privacy response team
Blaeu

Wilhelminaplein 1-40
3072 DE ROTTERDAM

(Rob van Eijk, PhD
rob@blaeu.com

European Data Protection Board
Rue Montoyer 30
1047 Brussels
Belgium

Date 30 October 2026
Reference Guidelines 02/2026 on Anonymisation
Subject Public consultation

Dear Mrs. Talus, Mrs. Vereecken,

I write in a personal capacity, as an expert working where law and technology meet. My comments on the draft guidelines concern how the flowchart in Annex 1 works in practice.

Before the EDPB published the guidelines, Benjamin de Vanssay and I worked out five use cases. They are set out in our report 'From Pseudonymised to Anonymous Data' (SSRN 6972998), which forms part of this submission as an annex. Five European companies supplied the descriptions. The cases cover the following.

- Scrubbed prompts sent to a generative AI provider that retains nothing
- Vehicle data for live traffic maps
- Hashed data with a secret salt for training anomaly detection
- Aggregated audience figures for an online publisher
- Ad models trained on synthetic data, generated from group counts with differential privacy

The publication of the guidelines led me to test the cases myself against the 'simplified approach' and the 'contextual approach'. That test shows where controllers lack firm guidance, and it is the subject of my submission.

1. The contextual approach works in practice and confirms the conclusions of our report. In my view, the data is anonymous for the intended recipient in four cases. In the case with hashed data, it remains personal data. The guidelines sharpen the analysis. Contracts only complement technical measures (paragraph 34), the EDPB advises against weighing a lack of motivation (paragraph 31), and on my reading of paragraph 36 a single demographic cell with too few people can turn a whole publication into personal data**. Two references in the flowchart are wrong, and I ask the EDPB to correct them. (a) Paragraph 104 does not**

exist, and (b) the step on violating only No Record Isolation, through which two cases pass, belongs to paragraphs 98–100.

2. The simplified approach rejects four of the five cases, because it disregards the recipient's means. The Court of Justice made those means decisive in the SRB judgment (C-413/23 P). Any data architecture that relies on the recipient's perspective therefore fails. This is consistent with paragraph 48 and footnote 36, but the flowchart does not reflect it. Answering yes to 'Is it obvious that the (extended) data is not anonymous?' leads straight to 'Consider data as personal'. It skips the move to the contextual approach that paragraphs 49 and 96 allow, and that the flowchart does offer at the bottom of the branch. Two cases end there. **I ask the EDPB to add that move, and to align paragraph 49 ('possible in theory') and the summary ('in the abstract') with paragraphs 48, 100, and 103.**

3. The case with differential privacy is the only case that fully passes the simplified approach. That technique limits what the data reveals about any one person, whoever the attacker is. This fits paragraph 71(b) and footnote 43, but the guidelines never mention differential privacy. Privacy budgets, minimum cell sizes, and thresholds for the number of contributors are also missing, although three of the five conclusions depend on those parameters. The European Medicines Agency (EMA) and Health Canada show that such parameters are workable. For clinical data released to the public, they recommend a maximum re-identification risk of 9 per cent, [A] which Health Canada equates to a minimum of eleven patients per cell. [B] **I would like the EDPB to recognise formal privacy guarantees as evidence under paragraph 71(b), with guidance on privacy budgets, how they add up across releases, and minimum numbers of people per cell.**

4. A party's role can matter more than its safeguards. Under paragraph 15, data remains personal data for a party that processes it on a controller's behalf, whatever that party can see. The more strictly an agreement obliges an AI provider to retain nothing, the more that provider resembles such a processor. The safeguard that protects the data can then keep it personal. **I ask the EDPB to explain how paragraph 15 relates to the SRB judgment.** There, the Court holds that pseudonymised data need not be personal data for Deloitte, which the SRB had engaged as an independent valuer, provided two conditions are met (paragraphs 21 and 77 of the judgment). The Court left unexamined the EDPS's argument that Deloitte was bound to the SRB by contract (paragraphs 93 and 116).

I offer the five cases as a set of practical examples for the final guidelines. They also show where the limits lie. Placing a secret salt with a third party does not help if the controller itself keeps the source database (case 3, paragraphs 20, 28, 34, and 60). This corrects a solution our report proposed.

Thank you for taking the time to read my observations.

