IMMUTA

WHITE PAPER

A Guide to the EU's Unclear Anonymization Standards

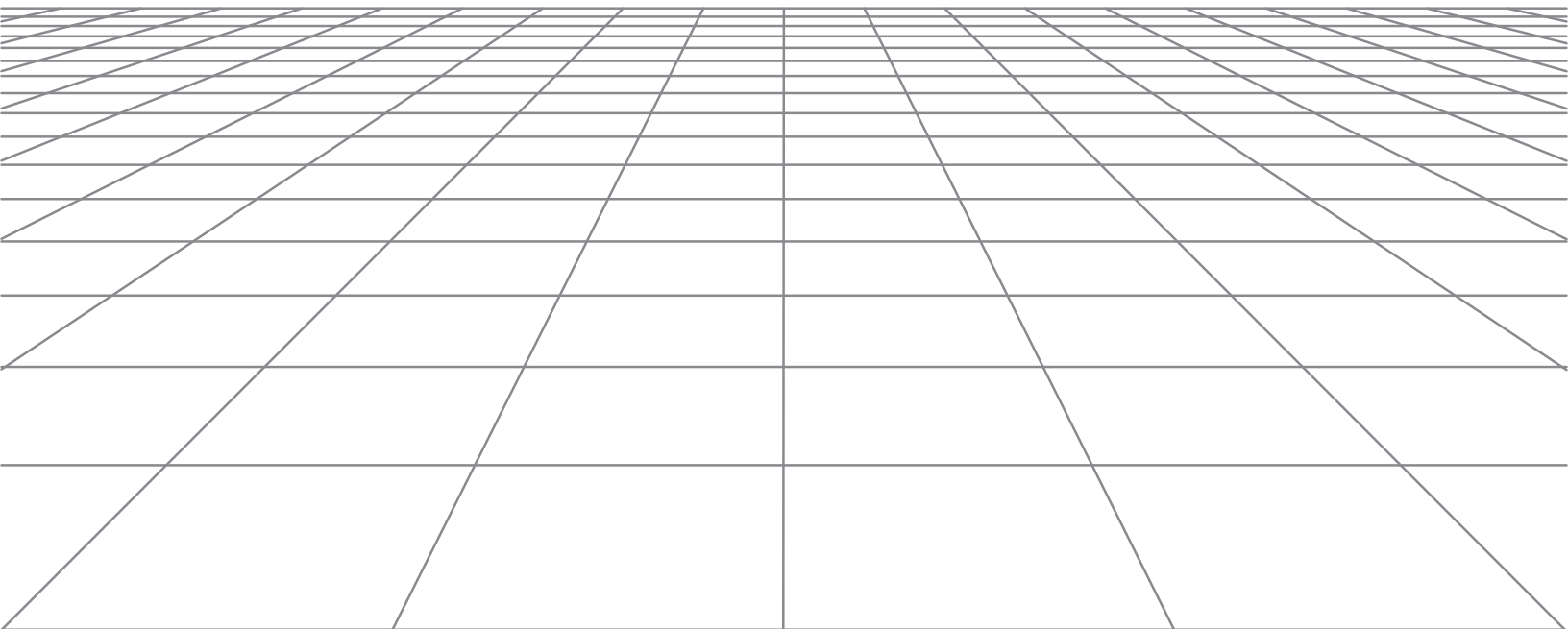


Table of Contents

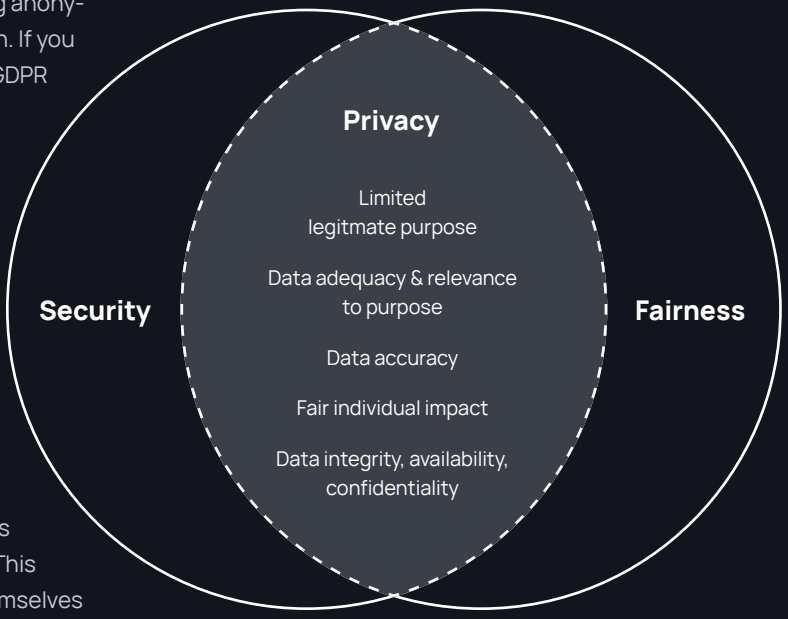
Introduction

The EU General Data Protection Regulation is among the most influential data privacy laws in the world - setting the standard, in many ways, for how global organizations implement their data privacy programs. However, the GDPR itself, and EU data protection laws more generally, suffer from one central problem: one of their most important provisions is unclear.

Specifically, the GDPR defines anonymous data as data which "does not relate to an identified or identifiable natural person or to personal data rendered anonymous" such that "the data subject is not or no longer identifiable." Data that meets this criteria is therefore not subject to the GDPR, making anonymous data the holy grail of data protection. If you can anonymize data, regulations like the GDPR simply no longer apply - not their onerous requirements on handling data, not even their very high fines. From a compliance standpoint, anonymous data makes your life easier.

The problem is that even though the GDPR specifically calls out anonymous data, and even though European data protection authorities (DPAs) have been publicly talking about anonymization for decades, it's unclear that anyone - regulators included - really knows what "anonymization" means in practice. This is something that even the regulators themselves seem to have acknowledged, with the Spanish DPA and the European Data Protection Supervisor (EDPS) releasing a joint document recently entitled "10 misunderstandings related to anonymization" aimed at clarifying these exact issues.

So why do DPAs find it hard to converge on a clear and uniform anonymization standard? Because anonymization only protects the confidentiality of data and EU data protection law is about much more than ensuring confidentiality. EU data protection laws are focused on placing individuals in control of their data and preventing illegitimate uses of that data. Therefore, the DPAs are not acting irrationally – their mission is just highly complex and broad. The danger, however, is that by adopting overly restrictive approaches to anonymization, data protection laws negatively impact legitimate activities and become unenforceable over time.



In the big picture, this uncertainty persists and can leave organizations attempting to anonymize their data in a deep bind, even when they don't equate anonymization to a free license. Our goal in this white paper is to outline why the state of anonymization remains so uncertain, and what organizations can actually do about it as they seek to anonymize their data.

