or the data subjects)[17].

Whatever the first challenge security and privacy teams choose to tackle, the data protection journey always starts with data classification. Let's introduce a few definitions that are useful to grasp before kicking off a data classification project.

| | DEFINITIONS |
|---|---|
| **Data class** | A data class is a group of data types used to drive data access and usage rules. |
| **Class property** | A class property is a trait of a specific class expressed in terms of security, privacy/data protection goals, such as confidentiality and criticality. Data classes do not necessarily have the same level of confidentiality or criticality. |
| **Conditional class property** | A class property is conditional when the property depends upon the characterization of a condition. This could be the combination of two classes within a row or a query, such as the classes **Identifier** and **Sensitive Attribute**, which are defined below. For example, a username is confidential when it is associated with a password. |
| **Absolute class property** | A class property is absolute when the property does not depend upon the characterization of a condition. For example, a social security number is always confidential, regardless of the other attributes associated with it within a row. |
| **Confidentiality** | Data is confidential if its use or disclosure must be restricted by processing activity type or data consumer attribute, in order to prevent leakage of information. Confidentiality is a class property. |

15 Information leakage is defined as the release of information, intentional or unintentional, to an untrusted environment. *See* NIST, *Security and Privacy Concerns for Information Systems and Organizations*, Special Publication 800-53 Revision 5 (Sept. 2020), https://doi.org/10.6028/NIST.SP.800-53r5.

16 Bias is defined as a value that is more likely to be chosen than another available value, as contrasted with 'unbiased.' *See* NIST, *Recommendation for Random Number Generation Using Deterministic Random Bit Generators*, Special Publication 800-90A Revision 1 (June 2015), http://dx.doi.org/10.6028/NIST.SP.800-90Ar1; NIST, *Towards a Standard of Identifying and Managing Bias in Artificial Intelligence*, Special Publication 1270 (March 2022), https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.1270.pdf.

17 *See, e.g.*, Sandra Wachter et. al, *Why Fairness Cannot Be Automated: Bridging the Gap Between EU Non-Discrimination Law and AI,* arXiv (May 12, 2020), arXiv:2005.05906; Sandra Wachter et. al, *The Legality of Fairness Metrics Under EU Non-Discrimination Law*, 123 West Virginia Law Review 3 (January 15, 2021).

| | |
|---|---|
| **Criticality** | Data is critical if it must remain uncorrupted, reliable, and readily available within an organization because its processing is important for the success of the organization's operations. Criticality is a class property.[18] |

Data confidentiality and data criticality sit on a spectrum. They are business and/or legal requirements; security and privacy frameworks impose a wide range of obligations upon covered entities, ranging from confidentiality, integrity, and accuracy, to quality and availability. These obligations are often phrased differently across frameworks and can be either horizontal (across all sectors) or vertical (sector-specific).

Data classification should be informed by applicable regulatory frameworks. This leads us to formulate our first three recommendations:

| | RECOMMENDATION | RATIONALE |
|---|---|---|
| 01 | When **Personal Information** is processed as a class, two subclasses should be distinguished: **Personal Identifiers** and other attributes, such as **Sensitive Attributes**. It is likely that these classes will partially overlap. | Generally speaking, the association of the data with an individual triggers the obligation to handle the data with care under data protection frameworks.[19] It is important to specifically detect the presence of personal identifiers within the data to validate the classification. Importantly however, under many data protection frameworks, the category of personal information is broader than the category of personal identifiers, at least when these identifiers remain in the clear.[20] Hence the importance of making sure that the **Personal Information** data class is broader than the **Personal Identifier** data class. This step is crucial for being in a position to execute data subject requests and either grant access to data, allow data subjects to port it, correct it, delete and/or restrict it, and terminate processing activities, as illustrated in the table below.<br><br>Data protection frameworks usually distinguish between 'regular' personal information and 'sensitive' personal information (sometimes called 'special' category data)[21]. These frameworks then impose additional restrictions upon the use and disclosure of this category of information. It thus makes sense to specifically track access and use of sensitive information. |

18 Criticality can be further down into integrity and availability. *See* NIST, *Standards for Security Categorization of Federal Information and Information Systems*, FIPS Publication 199 (Feb. 2004), https://nvlpubs.nist.gov/nistpubs/fips/nist.fips.199.pdf; NIST, *Volume I: Guide for Mapping Types of Information and Information Systems to Security Categories*, Special Publication 800-60 Volume I, Revision 1 (Aug. 2008), https://nvlpubs.nist.gov/nistpubs/legacy/sp/nistspecialpublication800-60v1r1.pdf.