Cordialement,
Rob van Eijk, PhD

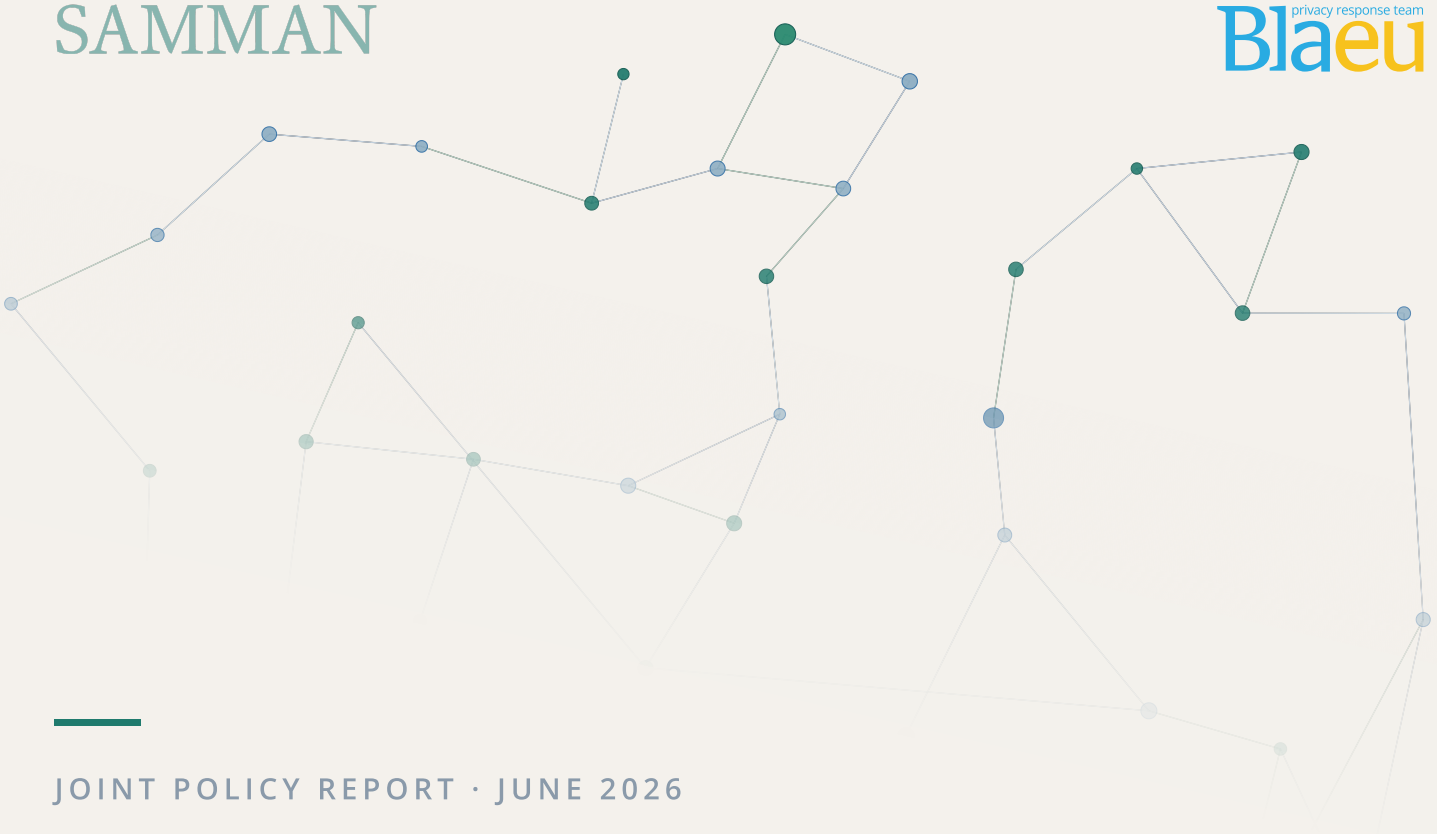
Notes

[A]: European Medicines Agency, *Anonymisation Report Form Instructions* (EMA/482639/2023, 21 May 2024) 6 https://www.ema.europa.eu/system/files/documents/other/anonymisation-form-instructions-en.pdf accessed 30 October 2026. See also European Medicines Agency, *European Medicines Agency policy on publication of clinical data for medicinal products for human use* (Policy 0070, EMA/144064/2019, 21 March 2019) https://www.ema.europa.eu/en/documents/other/policy-70-european-medicines-agency-policy-publication-clinical-data-medicinal-products-human-use_en.pdf accessed 30 October 2026.

[B]: Health Canada, *Public Release of Clinical Information: guidance document* (version 1.0, 12 March 2019) s 6.2 https://www.canada.ca/en/health-canada/services/drug-health-product-review-approval/profile-public-release-clinical-information-guidance/document.html accessed 30 October 2026. See also ISO/IEC 27559:2022 *Information security, cybersecurity and privacy protection – Privacy enhancing data de-identification framework* (ISO 2022).

SAMMAN

Blaeu privacy response team



JOINT POLICY REPORT · JUNE 2026

From Pseudonymised to Anonymous Data

A practical assessment framework under the SRB ruling, applied to real use cases provided by European companies

AUTHORS

Dr. Rob van Eijk · Team Blaeu

Benjamin de Vanssay · SAMMAN Law & Corporate Affairs

ABOUT THE FIRMS

SAMMAN Law & Corporate Affairs · CABINET-SAMMAN.COM

SAMMAN Law & Corporate Affairs is a Paris and Brussels law firm advising clients on French and European public and regulatory affairs. The firm operates at the intersection of law, policy, and institutional strategy, covering legislative, regulatory, and political dimensions across sectors.

Team Blaeu · BLAEU.COM

Founded in 2019, Team Blaeu is a Dutch advisory firm specialising in EU data protection, digital markets regulation, and artificial intelligence regulation. Team Blaeu helps organisations navigate GDPR compliance, AI Act governance, and enforcement proceedings at the national and European levels.

CONTENTS

Table of contents

About the firms 02
Preamble 05
1 Introduction 06
1.1 Pseudonymisation under the absolute approach to personal data 06
1.2 Pseudonymisation under the relative approach through the Court of Justice ruling 07
1.3 Pseudonymisation in practice and the need for evidence-based guidance 08
2 From the SRB Ruling and DPA Guidance to a Practical Assessment Framework 09
3 Use Case 1 — Scrubbed Prompts to a Generative AI Provider Under Zero Retention 11
3.1 How the Zero-Data-Retention Agreement Works 12
3.2 Identifier Removal and Zero Retention Block Singling Out and Linkage 12
3.3 Why Training Data Cannot Serve as a Re-Identification Route 13
3.4 Scrubbed Prompts Under Zero Retention 13
4 Use Case 2 — Building Live Traffic Maps from Anonymised Vehicle Signals 15
4.1 How Vehicles Create Floating Car Data 15
4.2 Random Batching and Gap Lengths Hide Where Journeys Begin and End 16
4.3 Time Randomisation and Backend Deletion Break the Data Trail 16
4.4 Converting GPS Points into Road Segments Removes Precise Location 17
4.5 Swarm Aggregation Dissolves Individual Vehicles into Traffic Patterns 17
4.6 Road Conditions Survive the Pipeline, Individual Drivers Do Not 17
5 Use Case 3 — Intelligent Anomaly Detection Features in Cloud Business Software 20
5.1 The Limitations of Standard Privacy Enhancing Technologies (PETs) 20
5.2 Salted Hashing and Strict Environmental Separation 21
5.3 The Patterns Reach Training, but the Controller Still Holds the Key 21
6 Use Case 4 — Delivering Aggregated Audience Statistics to a Publishing Platform 23
6.1 How the Measurement Provider Collects and Processes Data 23
6.2 Aggregation as the Main Privacy Safeguard 24
6.3 Four Convergent Barriers Place the Reports Outside GDPR Scope for the Platform 26
7 Use Case 5 — Training Ad Prediction Models Without Individual User Records 28
7.1 Replacing Direct Identifiers with Hashed or Random IDs 29
7.2 Pooling Records into Group Counts Removes Individual Data 29
7.3 Adding Statistical Noise Protects Each Person's Contribution 29
7.4 Reconstructing Training Signals from Group Counts Avoids User-Level Data 29
7.5 The Synthetic Dataset Under the Two-Step Assessment 30
8 Conclusions — What the Use Cases Show About Pseudonymous Data in Practice 32
8.1 What the Five Use Cases Share 32
8.2 Looking Ahead 32

Glossary of Technical Terms 34

Appendices 36

A CJEU SRB Litigation 36

A1 Single Resolution Board v European Data Protection Supervisor 36

A2 European Data Protection Supervisor v Single Resolution Board 36

A3 AG Opinion 36

B CJEU Case Law 37

B1 Scarlet Extended SA v SABAM 37

B2 Patrick Breyer v Bundesrepublik Deutschland 37

B3 Peter Nowak v Data Protection Commissioner 37

B4 Gesamtverband Autoteile-Handel e.V. v Scania CV AB 37

B5 OC v European Commission 37

B6 IAB Europe v Gegevensbeschermingsautoriteit 38

C Relevant GDPR Definitions and Recitals 38

C1 Recital 26 GDPR 38

C2 Article 4(1) GDPR — Personal Data 38

C3 Article 4(5) GDPR — Pseudonymisation 38

C4 Article 25 GDPR — Data Protection by Design and by Default 38

PREAMBLE

Preamble

The SRB ruling of the Court of Justice of the European Union,[1] has renewed the long-standing debate on the boundaries of the General Data Protection Regulation (GDPR). It confirmed that identifiability must be assessed by reference to "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."[2] That assessment is recipient-specific and therefore contextual. This report builds on that ruling and applies a more demanding standard where guidance and case law diverge, taking a precautionary approach.

Legal certainty requires clear criteria for determining when pseudonymised data become anonymous data. This study proposes a practical and evidence-based contribution to that assessment.

It draws on use cases provided by leading European industry players across the automotive, software, entertainment, legal tech, and advertising sectors, following extensive exchanges with technologists. These cases detail the technical architecture through which companies implement pseudonymisation in practice, and show the layers of safeguards the companies apply against re-identification.