Conflicting Regulatory Guidance in the EU

Let's start with the uncertainty around what "anonymization" truly means under EU data protection standards. Even though the GDPR discusses anonymization in Recital 26, we have to go back to opinions issued by the Working Party 29, an official body made up of representatives of EU DPAs, the European Data Protection Supervisor, and the European Commission (for the full context, the Working Party has been replaced by the European Data Protection Board, which is expected to issue guidance on the matter but has not done so yet).

In 2007, the Working Party 29 issued an opinion that clearly articulated the difference between "anonymization" on the one hand and "pseudonymization" on the other. The main difference between the two came down to re-identification. While pseudonymization is privacy-protective – for example, key-coding data so that identifiers are masked but can be uncovered if needed – it is also technically reversible. On the other hand, anonymization was defined in the following way: "disguising identities can also be done in a way that no re-identification is possible, e.g. by one-way cryptography, which creates in general anonymised data."

The problem was apparent even in 2007, because defining what is "possible" involves predicting the future, and therefore indicates some amount of uncertainty. Could developments in quantum computing, to take one extreme example, render many current cryptographic standards obsolete? It's possible. The question of possibility is a matter of risk tolerance, and in 2007 the Working Party 29 erred on the side of flexibility, writing that as long as "appropriate technical measures" have been put in place to prevent re-identification of data, that data can be considered anonymous. Here's the exact language from the 2007 opinion:

> Even if identification of certain data subjects may take place despite all those protocols and measures (due to unforeseeable circumstances such as accidental matching of qualities of the data subject that reveal his/her identity),the information processed by the original controller may not be considered to relate to identified or identifiable individuals taking account of all the means likely reasonably to be used by the controller or by any other person. Its processing may thus not be subject to the provisions of the Directive.