19 *See, e.g.*, Cal. Civ Code §§ 1798.100 (2020).

20 *See* Sophie Stalla-Bourdillon & Jonnalagadda, *A Data Protection Grammar*, Immuta Resources (2021), https://www.immuta.com/resources/a-data-protection-grammar/.

21 With this said, either the justification for processing sensitive information is harder to find, and/or individuals are granted more rights when sensitive information is processed. In particular, they can gain the ability to restrict processing activities to their core, i.e., processing activities that are necessary for the performance of the contract or the provision of services and/or to prevent certain categories of further processing such as granting access to third parties. By way of example, the California Privacy Rights Act (CPRA), an amendment to the California Consumer Privacy Act (CCPA), defines sensitive information as a subcategory of personal information which includes data elements that reveal a consumer's social security number, racial or ethnic origin, precise geolocation, genetic data or "the processing of biometric information for the purpose of uniquely identifying a consumer." A consumer can restrict an organization's processing of this category of data to certain enumerated purposes, such as those necessary to provide the goods or services, for example, restricting an organization's processing of an employee's social security number to identification for tax purposes. However, organizations are only subject to such obligations if their use of "sensitive personal information" implies "inferring characteristics about a consumer." Cal. Civ Code §§ 1798.100 *et seq* (2020).

02 When **Sensitive Attributes** and **Personal Identifiers** are associated together, the level of confidentiality for these data classes should increase.

When adopting a risk-based approach and focusing on individual harm, it is fair to state that the sensitivity of the data has a direct correlation with the intensity of the harm felt by individuals if their data is misused. However, the matter is further complicated by the fact that harm is also dependent upon the level of identifiability within the data.[22] What this means from a data classification perspective is that the combination of personal identifiers and sensitive information that should be restricted and carefully monitored first and foremost.[23]

03 Data classification should be performed dynamically: when data is at rest and when data is being used, i.e., when it is being queried.

Recommendation 4 is a consequence of Recommendation 3. Being exposed to sensitive data (i.e., having the means to access sensitive data) does not necessarily mean that sensitive data has been accessed. Performing data classification at the query level is useful to assess what has actually been consumed by the data consumer. For example, query results could be returning aggregates of sensitive attributes with no personal identifiers.

Crucially, defining data classes should also be informed by de-identification best practices.[24] This leads us to stress the importance of Recommendation 3. Although not all frameworks expressly acknowledge the importance of restricting the combination of identifiers and sensitive information to prevent harm,[25] this requirement should be considered implicit when the framework endorses a risk-based approach.[26] Of note, some frameworks adopt a limited approach to the definition of identifiers, which is not adapted to all data processing contexts.[27]

22 This is explicitly acknowledged by the methodology developed by the French regulator, the Commission Nationale Informatique et Libertés (CNIL). CNIL explains that, "severity represents the magnitude of a risk. It essentially depends on the level of identification of personal data and the level of consequences of the potential impacts." *See* CNIL, Methodology for Privacy Risk Assessment (2012), https://www.dataguidance.com/sites/default/files/methodology_for_privacy_risk_management_how_to_implement_the_data_protection_act.pdf. More recent guidance is less detailed on this point. *See, e.g.*, CNIL, Privacy Impact Assessment Methodology (PIA) (2018), https://www.cnil.fr/sites/default/files/atoms/files/cnil-pia-1-en-methodology.pdf.

23 This is not to say that the processing of sensitive information in the absence of identifiers could not generate harm, but harm is likely to be collective rather than purely individual in this situation. See Sophie Stalla-Bourdillon et. al, Warning Signs: The Future of Privacy and Security in the Age of Machine Learning, Immuta Resources (Sept. 2019), https://www.immuta.com/resources/warning-signs-the-future-of-privacy-and-security-in-the-age-of-machine-learning/.

24 De-identification best practices are described in both the technical literature (e.g, on statistical disclosure) and within regulatory guidance. It is crucial to distinguish two claims: 1) a state-of-the art de-identification technique has been applied upon the data and 2) the data is legally anonymized within the meaning of the applicable legal framework. Even if legal anonymization is not achieved (or even achievable), de-identification techniques are still useful to secure and minimize the amount of to-be-processed data. *See*, e.g., Kahled El Emam, *Guide to the De-Identification of Personal Health Information* (1st ed. 2013).