Taken together, they offer a concrete starting point for the practical and consistent application of legal requirements, moving beyond purely theoretical risk assessments. By identifying recurring patterns and effective safeguards from real life use cases, the study lays the foundation for a more structured and predictable evaluation of anonymisation techniques. It also opens the way for certification and the creation of a standardised assessment framework based on clear, operational, and verifiable criteria.[3]

This report is not legal advice. The authors are privacy practitioners, not a supervisory authority. The descriptions of the use cases are based on the state of technology at the time of writing.

1. Case C-413/23 P European Data Protection Supervisor v Single Resolution Board EU:C:2025:645.
2. Recital 26 GDPR.
3. Article 25(3) GDPR, Article 42 GDPR.

SECTION 01

1 · Introduction

Almost a decade after entering into force on 24 May 2016 and applying since 25 May 2018, the General Data Protection Regulation (GDPR) stands as one of the most ambitious and influential regulatory initiatives of the digital age. By adopting a principle-based, risk-based and accountability-driven framework, the GDPR sought to adapt the 95/46 Directive to the modern data uses of organisations, keeping strong safeguard for individuals, while preserving the free flow of data within the European Union. Stakeholders widely agree that its underlying philosophy and regulatory architecture remain relevant in today's rapid pace of innovation.

Yet, despite this regulatory leadership, the European Union is inexorably lagging in data-driven innovation, particularly in Artificial Intelligence (AI). As highlighted in recent policy debates, including at the highest political level,[4] Europe's competitiveness is constrained by a complex, fragmented regulatory environment that generates legal uncertainty.

Much of that uncertainty does not come from within the GDPR itself but more from its rigid interpretation by data protection authorities (DPAs) that have stayed away from defining the contours of the core concept of personal data. The question is basic but contested: Since non-personal data is defined negatively, as any data not qualifying as personal data under the GDPR, what counts as non-personal data and when does the GDPR stop applying? Leaving this question unanswered paralyzes the application of all regulation encouraging the use and sharing of non-personal data, such as the Data Act and the Digital Markets Act.

1.1 Pseudonymisation under the absolute approach to personal data

Pseudonymisation sits in an awkward position in EU data protection law. EU law widely recognises it as a key risk-mitigation tool and as an enabler for the safe processing of personal data. The GDPR repeatedly endorses it as a means to reduce risks, strengthen security, operationalise data protection by design, and support core principles such as data minimisation and purpose limitation. Several EU acts echo this recognition.[5] The GDPR does not however exempt pseudonymised data from any of the GDPR obligations. It is only when data is anonymised that the obligations of the GDPR do not apply. But unfortunately, anonymisation strips datasets of their value for further use. In practice, however, DPAs have failed to clearly help organisations understand the different criteria that would enable distinguishing pseudonymised from anonymised data. Above all, they have generally treated data as personal whenever potential re-identification of an individual is theoretically possible (absolute approach to personal data), even where controllers have no realistic legal or technical means to identify individuals and have deployed robust safeguards. This interpretation, grounded in guidance by DPAs, has grown at the opposite of the GDPR risk-based approach.

4. In his report on European Competitiveness, Mario Draghi points to regulatory complexity and fragmentation as major structural drags on EU competitiveness, generating legal uncertainty, raising compliance costs, and deterring investments in innovative sectors. Mario Draghi, 'The Future of European Competitiveness – A Competitiveness Strategy for Europe' (European Commission, 2024) https://commission.europa.eu/topics/competitiveness/draghi-report_en#paragraph_47059 accessed 10 June 2026.

5. Pseudonymisation is currently a key data processing enabler across seven major EU legal instruments: the GDPR, Data Act, Data Governance Act and Digital Services Act (responsible data sharing and reuse); the Digital Markets Act (platform interoperability and data exchanges); the NIS2 Directive (cybersecurity of information systems); and the AI Act (training of high-risk AI systems).

Three simultaneous policy developments compound this uncertainty. First, the Article 29 Working Party, the predecessor of today's EDPB, issued Opinion 05/2014 on Anonymisation Techniques (WP216, hereafter the 2014 anonymisation opinion),[6] and that opinion is overdue for an update. Second, the EDPB finalised its Guidelines 01/2025 on Pseudonymisation. Third, the ongoing Digital omnibus trilogue may revise the definition of personal data itself. Taken together, these developments undermine businesses' ability to reuse data for innovative purposes with legal confidence.

As a result, organisations apply full GDPR obligations regardless of actual risk, such as registers of processing, data protection impact assessments, data processing agreements, transparency, and consent requirements. This unnecessarily raises compliance costs and diverts engineering resources.

This ultimately constrains data reuse and innovation, as the qualification of data as personal significantly limits controllers' ability to repurpose data they already hold. And in many cases the research exemption would not apply, and further processing for research purposes would not be considered compatible with original purposes. Where data is processed on the basis of legitimate interests, further use requires a balancing test. Consent is the harder case. There, a new purpose requires obtaining new consent.

This is all the more impactful for European companies, especially tech companies who have more limited resources than their non-European competitors.

1.2 Pseudonymisation under the relative approach through the Court of Justice ruling

The Court of Justice of the European Union (CJEU) provided some relief in the SRB ruling.[7] The ruling clarified that the concept of personal data has clear limits. Pseudonymised data may, in certain circumstances, effectively prevent the identification of individuals and qualify as anonymous. In SRB itself, however, the Court found the data personal from the SRB's perspective as data controller because the SRB held all information needed to re-identify the authors of the comments. The ruling's significance lies in the principle it established for recipients of data. Yet, the CJEU decision also indicates that the risks of identifying individuals must be evaluated on a case-by-case basis, leaving considerable room for interpretation by DPAs and courts.

This Court's decision represents a unique opportunity to clarify the concept of pseudonymisation and anonymisation, and to ensure that controllers can benefit from clear criteria enabling them to identify which data fall under the GDPR and which do not. At the same time, it preserves a high level of protection for personal data while having a more pragmatic approach with data protection obligations, by excluding their application in situations where an actor has no means reasonably likely to be used to identify the individual.

Additionally, the Commission's proposal to define harmonised criteria for determining when pseudonymised data no longer qualifies as personal data pursuant to the CJEU case law would, in the same way, improve legal certainty.[8] It would also ensure better GDPR harmonisation by avoiding divergent interpretations of the SRB ruling by DPAs, national courts, and other regulatory authorities dealing with data pseudonymisation (competition, consumer, information, or telecom authorities for instance).

6. Article 29 Working Party, Opinion 05/2014 on Anonymisation Techniques (WP216, 10 April 2014). The Working Party brought together the national supervisory authorities under Article 29 of Directive 95/46/EC, and the EDPB succeeded it on 25 May 2018 under Article 68 GDPR. The EDPB endorsed a set of Working Party guidelines in its Endorsement 1/2018, but WP216 was not among them.
7. SRB Ruling (Case C-413/23 P).
8. Commission, 'Digital Omnibus' COM(2025) 837 final, art 41a. Proposed Article 41a.

1.3 Pseudonymisation in practice and the need for evidence-based guidance

This report aims to support the case for stronger anonymisation, including the development of Privacy Enhancing Technologies as complementary tools for lawful data reuse. It does so through a pragmatic and evidence-based methodology aimed at providing practical guidance. The goal is to define with greater precision where the GDPR applies and where it does not so as to ease the burden of compliance and enable European entities to reuse data with more confidence.

Harmful surveillance practices based on intrusive data processing activities fall squarely within the scope of the GDPR as well as relevant other EU and national laws, regardless of the technical steps that accompany them. The framework (hereafter: Assessment Framework) in this report is not a tool for reclassifying such practices as outside the GDPR. They warrant stronger enforcement, not less.