This line of reasoning mirrors the similar standards for anonymization that other regulatory frameworks have adopted around the world - allowing for some consequent level of risk of re-identification. The Federal Trade Commission, for example, has said essentially the same thing about reasonable risks of re-identification, as has the state-level California Privacy Rights Act (CPRA) in California. It also echoes the Health Insurance Portability and Accountability Act (HIPAA) de-identification standards, which created a de-identification method known as "expert determination" that explicitly allows for a very low chance of re-identification risk.[1] Because perfect anonymization is in many ways impossible, some risk of re-identification will always remain. The 2007 opinion was, in other words, reasonable and in line with other anonymization standards.

But then came the Working Party's 2014 opinion on anonymization techniques, which turned this analysis on its head and set the path for significant confusion about EU anonymization standards that exists to this day. In particular, the Working Party revisited the difference between anonymization and pseudonymization, and declared that a "specific pitfall is to consider pseudonymised data to be equivalent to anonymised data." Pseudonymity continues to allow for identifiability, the Working Party wrote, and "therefore stays inside the scope of the legal regime of data protection." To be fair, the Working Party 29 was reacting to a series of public anonymization mishaps, such as the AOL and the Netflix cases, in which data was supposedly anonymized and released in large numbers, only to find later that the data was identifiable after all.

The difference between anonymization and pseudonymization, in the new analysis, lay in the likelihood of re-identifiability - whether it's possible to derive personal information from de-identified data. However, as study after study has demonstrated, it's pretty much impossible to perfectly anonymize data, meaning some possibility of re-identification frequently remains. So how should organizations determine what is likely?

The Working Party 29 enumerated three specific re-identification risks:

- **Singling out**, or the ability to locate an individual's record within a data set.
- **Linkability**, or the ability to link two records pertaining to the same individual or group of individuals.
- **Inference**, or the ability to confidently guess or estimate values using other information.

The Working Party 29 stated that an anonymization solution that protected against each of these risks "would be robust against re-identification performed by the most likely and reasonable means the data controller and any third party may employ." In other words, an anonymization that protects against each of these three risks – singling out, linkability, and inference – would be satisfactory. (Note, however, that the Working Party was not exactly saying that passing this three-pronged test is a requirement in all circumstances, as we will explain in further detail below). If an organization could explain how their anonymization efforts prevented singling out, linkability, and inference, their anonymization solutions would therefore stand up to regulatory scrutiny.

1 Note that unlike other standards, HIPAA's standard is only concerned with the anticipated recipient, making the scope of anonymization somewhat narrower – it was not, in other words, addressed to all potential recipients of the data over an unlimited period of time.

So far, so good. The problem emerged, however, when the Working Party 29 went on to suggest that both aggregation and destruction of the raw data were also needed to ensure no reasonable risk of re-identification remained. Here's their exact language:

> It is critical to understand that when a data controller does not delete the original (identifiable) data at event-level, and the data controller hands over part of this dataset (for example after removal or masking of identifiable data), the resulting dataset is still personal data. Only if the data controller would aggregate the data to a level where the individual events are no longer identifiable, the resulting dataset can be qualified as anonymous.

In other words, only by aggregating data into group statistics and permanently deleting the original data could organizations have full confidence that their data is anonymized and therefore falls outside the scope of data protection regulations in the EU.[2]

Due to this U-turn, EU regulators continue to vacillate between the 2007 and 2014 standards to this day. Some have stated that some residual risk of re-identification is acceptable, so long as the right precautions are in place. Regulators like the UK's Information Commissioner's Office, or ICO (when the UK was part of the EU), took this track, as did the Irish DPA and others. But other DPAs, like France's Commission Nationale de l'Informatique et des Libertés (CNIL), have used a more absolutist language in their guidance.[3]

Organizations attempting to comply with these confusing standards and meet EU anonymization requirements are therefore stuck between a rock and a hard place. So what can organizations do?

2 If you find this paragraph confusing, you're not alone. Indeed, it is difficult to make sense of this paragraph because it appears fundamentally inconsistent with other parts of the very same opinion. The Working Party 29, for example, also stated in the same opinion that "data controllers should consider that an anonymized dataset can still present residual risks to data subjects. Indeed, on the one hand, anonymization and re-identification are active fields of research and new discoveries are regularly published, and on the other hand even anonymized data, like statistics, may be used to enrich existing profiles of individuals, thus creating new data protection issues."

3 One potential reason for this difference was that the French translation of the Data Protection Directive was missing the word "reasonable," making the French standard more restrictive than the EU standard.