25 *See, e.g.*, CCPA, GLBA (Gramm-Leach-Bliley Act), CDPA (Consumer Data Privacy Act), among others.

26 *See, e.g.*, Article 29 Data Protection Working Party, Statement on the role of risk-based approach in data protection legal frameworks, Adopted on 30 May 2014, WP 218.

27 For example, HIPAA's safe harbor de-identification test. HIPAA's safe harbor requires the removal of a limited list of indirect identifiers, although it calls for the removal of a non-exhaustive list of direct identifiers. Therefore, it is possible by relying solely upon the HIPAA list of indirect identifiers, and depending upon the context, to expose patients to high re-identification risks. *See* Latanya Sweeney et. al, *Re-identification Risks in HIPAA Safe Harbor Data: A study of data from one environmental health study* (Aug. 28, 2017), https://pubmed.ncbi.nlm.nih.gov/30687852/; Victor Janmey, *Re-Identification Risk in HIPAA De-Identified Datasets: The MVA Attack* (Dec. 5, 2018), https://pubmed.ncbi.nlm.nih.gov/30815177/.

# Mapping Entity Tags to Classification Tags

When data classification is done systematically within an organization, each data type or attribute within a data source should be classified. In practice, this means that each attribute will be labeled or tagged. In order to facilitate monitoring and auditing, and ultimately ensuring the classification process is transparent, it is important to grasp the distinction between entity tags and classification tags.

It is best practice to tag attributes at least twice, with both entity tags and classification tags.

| | DEFINITIONS |
|---|---|
| **Entity tag** | Tags that detail what the data is and are applied irrespective of applicable regulatory frameworks (e.g., name, email address, IP address). Entity tags are true across all frameworks.[28] |
| **Classification tag** | Tags that express data classes (e.g., personal identifiers or sensitive attributes) and their related properties. These tags drive how the data should be treated, and are, at least in part, dependent upon the prescriptions of the applicable regulatory frameworks.[29] |
| **Classification rule** | Classification rules group entity tags in data classes as expressed by classification tags. For example, a classification rule could provide that attributes tagged as names should also be tagged as personal identifiers. The mapping of entity tags to classification tags is made possible by the formulation of classification rules. |

28 Their detection can be easily automated as they are highly dependent upon format. Organizations will typically specify the required format or patterns when data collection occurs to ensure tagging consistency. For example, a social security number in the United States is always a string of nine numbers. Similarly, a date is a combination of a two-digit day, two-digit month, and two or four-digit year, though the order may differ depending on geographic location.

29 Automating the detection of classification tags is a less straightforward process. The same is true for the attribution of class properties.

Let's insist upon a fourth recommendation:

| | RECOMMENDATION | RATIONALE |
|---|---|---|
| 04 | Attributes should be tagged twice, with both entity tags and classification tags. | When adopting a risk-based approach and focusing on individual harm, it is fair to state that the sensitivity of the data has a direct correlation with the intensity of the harm felt by individuals if their data is misused. However, the matter is further complicated by the fact that harm is also dependent upon the level of identifiability within the data.[30] What this means from a data classification perspective is that the combination of personal identifiers and sensitive information that should be restricted and carefully monitored first and foremost.[31] |

Three types of decisions matter for the purpose of attaching classification tags and class properties to data types within data sources:

# What should count as a personal identifier?

Identifiers are data types or attributes that are beneficial to an attacker in their attempt to identify or reidentify an individual. They have the following properties:

- Availability

  An attribute is considered available if an individual's value can be discovered using public or otherwise reasonably attainable information, such as telephone directories, social media, or voter registration rolls.

  For example, an address or gender is considered available, whereas a laboratory test may not be. Replicable attributes, which are those that are consistently associated with an individual, are more likely to be available. Such attributes could include telephone number, name, or a congenital medical condition, whereas a highly variable attribute, like body temperature, is not replicable and therefore generally not considered to be an available attribute.

- Distinguishability

  An attribute is considered distinguishable if its value can be used to discriminate between individuals.

  For example, in a domestic context, a country would not be distinguishable, since all individuals in consideration share the value of this attribute, whereas a postal code could be since only a very limited number of individuals share the same postal code in some jurisdictions.