Pseudonymisation covers a wide range of techniques, each offering different levels of reducing re-identification risk. Applying a single pseudonymisation technique is not sufficient. When remaining data is rich enough to single out a person, we argue that companies need a layered approach. That is the lesson this report takes seriously.

This paper seeks to contribute detailed practitioner-provided case descriptions to show that the reliance on pseudonymised data brings real benefits to organisations, innovation, and society at large while keeping residual risks low and documented. These use cases should benefit from the CJEU longstanding case law that personal data is a relative concept and in particular the SRB ruling that pseudonymised data can sometimes fall outside of the scope of the GDPR. Our aim is not to suggest that the likelihood of re-identification is without foundation, but to encourage a policy debate grounded in concrete positive evidence rather than theoretical and unsupported claims.

Drawing on concrete use cases across sectors provided by European companies, with particular attention to AI development and data-intensive operational needs, it outlines how pseudonymisation operates in practice, showing how it reduces risks, and preserves data utility. It enables lawful data reuse while meeting the CJEU standards of effectively preventing the identification of individuals and therefore falling outside the scope of the GDPR.

This report aims to help the relevant institutions (the Commission, the EDPB, and any other relevant stakeholders) define the harmonised criteria for determining when pseudonymised data no longer qualifies as personal data.

This report came together by consulting European companies. The use cases are not purely theoretical. Each company contributed a description of a real project. Each use case is situated in a specific context and any assessment requires a clear reference point. To allow for consistent evaluation across use cases, we briefly introduce the Assessment Framework below. It provides a set of practical criteria for evaluating the robustness of pseudonymisation measures.

We present five use cases. Together they span legal technology, automotive, enterprise software (two cases), media measurement, and advertising technology. The first covers Sending Scrubbed Prompts to a Generative AI Provider Under Zero Retention (Use Case 1, Section 3). The second covers Building Live Traffic Maps from Anonymised Vehicle Signals (Use Case 2, Section 4). The third covers Development of Intelligent Anomaly Detection Features in Cloud Business Software Using Relational Data (Use Case 3, Section 5). The fourth covers Delivering Aggregated Audience Statistics to a Publishing Platform (Use Case 4, Section 6). The fifth covers Training Ad Prediction Models Without Individual User Records (Use Case 5, Section 7). In Section 8, we close with conclusions.

SECTION 02

2 · From the SRB Ruling and DPA guidance to a Practical Assessment Framework

Pseudonymisation primarily relates to the technical and organisational measures that put data protection principles of data protection by design and by default into effect under Articles 25(1) and 25(2) GDPR. Anonymisation goes to the definition of personal data in Article 4(1) GDPR. When data is anonymised and leaves the data subject not or no longer identifiable, the GDPR does not apply. Therefore, the threshold for rendering data anonymous and out of scope of the GDPR is much higher than for pseudonymisation. The two concepts, personal data and anonymous data, carry distinct functions and distinct legal consequences. Recital 28 reinforces the point. It states that "The application of pseudonymisation to personal data can reduce the risks to the data subjects concerned and help controllers and processors to meet their data-protection obligations. The explicit introduction of 'pseudonymisation' in this Regulation is not intended to preclude any other measures of data protection".

Article 4(5) describes a relationship, not a fixed property of the data. "'Pseudonymisation' means 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".[9] Pseudonymisation presupposes that the additional information needed to re-attribute the data (still) exists, held separately, and under technical and organisational controls. Pseudonymous data can therefore be personal data for the party that holds that additional information and anonymous for a recipient who does not and who has no means reasonably likely to obtain it. This is the heart of the SRB ruling. The Court held that pseudonymised data are not personal in all cases and for every party, and that identifiability turns on the means available to the specific recipient.[10] Our framework prioritises that ruling as the governing legal test. We then add the stricter bar defined in the 2014 anonymisation opinion on top, as a precautionary layer, for the reasons set out below.

The SRB ruling also sets the operative test for a recipient, and it attaches two conditions. The Court held that pseudonymised data transferred to a recipient may be non-personal for that recipient. First, the recipient must not be in a position to lift the measures during its processing of the data. Second, those measures must in fact prevent the recipient from attributing the data to a natural person, including by cross-checking against other factors.[11] Both conditions have to hold. A measure that the recipient can undo on its own, or that leaves a route to identification through other data, fails the test. We apply these two conditions to every use case.

We assess anonymisation in the hands of a recipient against seven criteria. The first six are mainly based on the 2014 anonymisation opinion and ask how hard re-identification would be in practice. The seventh is based on the SRB ruling

9. See also Appendix C3.

10. Case C-413/23 P, EU:C:2025:645, paras 75 to 77 and 86.

11. Case C-413/23 P, para 77. The Court set two cumulative conditions for the recipient. "As regards Deloitte, to which the SRB transmitted pseudonymised comments, the technical and organisational measures referred to in Article 3(6) of Regulation 2018/1725 may, as the SRB essentially submits, have the effect that, for that company, those comments are not personal in nature. However, that presupposes, first, that Deloitte is not in a position to lift those measures during any processing of the comments which is carried out under its control. Second, those measures must in fact be such as to prevent Deloitte from attributing those comments to the data subject including by recourse to other means of identification such as cross-checking with other factors, in such a way that, for the company, the person concerned is not or is no longer identifiable."

and asks which means are reasonably likely to be used, by the recipient or by another person, which is the question Recital 26 GDPR directs the assessment to answer.

- Datasets containing pseudonymised data should be kept separate from directly identifying data about a natural person.[12]
- Anonymisation should be irreversible for the recipient. The framework judges irreversibility from the recipient's position, not in the abstract. Where the recipient cannot lift the measures during its processing and holds no means reasonably likely to reverse them, the data are irreversible for that recipient, even where the controller keeps the means to reverse them for its own purposes.[13]
- The likelihood of re-identifying a natural person should be insignificantly low.[14]
- The likelihood of singling out a natural person should be low. A purely theoretical possibility of singling out does not meet this threshold.[15]
- The likelihood of linking records relating to a natural person, in a dataset or across datasets, should be negligible.[16]
- The likelihood of inferring, with significant probability, information about a natural person from the data should be negligible.[17]
- The assessment should account for the means reasonably likely to be used, weighing the costs of identification, the time it requires, the technology available at the time of processing, and technological developments, including the availability of machine-readable datasets, generative AI, and agentic AI.[18]

The SRB and the Breyer ruling both ask whether specific means of re-identification are reasonably likely to be used by a specific recipient.[19] In Breyer, the Court found that dynamic IP addresses were personal data for the controller because it held *legal* means to obtain identifying information from a third party. The reasoning runs both ways. Where legal access to that information is clearly not available, even though the information exists somewhere, the data are not personal for that recipient. We remark that the CJEU does not work with probability thresholds, technical risk scores, or residual risk metrics.

One limit follows directly from Recital 26 and from Paragraph 85 of the SRB ruling. The assessment is recipient-specific, and the verdict belongs to the recipient, not to the data. A finding of anonymous holds for one recipient in one context. It does not survive an onward transfer to a party that does hold means reasonably likely to identify the data subject. Recital 26 asks about the means available to the controller or to another person, and the ruling confirms that data which are not personal for one recipient may still be personal in the hands of another. Each new recipient needs its own assessment.