01 Give Up and Embrace Pseudonymization

The first option is to give up on the project of anonymizing data entirely, and simply consider all de-identified data as pseudonymous. While pseudonymous data does not fall outside the scope of EU data protections because re-identification is still possible, the compliance burden on pseudonymous data can be significantly lighter, assuming: the processing purpose is legitimate; a legal basis is established (or the secondary purpose is considered to be compatible with the initial purpose); and the data controller is not in a position to identify individuals (making most individual rights virtually non-existent, except the rights to information and to object).

Standards for how to implement pseudonymization techniques vary, but many overlap with anonymization practices under other legal frameworks outside the EU. Here, for example, is how the European Union Agency for Cybersecurity describes pseudonymization techniques:

> The choice of a pseudonymisation technique and policy depends on different parameters, primarily the data protection level and the utility of the pseudonymised dataset (that the pseudonymisation entity wishes to achieve). In terms of protection, as discussed in the previous sections, RNG, message authentication codes and encryption are stronger techniques as they thwart by design exhaustive search, dictionary search and guesswork. Still, utility requirements might lead the pseudonymisation entity towards a combination of different approaches or variations of a selected approach. Similarly, with regard to pseudonymisation policies, fully- randomized pseudonymisation offers the best protection level but prevents any comparison between databases. Document-randomized and deterministic functions provide utility but allow linkability between records. Specific solutions might be applicable, depending on the identifiers that need to be pseudonymised.

If this seems a lot like anonymization to you, you're not alone. The more EU regulators have attempted to clarify the difference between pseudonymization and anonymization at a technical level, the closer each ends up seeming to the other. These similarities at a technical level, combined with the 2014 guidance from Working Party 29, have led some organizations to give up on anonymization entirely, at least until EU DPAs provide further clarity – an entirely reasonable response given the significant confusion.

That said, applying EU data protection standards to all types of pseudonymous data, irrespective of the strength of the pseudonymization process, can be problematic when data needs to be accessed quickly and shared among different types of stakeholders. Those who are using pseudonymous data for their own purposes are "controllers" under EU data protection laws and have to make sure they tick all the right boxes before processing the data. It's also unclear how pseudonymization can help justify data transfers to third countries with no adequacy decisions. Researchers in the medical space especially, such as this group, have been quite public about the problems this causes.

So if organizations aren't willing to give up on anonymization entirely, what else can they do? They have a few options.

02 Argue the Risks of Re-identification Are Sufficiently Remote

The next option lies in arguing that the means of re-identification are not reasonably likely. This would mean relying more heavily on the Working Party 29's 2007 opinion than on its 2014 opinion, or at least ignoring the most problematic paragraphs of the 2014 opinion and highlighting the following statement: "Whenever a proposal does not meet one of the criteria [i.e., singling out, linkability, inference], a thorough evaluation of the re-identification risks should be performed."

Some call this line of argument a "risk-based approach," which acknowledges that some risk of re-identification is acceptable even for anonymized data. The question becomes how can organizations argue that even though risk remains for re-identification, it is sufficiently remote and therefore their data is anonymous?

For starters, the 2014 Working Party guidance itself refers to the importance of context, stating that "account must be taken of 'all' the means 'likely reasonably' to be used for identification by the controller and third parties, paying special attention to what has lately become, in the current state of technology, 'likely reasonably' (given the increase in computational power and tools available)."

More fundamentally, a risk-based approach implies adopting an attacker-centric definition of anonymization, which appears compatible with the legal test. Indeed, the legal test will focus on assessing the re-identification means reasonably likely to be used by the controller or another person, i.e, an attacker. In order to anticipate attackers' behavior, de-identification experts rely upon risk models to guide their selection of data and context controls.

Other regulators have suggested that they are aligned to this exact approach. In the healthcare data context, for example, the European Medicines Agency has acknowledged that a risk-based approach is a valid alternative to the mitigation of the three re-identification risks, writing that:

> "According to the Opinion 05/2014 on anonymization techniques of the Art. 29 WP, two options are available to establish if the data is anonymized. Either through the demonstration of effective anonymization based on three criteria: [1] Possibility to single out an individual. [2] Possibility to link records relating to an individual. [3] Whether information can be inferred concerning an individual . . . or, whenever a proposal does not meet one of these criteria, through an evaluation of the identification risks." [4]

The French DPA CNIL seems to be somewhat open to this approach as well.