30 This is explicitly acknowledged by the methodology developed by the French regulator, the Commission Nationale Informatique et Libertés (CNIL). CNIL explains that, "severity represents the magnitude of a risk. It essentially depends on the level of identification of personal data and the level of consequences of the potential impacts." *See* CNIL, Methodology for Privacy Risk Assessment (2012), https://www.dataguidance.com/sites/default/files/methodology_for_privacy_risk_management_how_to_implement_the_data_protection_act.pdf. More recent guidance is less detailed on this point. *See, e.g.*, CNIL, Privacy Impact Assessment Methodology (PIA) (2018), https://www.cnil.fr/sites/default/files/atoms/files/cnil-pia-1-en-methodology.pdf.

31 This is not to say that the processing of sensitive information in the absence of identifiers could not generate harm, but harm is likely to be collective rather than purely individual in this situation. See Sophie Stalla-Bourdillon et. al, Warning Signs: The Future of Privacy and Security in the Age of Machine Learning, Immuta Resources (Sept. 2019), https://www.immuta.com/resources/warning-signs-the-future-of-privacy-and-security-in-the-age-of-machine-learning/.

Data re-identification probability can be reduced by interfering with either availability or distinguishability. To this end, (potentially) identifying attributes can be categorized into two distinct groups[32]:

- Direct identifiers[33]

  These are available attributes that are uniquely associated with an individual. Since these attributes are unique, they are highly distinguishable.[34]

  Statistical approaches may refrain from considering availability and instead classify attributes based on uniqueness alone; this can only help privacy because it broadens the scope of what is considered to be a direct identifier. There are good reasons to approach classification in this manner, for instance as a hedge against the future where such attributes may become available, or where third parties may collude to link records. These include, but are not limited, to medical record numbers, social security numbers, and full facial images.

- Indirect identifiers

  These are available attributes that are not uniquely associated with an individual, but provide some level of distinguishability when attempting to identify an individual's record.

  Such attributes include, but are not limited to, postal code, gender, and age. When multiple indirect identifiers are combined, the resulting record could be highly specific and therefore highly distinguishable.

This brings us to our fifth recommendation:

| | RECOMMENDATION | RATIONALE |
|---|---|---|
| 05 | The data class **Personal Identifier** should be broken down into two subclasses: **Direct Identifiers** and **Indirect Identifiers**. | Direct and identifiers do not carry the same re-identification risks and could be treated with different security/privacy techniques. When both direct and indirect identifiers are treated in such a way that re-identification risks are deemed remote or very small, the remaining attributes usually cease to be classified as personal information or data.[35] |

32 It is important to note that the success of some real world attacks described in the scientific literature does not make the distinction between direct and indirect identifiers moot. De-identification methods remain essential risk mitigation tools, even if it has been demonstrated that the release-and-forget model is hardly a workable one. This is because there is an alternative to the release-and-forget model, i.e., the trusted environment model. The UK Health Data Research Alliance Paper (8 December 2021) 'Principles and Best Practices for Trusted Research Environments' sets out an approach for a trustworthy ecosystem of data access for health research building upon the UK Health Data Research Alliance Paper (21 July 2020) 'Trusted Research Environments (TRE) Green Paper.'

33 Here, we use information-theoretic notions for direct and indirect identifiers since these are relevant for a risk-based approach, which should be distinguished from the notion of direct identifier taken to produce a limited dataset under 45 C.F.R § 164.514(e)(2).

34 There is debate as to how to treat random strings replacing identifiers. As long as the random string is not used in conjunction with another identifier, there is an argument that such treatment should not be considered to be a direct identifier. This is because the string has no intrinsic or extrinsic meaning, except when it is used as a joint between different data sources.