12. Article 4(5) GDPR.
13. See the 2014 anonymisation opinion. See also, Case C-413/23 P, para 77, on irreversibility assessed from the recipient's position.
14. See the SRB Ruling (Case C-413/23 P), para 82, “In addition, the Court has previously held that a means of identifying the data subject is not reasonably likely to be used where the risk of identification appears in reality to be insignificant, in that the identification of that data subject is prohibited by law or impossible in practice (…)”. And Case C-413/23 P, Opinion of Advocate General Spielmann, para 57, “Thus, it is only where the risk of identification is non-existent or insignificant that data can legally escape classification as ‘personal data’.“
15. Singling out, linkability, and inference are the key concepts of the 2014 anonymisation opinion.
16. ibid
17. ibid
18. Article 4(5) GDPR, Case C-413/23 P, paras 79-87, and the Glossary.
19. Case C-582/14 Patrick Breyer v Bundesrepublik Deutschland EU:C:2016:779.

The 2014 anonymisation opinion treats irreversibility as a requirement and sets singling out, linkability, and inference as separate tests. It does not mandate a particular technical method, but it asks the controller to satisfy each test on the evidence. We add this stricter bar on top of the SRB test as a matter of *precaution*. In practice the framework runs in two steps. First, we apply the SRB test. If the recipient holds means reasonably likely to identify the data subject, the data are personal, and the assessment ends there. Where a case falls short at Step 1, we may still remark on a Step 2 criterion that teaches something useful. Such remarks are asides. They carry no weight in the verdict. Second, where the SRB test points to anonymous, we apply the 2014 anonymisation opinion bar before we accept that result. Data that pass the SRB test but fail the 2014 anonymisation opinion bar do not clear our framework. We treat them as most likely not anonymous, because precaution should decide the close cases against a reduction in scope.

This matters for the results that follow. The framework is built to return a negative verdict, and in some use cases it does. Strong measures do not settle the question on their own. In some instances, where reversibility survives in the controller's hands and the recipient sits close to the means, or where residual context in free text keeps an inference route open, the precautionary bar places the data on the personal side of the line despite the measures applied. We mark those cases as most likely not anonymous and explain why in the relevant sections. We believe these use cases are of particular importance since they contribute to the stated goal of laying the foundation for a more structured and predictable evaluation of anonymisation techniques.

USE CASE 1

3 · Use Case 1 — Sending Scrubbed Prompts to a Generative AI Provider Under Zero Retention

ASSESSMENT

Assessed under the framework in Section 2, the scrubbed prompts clear both steps. We believe they are most likely anonymous in the hands of the generative AI provider as the recipient, while they stay personal for the legal intelligence provider that holds the accounts.

Case Summary. *This case was provided by a European legal intelligence provider in support of privacy-by-design AI chatbots adoption. When professional users interact with its legal chatbot, their prompts may contain (sensitive) personal data. The use case shows how technical safeguards can remove identifiers from those prompts and ensure they are deleted immediately after processing. As a result, the AI provider has no realistic way to know who sent the prompt or to link it to an individual. That holds for this provider only, and for prompts, not attachments. Treating such data as non-personal, where these safeguards are effective, can support privacy-by-design AI adoption in Europe, especially for professional services that rely on external AI models.*

The organisation is a legal intelligence provider and offers a legal chatbot service to professional users. As part of its service, the Organisation transmits user-generated prompts to frontier generative AI models (the most capable available models). Before sending the prompt, the Organisation scrubs all direct identifiers from the

prompt content. Prompts carry no name, email address, user ID, or other unique identifier. They contain only the substance of a legal or factual question.

This use case examines how a legal intelligence service acts as a pseudonymising data controller when transmitting user prompts to frontier generative AI models.

3.1 How the Zero-Data-Retention Agreement Works

The Organisation applies a zero-data-retention (ZDR) agreement to all generative AI providers. Prompts are processed in volatile memory (temporary working memory that disappears when processing ends) and are not stored, logged, indexed, or retained beyond the duration of inference. The Organisation also contractually and technically disables all provider-side features that could enable re-use or derivation of value from prompt content, including abuse monitoring pipelines, product improvement processes, and model training.

The standard terms of any generative AI platform make clear that providers of AI models may reserve the right to scan prompts for, e.g., content safety, log session metadata for billing, cache prompt-response pairs to improve the user experience, and collect interaction data for training. Sometimes content is routed for human review. The point is that access to the AI model is more than the model itself. The model is part of an AI architecture provided by the platform hosting it.

The ZDR agreement blocks these channels. It does not stop at a contractual prohibition. The provider must enforce the restrictions through technical controls at infrastructure level, so a system update or a renegotiated contract cannot silently reverse them. An audit right or an independent attestation of the provider's retention controls would let the Organisation verify that enforcement, and we regard such a clause as good practice for any ZDR agreement.

3.2 Identifier Removal and Zero Retention Block Singling Out and Linkage

The prompts received by the generative AI provider contain no direct identifiers and are not retained beyond the inference process in the frontier AI model. This combination structurally prevents two of the three core re-identification vectors (singling out and linkability). The third (inference, or re-identification through the model's training data) is addressed in section 3.3.

Singling out requires an adversary to combine quasi-identifiers,[20] to narrow a dataset to a single individual. The prompts transmitted by the Organisation carry no identifiers and concern, by design, legal and factual subject matter rather than personal information about the user of the service. No quasi-identifier is available from which to single out a natural person.

A fair question follows. What if the same person sends similar prompts to other services that leave identifiers in place? Someone could then try to match those named prompts to the stripped ones, using the words in each prompt and the time it arrived. Put that way, the content and the timing begin to act like quasi-identifiers, even though the prompt itself carries no name or account.

We believe that match may still fail for this provider. Since it holds none of those outside logs and has no lawful way to get them, and it keeps nothing of its own, there is no stored prompt to line up against anything. With only one side of the pair, there is nothing to cross-match.

Linkability requires access to stored records that can be cross-referenced against external datasets. The ZDR policy eliminates the data accumulation layer on which linkage attacks structurally depend. Without persistent data, there is nothing to link.

20. See Glossary.

3.3 Why Training Data Cannot Serve as a Re-Identification Route

Re-identification techniques often rely on data accumulation, cross-referencing, and pattern recognition across stored datasets. In the absence of any persistent data, these techniques cannot be applied. There is one more attack vector to consider. The prompt still exists while it moves from the Organisation to the provider. An attacker on the network could try to capture it during this transfer. To prevent this, the Organisation should send every prompt over an encrypted channel, so no one can read the data in transit. This is a security measure and not a question of identifiability, but it covers the only moment that zero retention does not.

Another question arises from the fact that generative AI providers operate models trained on vast datasets, which may include publicly available legal content. One might ask whether such training data could serve as a vector for re-identification (for instance, by correlating a query's substance or style with identifiable individuals whose writings or cases appear in the training corpus). Technically, large language models do not work like a database lookup that returns a stored record. The real question is whether an attacker could rebuild a removed identifier from the words left around it. Recent research shows this may be possible.[21] Legally, any deliberate use of training data for re-identification purposes would fall entirely outside the contractual scope agreed with the Organisation's platform service.

3.4 Scrubbed Prompts Under Zero Retention

We apply the Framework in two steps. We run the SRB test first, then the 2014 anonymisation opinion precautionary bar. Only a case that clears both reaches a verdict of most likely anonymous.