If you're interested in examining how these attacks work, you can read another article we've written on a call for risk-based assessments. At a minimum, the goal should be that no situationally-relevant attacker is reasonably in a position to re-identify the individuals within the transformed data set.

As if things weren't complicated enough, when producing the list of situationally-relevant attackers, EU DPAs don't always seem to know what to make of the difference between the position of the data controller and that of third parties. This shows that there is confusion about what a risk-based approach really is, and not all regulators are well versed in de-identification methods.

More specifically, some have mandated that "anonymisation procedures must ensure that not even the data controller is capable of re-identifying the data holders in an anonymised file," as the Spanish DPA and the EDPS have argued. Others, like the Irish Data Protection Commission, have written that "the anonymisation process [should] prevent the singling out of an individual data subject, even to someone in possession of the source data."

What, then, should we make of the controller's ability to re-identify the data? There are a few options, although not all of them are consistent with a risk-based approach. The Working Party 29's view expressed in 2014 seems clear: a data controller should not be considered an attacker. We should thus conclude that what really matters is that the data controller is not in a position to single out individuals within the supposedly anonymized data by using the raw data or publicly available information only. In other words, we should assume that the data controller has performed a state-of-the-art data transformation technique to treat both direct and indirect identifiers. Notably, this interpretation does not necessarily exclude anonymization for internal purposes.

It's worth noting that the ICO has developed a specific test, called "the motivated intruder test," to be used for assessing anonymization standards. This test states that, as a rule of thumb when appropriate process firewalls are in place, the attacker is not an insider, has no prior knowledge, and is not an expert (although it is acknowledged that the motivated intruder can be made more sophisticated depending upon the use case at hand).

4 It is worth noting that, in the healthcare space, this line of argument is easier to make because while the data is circulating freely, all the organizations using that data are usually under a confidentiality obligation. It is therefore easier to impose some limitations on what is reasonably likely.

03 Trusted Third Parties

The next option is to rely on what are called "trusted third parties," or TTPs, which can help serve as intermediaries between organizations possessing the raw data and those who seek to use anonymous data. Specifically, when one party wants to share anonymous data with a secondary organization, a trusted third party can "broker" the exchange, particularly when linking data sets could temporarily increase re-identification risks. By definition, the trusted third party is not considered to be a situationally-relevant attacker and is in charge of implementing de-identification techniques on the raw data, which remains under the control of the original party while sharing the de-identified data with the secondary organization.

In 2013, the Working Party 29 addressed this arrangement in an opinion on purpose restrictions, and described trusted third parties as operating "in situations where a number of organisations each want to anonymise the personal data they hold for use in a collaborative project," which can be used "to link datasets from separate organisations, and then create anonymised records for researchers." Introducing a third party to perform the de-identification and to keep the raw data separate, the Working Party 29 suggested, seems to be another useful method to achieve anonymization.

The 2013 opinion went on to described the possibility of something called "complete anonymisation":

> "In many situations, anonymisation may help public sector bodies comply with data protection law whilst at the same time enabling them to make the necessary data available for reuse. Indeed, when this is possible, 'complete' anonymisation (and a high level of aggregation) of personal data is the most definitive solution to minimize the risks of inadvertent disclosure."

One year earlier, the UK DPA itself described this type of arrangement as "particularly effective where a number of organisations each want to anonymise the personal data they hold for use in a collaborative project." Indeed, the ICO went to great lengths to describe how this approach enables anonymization:

"A trusted third party is an organisation which can be used to convert personal data into an anonymised form. . . .The personal data can then be anonymised in 'safe', high security conditions and to an agreed specification – allowing the subsequent linkage of anonymised individual-level data, for example. The great advantage of a TTP arrangement is that it allows social science research to take place – for example using anonymised data derived from health and criminal justice records – without the organisations involved ever having access to each others' personal data. Security, anonymisation and anti-re-identification measures taken by the TTP should be covered on agreement."

Inserting a third party into the de-identification process is therefore one central way to bolster the claim to anonymity, and can facilitate the creation of anonymized data when data from different sources is being linked together. The inclusion of a TTP is thus a key component of a broader risk-based approach.