35 It should be noted that both secondary legislation and case law suggest more or less clearly that the de-identification or anonymisation of personal information or data is a valid way to achieve deletion. See California Consumer Privacy Act Regulations, Cal. Code Regs. tit. 11, § 999.313(d)(2) (2020); Data Protection Authority [DPA] Dec. 5, 2018, 123.270/0009, https://www.ris.bka.gv.at/Dokumente/Dsk/DSBT_20181205_DSB_D123_270_0009_DSB_2018_00/DSBT_20181205_DSB_D123_270_0009_DSB_2018_00.html (Austria); Αρχή προστασίας δεδομένων προσωπικού χαρακτήρα [ΑΠΔΠΧ] [Hellenic Data Protection Authority] [HDPA] 06/2020 (Greece). ***Compare with*** Commission Nationale de l'Informatique et des Libertés [National Commission for Computing and Liberties], June 14, 2021,SAN-2021-008 (France), https://www.legifrance.gouv.fr/cnil/id/CNILTEXT000043668709 ("[T]he company does not delete the personal data but only deactivates the account in question, preventing the person from logging on and blocking the sending of commercial prospecting. The delegation thus noted the presence in the database of the personal data of a customer of the company (name, first name and e-mail address) who had previously made a request for deletion by e-mail. Access to his account had simply been deactivated."). ***See also*** the Federal Trade Commission (FTC)'s 2016 settlement with dating site AshleyMadison.com where the site was penalized for failing to fully delete customers' data even after the customer had paid for "Full Delete" rather than "Basic Deactivation." While this settlement was based on infringement of the prohibition on unfair or deceptive commercial practices rather than any data protection statutes, it highlights the expectations that data subjects have when they request deletion. F.T.C. v. Ruby Corp., No. 16-cv-02438 (D.D.C. Dec. 14, 2016).

Some frameworks have their own list of identifiers when they govern de-identification practices, such as the Health Insurance Portability and Accountability Act (HIPAA)[36]. Developing a framework-specific subclass of identifiers might thus make sense in some contexts.

# What should count as sensitive attributes?

There are various definitions for sensitive information, some of which can be extremely broad depending on whether the framework is vertical or horizontal.[37] For example, the CCPA is a horizontal framework, and its category of sensitive information comprises personal information that reveals a wide range of individual characteristics, such as racial or ethnic origin.[38] However, the CCPA limits the effect of sensitive information characterization by granting consumers a right to restrict the processing of sensitive information in limited circumstances, such as when the processing aims to infer characteristics about the consumer only.[39]

The US Federal Trade Commission (FTC) includes within its definition of sensitive information unique identifiers such as social security numbers, financial, health, children's, and geolocation information.[40]

The language of the GDPR, which is also an horizontal framework, is more homogeneous than that of the CCPA[41] but is also narrower.[42] The effect of the characterization of special category data is nonetheless much more significant. It reduces the range of justifications available by adding a second layer of legal bases.[43]

36 *See* 45 C.F.R. § 164.514(b)(2). HIPAA Safe Harbor lists 18 categories of identifiers, with the 18th category comprising unique identifying numbers, characteristics, or codes. While HIPAA's definition of direct identifiers is thus open, its list of indirect identifiers is closed, which is problematic in some contexts.

37 A framework is horizontal when it applies across sectors, unless an exception applies. A framework is vertical when it governs one sector, which usually implies a relatively narrow definition of the covered data and the covered entities. ***See, e.g.***, HIPAA, GLBA, or FERPA in the United States.

38 Note, however, that health information is defined slightly differently within CCPA. Cal. Civ Code § 1798.140(ae)(2) comprises "personal information collected and analyzed concerning a consumer's health."

39 Cal. Civ Code § 1798.121 (2020). For an interpretation of the category of health data under the GDPR *see* Case C-184/20, OT v. Vyriausioji tarnybinės etikos komisija, ECLI:EU:C:2022:601 (Aug. 1, 2022). *See also* Article 29 Data Protection Working Party, Annex to Letter to Mr. Timmers, 5 February 2015, https://ec.europa.eu/justice/article-29/documentation/other-document/files/2015/20150205_letter_art29wp_ec_health_data_after_plenary_annex_en.pdf ("Raw, relatively low privacy impact personal data can quickly change into health data when the dataset can be used to determine the health status of a person.")

40 *Protecting Consumer Privacy in an Era of Rapid Change*, Fed. Trade Comm'n (Mar. 2012), https://www.ftc.gov/sites/default/files/documents/reports/federal-trade-commission-report-protecting-consumer-privacy-era-rapid-change-recommendations/120326privacyreport.pdf.

41 GDPR, art. 9(1). Note the difference between the GDPR and the CCPA definition: political opinions are not considered to be sensitive under the CCPA framework.