Step 1, the SRB test. The SRB ruling sets two cumulative conditions for the recipient, here the generative AI provider. The first asks whether the provider can lift the measures during its processing. We believe it cannot. The Organisation strips all direct identifiers before the prompt leaves its systems, and the ZDR agreement requires the provider to enforce zero retention through technical controls at the infrastructure level, not through a contract term alone. The Organisation cannot inspect the provider's infrastructure, so this part of the assessment rests on the provider honouring that obligation. The provider never receives the identifying information, so it cannot restore what it never held. The second condition asks whether the measures in fact prevent the provider from attributing the prompt to a natural person, including by cross-checking against other factors. We believe they do. The prompt carries no name, no email address, and no user ID, and the provider holds no accumulated dataset against which to cross-check. Re-identifying a user from an isolated, identifier-free prompt processed without retention would demand disproportionate computational and investigative effort. Both conditions hold, so in our view the SRB test points to anonymous for this recipient.

Step 2, the 2014 anonymisation opinion precautionary bar. Passing the SRB test is not the end of our assessment. We then test the data against the stricter 2014 anonymisation opinion criteria before we accept the result.

The data clear both steps. We therefore believe the scrubbed prompts are most likely anonymous in the hands of the generative AI provider, within the meaning of Recital 26 GDPR.

We remark to keep three limitations in mind. First, the assessment covers prompt-based processing only. It excludes attachments, which may carry content of a different nature and may need their own identifiability

21. Lucas Georges Gabriel Charpentier and Pierre Lison, 'Re-identification of De-identified Documents with Autoregressive Infilling' (2025) Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) 1192. The study recovered up to 80% of the masked details in court rulings and other documents. It did so by matching the remaining text against a large store of outside knowledge. This provider has no such store, and it keeps no prompt, so the attack has nothing to work with.

assessment. A scanned document or photo run through Optical Character Recognition, a tool that turns images of text into machine-readable text, could hold personal data that prompt-scrubbing never sees. Second, the verdict is recipient-specific. It holds for this provider under this agreement, and it does not apply to a recipient that does hold means reasonably likely to identify the sender. Third, scrubbing removes direct identifiers only. The substance of a question can still carry details that point to a person. This provider cannot act on those details, because no prompt survives inference and it holds nothing to compare a prompt with.

USE CASE 1 — SRB FRAMEWORK ASSESSMENT	Case C-413/23 P
CRITERION	ASSESSMENT
Separation	The Organisation strips all direct identifiers before sharing the prompt. The provider receives only prompt content and holds none of the identity data retained by the Organisation. The separation is structural, not merely organisational.
Irreversibility	The provider processes prompts in volatile memory and discards them at the end of inference. No persistent record survives at any point in the chain. With nothing to reverse, the architecture itself enforces irreversibility, and that holds for the provider rather than resting on a key kept elsewhere.
Singling out	Singling out needs a dataset to narrow. The zero-retention policy means no prompt survives beyond the duration of inference. Even a prompt with detailed factual content about a specific matter cannot single out its sender, because the provider keeps no dataset in which to run that operation.
Linkability	No records persist beyond inference. Without stored records, cross-dataset linkage is not possible in practice.
Inference	Large language models do not work as retrieval systems, and no one can reverse-engineer their statistical weights to attribute a prompt to a specific person. The related membership inference risk,[22] which tests whether a prompt formed part of the training data, also fails here. The agreement prohibits the provider, by contract and by technical control, from using prompts for training or model improvement. These prompts never enter the training corpus, so there is no membership to infer.

22. See Glossary.

USE CASE 2

4 · Use Case 2 — Building Live Traffic Maps from Anonymised Vehicle Signals

ASSESSMENT

Under the Section 2 framework, the floating car data clear both steps in the hands of the third-party provider as the recipient, where the minimum contributor threshold runs. Sparse-road data that escape that threshold may remain personal data.

Summary. *This case was provided by a European automotive player in support of boosting the uptake of safer and smarter mobility services. Connected vehicles can generate data on road conditions, traffic and safety risks, such as black ice, which may initially include location and vehicle-related signals. The use case shows how short-lived pseudonyms, random batching, time delays, deletion of raw data and aggregation across many vehicles can prevent the third-party provider from linking the data back to a specific vehicle or driver. As a result, the provider receives information about road segments and traffic patterns, not individual journeys. Treating such data as non-personal, where these safeguards are effective, can support safer and smarter mobility services in Europe while ensuring that the system sees the road, not the driver. On quiet roads with few vehicles, a minimum contributor threshold suppresses the data point, and data that escape that threshold may remain personal.*

This use case examines whether Floating Car Data (FCD) constitutes personal data in the hands of a third-party provider that receives it. FCD is an established term in EU policy and industry, covering sensor data that moving vehicles collect and contribute to a shared pool. The concept features in discussions around the European Mobility Data Space.

For example, a driver approaching a patch of black ice cannot see it until it's too late. Millions of connected vehicles address this gap. They continuously share friction and performance data across fleet pools, supporting real-time hazard detection, and predictive maintenance. The same vehicles map road surfaces and traffic conditions as they drive, feeding services that fixed roadside infrastructure could never cover at the same scale.

Before sending the data, the system removes direct identifiers and replaces them with short-lived pseudonyms. Data aggregates across multiple vehicles per data point. Temporal randomisation and early deletion of raw data do the rest. At scale, overlapping observations create a dense data cloud in which individual vehicles become indistinguishable, making attribution practically infeasible.

4.1 How Vehicles Create Floating Car Data

Each vehicle in the fleet captures data continuously while moving. Onboard cameras and sensors record timestamps, GPS positions, and surface conditions, such

as road quality and friction data. The vehicle transmits this data to the manufacturer's backend system within seconds of capture.

The data uses a hashed vehicle identifier (a hash of the Vehicle Identification Number). Hashing is a mathematical way of scrambling a piece of information into a fixed-length string of letters and numbers that cannot be easily reversed.[23] The VIN is a unique 17-character code that identifies each vehicle as it leaves the production line. The hash equivalent of a VIN, e.g., 'WVWZZZ1JZXW000001', is a string '38657c765e00f5c9f487540df0aa7b36e08a0ec6db94de4658cb47eacca32ff3'. The manufacturer's backend receives data containing this hashed identifier and can link packets from the same vehicle for up to 30 minutes. A contractually binding job instruction governs what staff may do with that data during the window, adding an organisational control on top of the technical one. After that window closes, the manufacturer deletes the hashed identifier from the data. No link between the data and the originating vehicle remains.

Another privacy control worth mentioning is that individual drivers can deactivate data sharing from inside the vehicle, giving them direct control over their participation.

4.2 Random Batching and Gap Lengths Hide Where Journeys Begin and End

The vehicle does not stream data continuously. The manufacturer's backend instructs the in-vehicle system to group data into batches, with deliberate random gaps between batches.

Randomisation works as follows. Speed determines the ranges. At lower urban speeds, batches span between two and five kilometres, with average gaps of around five hundred metres. At motorway speeds above, e.g., 110 kilometres per hour, batches extend to between eight and twelve kilometres, with gaps of over one and a half kilometres. The system adds random variation on top, so no fixed pattern emerges.