04 Emerging Technological Approaches

All of the anonymization options we list above are heavy on processes – assessing all the reasonably likely ways data could be re-identified, as in the risk-based approach, or inserting third parties to manage the data, as in the use of trusted third parties. There is, however, one additional method, which relies on a group of emerging technologies to help automate the de-identification process and streamline and accelerate the risk-based approach. It's worth noting that because these technologies are still emerging, and in many senses still being proven in real-world settings, organizations often lack the resources needed to integrate these technologies within their processing practices and demonstrate that they support anonymization under the EU frameworks. That said, there are clear signs these technologies might stand up to regulatory scrutiny.

Take, for example, synthetic data, which consists of creating new data from a sample set of data and preserving the correlations within the sample set without recreating any direct identifiers. The use of synthetic data has been growing in the healthcare space in particular, offering a promising way to extract trends from health data without directly using patient identifiers. Indeed, one such solution has even been designated as anonymous data under GDPR standards by the CNIL. The technology is still in its infancy and does not necessarily eliminate all re-identification risks, so it remains to be seen how useful it can be in real-world settings; however, EU DPAs do seem open to labeling such data anonymous under data protection standards.

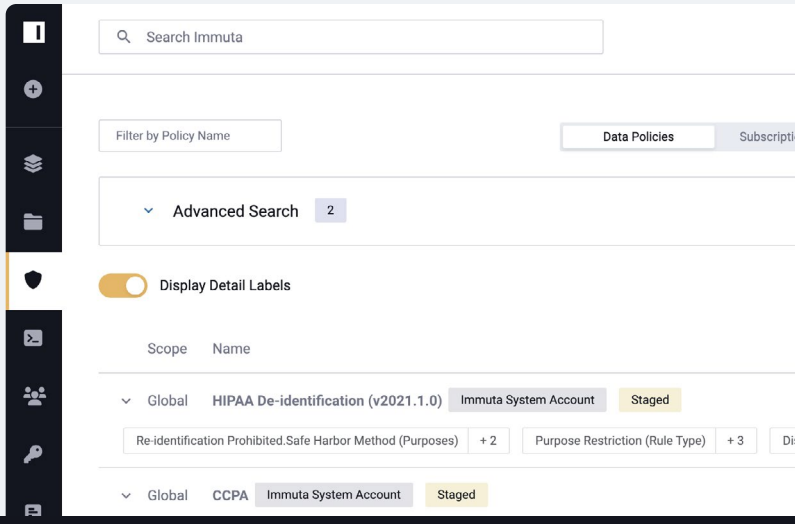
Differential privacy is a mathematical privacy framework that holds promise for anonymization. This method inserts controlled randomization into a data analysis process, resulting in mathematically guaranteed limits on the amount of personal information inferrable by any attacker. (For a more thorough overview of differential privacy, see here.) While EU DPAs have yet to formally opine on differential privacy, we believe they're likely to look favorably upon the technique, which the US Census Bureau now uses to protect the privacy of respondents' data. Indeed, differential privacy is currently being litigated in Alabama over concerns that uncertain census accounting may result in unequal congressional representation, in what we believe is likely to generate clearcut legal precedent for the technique's guarantees. One amicus brief in support of the US government's use of differential privacy, filed by the non-profit EPIC, even goes so far as to declare that "differential privacy is the only reliable technique for defeating current and future re-identification attacks."

Synthetic data and differential privacy are not the only techniques with a promise for anonymization: some tout the benefits of federated learning, which, if implemented with an eye towards compliance, can serve a similar function as trusted third parties, although in practice it tends to be used as a data minimization technique rather than an anonymization technique. Although there are often significant technical barriers in practice and deployment, another technique known as secure multi-party computation can be used to design multi-party data processing protocols that simulate use of a trusted third-party without actually having one. In that sense, the future holds new opportunities for meeting the requirements of EU data protection laws, even if such mandates remain unclear.

What to Do?

Absent further clarifications from EU regulatory authorities themselves, there is no one-size-fits-all approach to anonymization for organizations seeking to comply with EU data protection standards. That said, there are a host of concrete options - and clear arguments - these organizations can use to maximize their data's value while protecting its confidentiality. While there is much cause for confusion, in other words, there's also cause for optimism.

To start transforming the way you access and share your sensitive data Request a demo today.



About Immuta

Immuta is the **market leader in Data Access**, providing data teams one universal platform to control access to analytical data sets in the cloud. Only Immuta can automate access to data by discovering, protecting, and monitoring data. Data-driven organizations around the world trust Immuta to speed time to data, safely share more data with more users, and mitigate the risk of data leaks and breaches. Founded in 2015, Immuta is headquartered in Boston, MA.

IMMUTA

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