42 GDPR, art. 9.

43 GDPR, art. 9. *See also* Article 29 Data Protection Working Party, Advice paper on special categories of data ("sensitive data"), April 2011, http://ec.europa.eu/justice/policies/privacy/docs/wpdocs/others/2011_04_20_letter_artwp_mme_le_bai l_directive_9546ec_annex1_en.pdf ("[M]isuse of these data in general, is likely to have more severe consequences for the individual's fundamental rights, such as the right to privacy and non-discrimination, than misuse of other, "less sensitive" types of personal data. Misuse of health data, including drawing incorrect or unreliable conclusions, may be irreversible and have long-term consequences for the individual as well as his or her social environment.").

This brings us to our sixth classification recommendation:

| RECOMMENDATION | RATIONALE |
|---|---|
| 06 The data class **Sensitive Attribute** should be broadly defined and then broken down into sub-classes. | There is a consensual view that the GDPR class of special category data excludes several sensitive data types. Therefore, it makes sense to adopt a broad definition of sensitive personal data and further break down this class by taking into account the main functions of the organization for which the data classification is being built. By way of example, when controllers are healthcare providers it's a good idea to distinguish between transactional or financial data, login details, and medical records. With this said, when the data originates from a healthcare provider, a sticky definition of health data will usually be needed to cover not only medical data but also information used by the entity to register the individual for healthcare services as well as payment data.[44] |

As mentioned above, identifiers and sensitive attributes can overlap. This is the case with social security numbers for example, which are unique and replicable, but are also mentioned within the CCPA's list of sensitive information.[45]

# Which property should be associated with which class?

There are two types of frameworks that are particularly relevant to answer this question (when they are not merged together): privacy and data protection frameworks, and breach notification frameworks. A breach notification obligation requires notifying the regulator and/or affected individuals when a data breach happens and meets a predetermined threshold. In comprehensive and horizontal data protection frameworks, breach notification obligations are often already included.[46] In other jurisdictions, usually where there is no horizontal privacy or data protection framework, breach notification obligations are found within bespoke frameworks.[47] Importantly, although breach notification laws are more limited in terms of their effects than horizontal privacy or data protection laws, they tend to adopt a broad definition of covered data, which usually largely overlaps with personal information.[48]

Many US frameworks have a provision specifically excluding publicly available information from the definition of "personal information"[49]. Still, it's important to remember that publicly available information can still

44 *See, e.g.*, GDPR, recital 35 ("Personal data concerning health should include all data pertaining to the health status of a data subject which reveal information relating to the past, current or future physical or mental health status of the data subject. This includes information about the natural person collected in the course of the registration for, or the provision of, health care services as referred to in Directive 2011/24/EU of the European Parliament and of the Council to that natural person; a number, symbol or particular assigned to a natural person to uniquely identify the natural person for health purposes; information derived from the testing or examination of a body part or bodily substance, including from genetic data and biological samples; and any information on, for example, a disease, disability, disease risk, medical history, clinical treatment or the physiological or biomedical state of the data subject independent of its source, for example from a physician or other health professional, a hospital, a medical device or an in vitro diagnostic test."). The status of payment data is less clear under the GDPR. Compare with HIPAA, 45 C.F.R § 164.514(e)(2).

45 Cal. Civ Code §§ 1798.140(ae) (2020).

46 *See, e.g.*, GDPR, art. 33, 34.

47 *See, e.g.*, US state laws such as Wyoming (Wyo. Stat. Ann. §§ 40-12-501, 40-12-502 (2015)), Arkansas (Ark. Code Ann. §§ 4-110-101 et seq. (2005)), Idaho (Idaho Code Ann. §§ 28-51-104 et seq. (West 2006)), among others.

48 *See, e.g.*, Colorado Breach Notification Law (Colo. Rev. Stat. Ann. § 6-1-716 et seq.); Missouri Breach Notification Law (Mo. Rev. Stat. § 407.1500); West Virginia Breach Notification Law (W. Va. Code Ann. §§ 46A-2A-101 et seq. (2008).)