The design protects the origin and destination of every trip against any third party or bad actor watching the data stream. The vehicle sends nothing for the first five hundred metres of any trip and transmits no signal to mark the end of a journey. If a batch is incomplete due to randomisation, it is not sent. Only complete, fully formed batches ever leave the vehicle.

Two terms appear throughout this use case and carry different meanings. A *batch* is the collection window, the stretch of road over which the vehicle gathers data before a gap begins. A *container* is the transmission package. Once a batch is complete, the system wraps the data into a container, attaches the hashed VIN, and sends it to the manufacturer's backend. The two terms describe the same unit of data at different stages.

4.3 Time Randomisation and Backend Deletion Break the Data Trail

The manufacturer implements two measures before forwarding any container to the third-party provider.

First, it applies a random time delay to each container independently. Containers can arrive at the third-party system in a different order than the vehicle that produced them. Reversing that reordering would demand substantial computational effort.

Second, once it forwards each container, it deletes that container from its own backend. The manufacturer retains no copy.

Before any container leaves its systems, it has already stripped the hashed vehicle identifier from every container. The manufacturer measures how quickly this stripping completes across the fleet. For example, half of all data loses its vehicle identifier within just over two

23. A hashed VIN resists casual inspection, but a VIN is a short, structured, and enumerable input, so an unsalted hash of it is weak against a brute-force search across the limited input space. This bears on the manufacturer, which holds the hashed identifier within its 30-minute window. It does not change the position of the third-party provider, which never receives the hashed VIN and so has nothing to brute-force.

hours, three-quarters within eight and a half hours, nine in ten within less than two days, and ninety-five percent within six days. The third-party provider receives only stripped data, with no reference to a VIN.

An adversary who defeated the reordering would still have no way to say which vehicle produced the sequence. Nothing ties it to any vehicle.

4.4 Converting GPS Points into Road Segments Removes Precise Location

The third-party provider unpacks each container, matches its GPS coordinates to the nearest road segment on a digital map, and then deletes the raw container. Transforming vehicle data into a road segment this way is a crucial step. The provider stores no raw data. Its output is a map layer or safety signal that reflects the combined observations of many vehicles onto a road segment, not a record of any individual vehicle's movement.

The third-party provider may also combine the road segment data it receives with FCD from other vehicle manufacturers to improve map coverage. This further dissolves any individual vehicle's contribution into a larger pool fed by multiple manufacturers.

4.5 Swarm Aggregation Dissolves Individual Vehicles into Traffic Patterns

Swarm aggregation,[24] relies on many vehicles simultaneously reporting observations for the same road segments. The fleet generates overlapping observations from different vehicles across every point of the road network. At this scale, no single vehicle produces a data point that stands apart from the rest.

Overlapping reports from many vehicles create a data cloud. An individual vehicle's contribution is one among many for any given location. Even a targeted attempt at attribution would face randomised batch boundaries, random gaps, time-delay reordering, and the sheer density of other vehicles' contributions.

One scenario is worth addressing directly. On sparse roads with light traffic, such as rural routes at off-peak hours, the swarm effect weakens. Few vehicles contribute data for the same segment, and a single vehicle's observation may not dissolve into a larger group.

AI may find patterns of individual vehicle behaviour that human analysts miss. The architecture addresses this directly. The third-party provider applies a minimum contributor threshold at the output level, retaining a road segment observation only when a set number of vehicles have contributed to it within a given time window. Below that threshold, the provider suppresses the data point rather than storing it. The edge case of a single vehicle on a quiet rural road at an off-peak hour leaves no trace in the dataset. Combined with immediate container deletion, no individual observation survives, even where the swarm is thin.

4.6 Road Conditions Survive the Pipeline, Individual Drivers Do Not

We run the SRB test first, then the 2014 anonymisation opinion precautionary bar. The assessment takes the third-party provider's position as a recipient, because the manufacturer holds the hashed identifier within its 30-minute window and the data stay personal in its hands.

Step 1. The SRB ruling sets two cumulative conditions for the recipient, here the third-party provider. The first asks whether the provider can lift the measures during its processing. We believe it cannot. The manufacturer strips the hashed vehicle identifier from every container before forwarding it, so the provider never receives a VIN and cannot restore one it never held. The provider also cannot undo the random batching, the suppressed trip boundaries, or the time-delay reordering, because the in-vehicle system and the manufacturer's backend apply those upstream. The second condition asks whether the measures in fact prevent the provider from attributing a container to a natural person, including by cross-

24. See Glossary.

checking against other factors. On busy roads they do. The containers carry GPS coordinates and timestamps but no identifier, the first five hundred metres of every trip never leave the vehicle, no signal marks a journey's end, and the provider map-matches each container and then deletes the raw data. There is no persistent trace to cross-check against an external location dataset. The deletion is the provider's own practice, so it carries no weight under Step 1. We count it as Step 2 evidence, and the upstream measures alone still satisfy both conditions on busy roads.

One scenario needs care. On a sparse rural road at an off-peak hour, a single vehicle's observation can stand apart, and a unique pattern can re-open a route to attribution. The minimum contributor threshold answers this directly. The provider keeps a road segment observation only where a set number of vehicles contributed within the time window, and suppresses the data point below that threshold. The threshold is the provider's own control, so we count it as Step 2 evidence, not as a Step 1 measure. Where it runs, the sparse-road risk closes and the verdict of most likely anonymous holds for this recipient. Where it does not, the sparse-road data may remain personal data, and we believe should be treated that way.

Step 2. We then test the data against the 2014 anonymisation opinion criteria.

The data clear both steps where the minimum contributor threshold runs. We believe the floating car data are most likely anonymous in the hands of the third-party provider as the recipient, within the meaning of Recital 26 GDPR, and they remain personal in the hands of the manufacturer that still holds the hashed identifier.

Two important qualifications frame that conclusion. First, it depends on the minimum contributor threshold. The report describes a generic use case, so it names no threshold value and no time window. Those parameters differ from one system to the next. The verdict therefore assumes that the provider sets both at a level that in fact suppresses single-vehicle observations, and can show this on request. On a sparse rural road at an off-peak hour, without that suppression, a single vehicle's pattern can survive, and the data may stay personal. Second, the verdict is recipient-specific. It holds for this provider in this pipeline, and it does not apply to a recipient that does hold means reasonably likely to identify the vehicle or its driver.

USE CASE 2 — SRB FRAMEWORK ASSESSMENT	Case C-413/23 P
CRITERION	ASSESSMENT
Separation	The provider never receives a vehicle identifier. The manufacturer strips the hashed VIN from every container before forwarding it, and the random time delay scrambles arrival order, so any attempt to reconstruct a vehicle's sequence would demand substantial computation and would also breach the contract. The separation is structural, not merely organisational.
Irreversibility	No VIN reaches the provider, and the provider deletes each raw container once it has matched the GPS points to a road segment. Nothing persists to reverse, and this holds for the provider rather than resting on a key kept elsewhere. The manufacturer keeps the means to link packets for up to 30 minutes, but that capacity sits with the controller, not the recipient.
Re-identification	The likelihood of re-identifying a natural person most likely stays insignificantly low. An attacker would need to defeat the random batching, the reordering, and the swarm density at once, and the deleted raw containers leave nothing to attack.
Singling out	The containers carry no quasi-identifiers. There is no name, no persistent device ID, and no journey start or end point to anchor a record to one vehicle. Random batch lengths and suppressed trip boundaries offer no surface for re-identification, and AI pattern recognition meets the same structural gap.
Linkability	No persistent pseudonym connects observations across containers, and the provider holds no raw GPS trace to match against external datasets. Each road segment output reflects the combined contribution of many vehicles, and the provider may pool it further with floating car data from other manufacturers, which dissolves any single vehicle's share even more.
Inference	The data describe road segments, not drivers. The aggregated output says nothing about where a particular vehicle went, how fast it travelled, or who sat behind the wheel.