49 *See, e.g.*, CCPA, § 1798.140(o)(2); 16 C.F.R. § 313.3(n-p) (2000); Colorado Privacy Act, Colo. Rev. Stat. §§ 6-1-1301 et seq. (2021).

generate risks depending on how it is used to infer other characteristics about an individual. For example, some churches in the US publish publicly available membership directories that can contain church members' names, phone numbers, physical addresses, and email addresses. Using this directory, an organization could

infer additional sensitive attributes, such as an individual's religious or philosophical beliefs, which could then be used to cause harm. More generally, both membership-based and attribute-based inference could still be a concern with publicly available information.[50] This explains why other frameworks do not exclude publicly available from their remit and still require some form of governance.[51]

Note that the FTC enforced action of its Health Breach Notification Rule for the first time in early 2023 against a digital health company that advertises, distributes, and sells health-related products and services directly to consumers, for unauthorized disclosures of consumers' personal health information to third parties.[52]

These bring us to three additional recommendations:

| | RECOMMENDATION | RATIONALE |
|---|---|---|
| 07 | Confidentiality should be broken down into several levels. At a minimum, data classes should be associated with a high level of confidentiality when they fall within the domain of breach notification obligations or disclosure reporting obligations. | Not all attributes are associated with the same level of confidentiality within an organization. Attributes and/or combination of attributes that fall within the domain of breach notification obligations should be considered confidential. The same should be true for attributes or combinations of attributes that, if disclosed, must trigger the production of disclosure reports.[53] |
| 08 | The number of confidentiality levels should depend upon the strength of the de-identification techniques that can be implemented upon the data. | In its standards for security categorization of federal information and information systems, NIST recommends distinguishing three confidentiality levels – low, moderate, and high.[54] However, when personal information is at stake, the number of levels usually depends upon one's conception to de-identification or anonymization, and the presence of sensitive attributes. If it is reasonable to assume that sensitive information can be effectively de-identified or anonymized without context controls, then three levels may suffice: low, medium, and high. If this assumption is unreasonable, or if de-identification or anonymization still require context controls, then four levels might be preferable: low, medium, restricted, and high. |

50 Sophie Stalla-Bourdillon & Alfred Rossi, *Aggregation, Synthesisation, and Anonymisation: A Call for a Risk-Based Assessment of Anonymisation Approaches*, in 13 Data Prot. and Privacy (Dara Hallinan, Ronald Leenes, & Paul De Hert eds., 2021).

51 *See, e.g.*, GDPR art. 2.

52 F.T.C. vs. Goodrx Holdings Inc., Case 3:23-cv-00460-DMR (filed 02/01/23).

53 *See, e.g.*, 45 C.F.R. § 164.528(a-d) (2022).

54 NIST, *supra* note 13.

## 09 Criticality should be broken down into several levels.

The level of criticality usually depends upon the needs of the business function for which the data classification is being built.[55] As with levels of confidentiality, NIST's standards for security categorization of federal information and information systems, recommends distinguishing three availability levels – low, moderate, and high.[56]

55 *See*, NIST's categorization of availability. *Id.*

56 NIST, *supra* note 13.

# Building Policy Tables

By associating data classes with properties, purposes, data treatments, and user attributes, it becomes possible to set the ground for operationalizing the principles of the least privilege, purpose limitation, and data minimization in a systematic fashion.

We recommend accomplishing this by building multi-dimensional policy tables, which make it possible to easily visualize both data classes and data treatment rules, and to quickly deduce the rationale for the data treatment rules. These tables should be composed of three innate data class dimensions[57] that are essential for the formulation of fine-grained data access rules, and at least one consequential dimension,[58] (i.e., data treatment), as illustrated in Table 2.

| DATA CLASS | CLASS PROPERTY | PROCESSING PURPOSE | USER ATTRIBUTE | TREATMENT |
|---|---|---|---|---|
| Direct Identifier | 1. Confidential - Medium<br>2. Confidential - High in combination with sensitive<br><br>Critical - High | Payment<br>Treatment<br>R&D* | Finance<br>Primary Care<br>Data Science<br>Data engineer | Clear<br>Clear<br>Masked |
| Indirect Identifier | 1. Confidential - Medium<br>2. Confidential - High in combination with sensitive<br><br>Critical - Medium | Payment<br>Treatment<br>R&D* | Finance<br>Primary Care<br>Data Science | Masked<br>Clear<br>Clear |
| Sensitive Attribute | 1. Confidential - Restricted<br>2. Confidential High in combination with Identifiers<br><br>Critical - High | Payment<br>Treatment<br>R&D* | Finance<br>Primary Care<br>Data Science | Masked<br>Clear<br>Clear |

Table 2. An exemplar of a policy table

57 An innate data class dimension is independent from other dimensions.

58 A consequential or acquired data class dimension is a dimension that is derived from innate dimensions.

*To comply with regulations such as GDPR, R&D will have to be broken down into more specific sub-purposes, which should make the consequences for the data subjects clear.

We have already covered class properties in previous sections.

- ### Processing Purposes

  Data risks are directly dependent upon processing purposes. Expressing processing purposes makes it possible to tailor the amount of data to the data consumer's needs. The more granular the purpose, the more granular the implementation of the least privilege, purpose limitation, and data minimization principles.

  What both best practice and guidance show is that it is useful to build a hierarchy of purposes to define data needs and associated related processing impact.[59]

- ### Data Consumers' Attributes

  Although not every organization is structured identically, it is likely that users' job descriptions will vary within an organization, and not all data consumers will be required to perform the same set of processing activities on the data. In fact, what the data mesh approach mandates is that each data domain expresses its own needs.[60] Therefore, it is often useful to map processing purposes with user attributes to ensure data consumers are able to perform a relevant set of purposes that are generally aligned with their job descriptions or more broadly aligned with the business function (or data domain in which their position is located). User attributes are therefore a useful dimension to refine the operationalization of the least privilege, and the data minimization and purpose limitation principles.

- ### Data Treatment

  A data treatment is a measure or set of measures that impacts how the data appears to a data consumer, such as how many columns and rows are available to the data consumer or whether the attributes have been transformed through masking techniques. Data treatment is dependent upon the class property, processing purpose, and user attributes. It is thus a consequential dimension and a direct expression of the least privilege, purpose limitation, and data minimization.

  The choice of the data treatment represents a tradeoff between data protection and data utility.

It is possible to further complete policy tables by specifying additional consequential dimensions, such as restrictions concerning data consumption and communication exchange tools or tool treatment. For example, one internal rule could be that sensitive attributes should never be shared via instant messaging. The same is true with data consumption actions, such as read, write, and delete. Each data consumer could also be associated with a list of data actions that are dependent upon the processing purposes and associated user attributes. For example, an internal rule could state that users with the attribute 'Data Science' are only able to read data, whereas users with the attribute 'Data Engineer' are able to write data.

59 Article 29 Data Protection Working Party, *Opinion 03/2013 on Purpose Limitation*, Article 29 Data Protection Working Party, p. 19 (April 2, 2013), https://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2013/wp203_en.pdf. ("For 'related' processing operations, the concept of an overall purpose, under whose umbrella a number of separate processing operations take place, can be useful. That said, controllers should avoid identifying only one broad purpose in order to justify various further processing activities which are in fact only remotely related to the actual initial purpose.").

60 *See* Zhamak Dehghani, *How to Move Beyond a Monolithic Data Lake to a Distributed Data Mesh*, martinFowler.com (May 20, 2019), https://martinfowler.com/articles/data-monolith-to-mesh.html.

Thus, here our last recommendation:

| | RECOMMENDATION | RATIONALE |
|---|---|---|
| 10 | Multi-dimensional policy tables should be built to operationalize the principles of least privilege, purpose limitation, and data minimization. | Data classification must be both granular and simple enough to drive enforceable fine-grained data access control and usage rules. Policy tables are a means to achieve this goal. |

# Conclusions

For data classification to really add value to the building of a controlled environment, it should initially be processing system agnostic. Data classification should also be granular enough to enable the formulation of fine-grained access control and usage rules, while being simple enough to easily draw the overall picture of data consumption use cases. This is a must-do to operationalize core requirements of data protection frameworks and/or standards, such as the least privilege, purpose limitation, and data minimization principles.

In this white paper, we set forth a four-dimensional data classification primitive for producing comprehensive fine-grained access control and usage rules that can be easily represented within policy tables, and made ten recommendations to data security and privacy teams ready to tackle the data classification challenge. These provide a strong foundation from which to start a comprehensive data protection journey, so organizations can unlock more value from their data without compromising its security and privacy.

Schedule a demo with our team today.

## About Immuta

Immuta enables organizations to unlock value from their cloud data by protecting it and providing secure access. The Immuta Data Security Platform provides sensitive data discovery, security and access control, data activity monitoring, and has deep integrations with the leading cloud data platforms. Immuta is now trusted by Fortune 500 companies and government agencies around the world to secure their data. Founded in 2015, Immuta is headquartered in Boston, MA.

IMMUTA

25 Thomson Place, 4th Floor, Boston, MA 02210 immuta.com (800) 655-0982