USE CASE 3

5 · Use Case 3 — Development of Intelligent Anomaly Detection Features in Cloud Business Software

ASSESSMENT

This use case falls short of meeting the criteria of the Assessment Framework. It would however only require an adjustment to fully pass it, i.e. having the controller entrust the keys to a third party. We still chose to include it deliberately to show that the test threshold is very high, but not unreachable in practice.

Case Summary. *This case was provided by a European cloud business software provider in support of privacy-preserving AI feature development. Business software increasingly includes AI features that detect unusual patterns in customer data and flag them for action. To build these features, AI models must train on real business data. That data initially contain information that can identify the individual users. The use case shows how the provider scrambles customer identifiers before the data reaches the AI development team. It keeps the key that would unscramble them locked in a separate system, so the developers have no practical way to re-identify anyone. The development team learns from the patterns in the data without ever seeing who the data is about. This case shows even strong safeguards fall short once the data controller keeps the key, so the data stay personal in its hands.*

This use case examines how a cloud business software provider processes relational customer data to develop intelligent features, such as anomaly detection, while testing the boundaries of anonymisation.

Intelligent features that detect deviations from "normal" behaviour and propose mitigating activities are now a basic expectation in modern business software. To build such functionality, AI models must be trained on large sets of real relational (tabular) data to identify specific patterns. While the individual data subject is never of interest, the referential integrity of the data must be maintained. The AI must learn that the same entity involved in "Situation A" is also involved in "Activity B" across different database tables and events. Specific terms of service permit the use of customer data for this purpose, provided it is in an anonymised form.

5.1 The Limitations of Standard Privacy Enhancing Technologies (PETs)

The original cloud service data contains direct identifiers (names, email addresses) and internal database IDs. Extracting this data for AI training requires robust privacy enhancing techniques. However, standard PETs do not suffice for anomaly detection.

Applying differential privacy with a strict privacy budget (epsilon),[25] adds significant statistical noise and that noise destroys much of the data's utility. k-Anonymity,[26]

25. ibid
26. ibid

fails for the opposite reason. It hides each person in a crowd of lookalikes, so it removes the very outliers an anomaly detector needs. The provider therefore needs a different architecture.

5.2 Salted Hashing and Strict Environmental Separation

Since direct identification is irrelevant to the training objective, the organisation replaces internal identifiers with pseudonyms. To preserve the necessary linkability across tables (referential integrity) without exposing the data subjects, the provider applies salted hashing to the internal identifiers.

Crucially, the salt value is stored exclusively within the highly secure, live cloud service environment.[27] The hashed data are then transferred to a strictly segregated AI training environment. The internal data scientists and developers working in the AI training environment have absolutely no access to the live cloud environment, the original data, or the salt value.

5.3 The Patterns Reach Training, but the Controller Still Holds the Key

This use case presents a more complex fact pattern than the previous two use cases. The architecture intentionally preserves linkability within the dataset, and whoever holds the salt can technically reverse the hash. Both features are necessary for anomaly detection to function.

Step 1. The Court says that (we paraphrase) the SRB holds the additional information that lets it attribute the comments to a person, so the comments stay personal for the Board in spite of pseudonymisation.[28] Next, the Court turns to Deloitte, which is a separate company to which the SRB transmitted pseudonymised comments. For Deloitte the comments may be non-personal, but only if Deloitte cannot lift the measures during processing carried out under its own control, and only if the measures in fact stop Deloitte attributing the comments to a person.[29]

This use case has no “Deloitte”-entity. The data scientists work inside the same organisation, the provider keeps the salt in its production environment, and the data move between two entities of one company, not between two companies. So Paragraph 77 of the SRB ruling never engages. The provider controls the processing in the training environment, and the provider can lift the measures whenever it chooses, because it holds the salt. The exception needs a separate recipient processing under its own control, and there is none here. That leaves Paragraph 76 of the SRB ruling, the rule and not the exception. The provider is the pseudonymising controller that keeps the key, so the salted-hash data stay personal in its hands, training environment included.

Paragraph 78 of the SRB ruling closes the remaining gap when we read it together with the means test. Paragraph 16 of the SRB ruling counts pseudonymised data as information on an identifiable person where additional information can attribute them to that person. The Court links the recital to the means reasonably likely to be used, and treats the cost and time of identification as objective factors. For this provider the test needs no weighing at all. It keeps the salt live in its own production environment and can re-link hash to identifier on demand. Step 1 therefore, to the best of our understanding, points to personal data.

Step 2. The data fall short at Step 1, and the framework ends there. For completeness we still remark on two criteria, because they show what the architecture does well and where it stops.

The data fall short at Step 1, and the irreversibility remark points the same way. We believe the salted-hash

27. A single secret salt behaves like one master key. It protects the whole dataset, and it also gathers the whole risk in one place. One leak of that salt, by error, by an insider, or by a compromise of the production system, reverses every hashed identifier at once.

28. Case C-413/23 P, para 76.

29. Case C-413/23 P, para 77.

training data are personal data in the hands of the provider, and therefore most likely not anonymous, within the meaning of Recital 26 GDPR. This is not a close case decided by precaution. It is a direct reading of Paragraph 76 and Paragraph 16, the provider keeps the key, so the provider holds personal data, and the segregated training environment is one part of that same provider. Had the provider shared the data with an independent recipient that cannot obtain the salt, Paragraph 77 of the SRB ruling would engage, and the data could most likely qualify as anonymous in hands of the recipient.

This use case earns its place in the wider privacy discussion. It brings together the strongest mix on offer, (1) cryptographic separation, (2) an organisational split, and (3) a binding contract for the data scientists, and it still lands on the personal side, because the provider keeps the salt. Here we meet the limit the SRB ruling draws around a controller that holds the key. Beyond that ruling there is little case law to guide such a design, so it sits in untested territory, and only firmer guidance or an agreed certification standard will show developers how much closer to anonymity an architecture like this can move.

USE CASE 3 — SRB FRAMEWORK ASSESSMENT Case C-413/23 P

CRITERION	ASSESSMENT
Separation	The separation between the training environment and production is structural and cryptographic, and it is genuine. Three layers hold the salt away from the data scientists, technical controls, an organisational boundary, and a contract. This is the strongest part of the architecture, and it does real security work. It does not make the salt disappear from the controller that owns both environments.
Irreversibility	This criterion decides the 2014 anonymisation opinion cross-check, and it falls short. The report argues that without the salt, reversing the hash by brute force is computationally infeasible, and we believe that for the data scientists in isolation that is broadly fair. The 2014 anonymisation opinion do not read irreversibility relative to one team within the controller, though, and the provider keeps the salt live in production for its own purposes, so it can re-link hash to identifier on demand. The reversal key is a permanent feature of the system, not a residual risk, and data the controller can re-link do not count as irreversibly anonymised.

USE CASE 4

6 · Use Case 4 — Delivering Aggregated Audience Statistics to a Publishing Platform

ASSESSMENT

Under the Section 2 framework, the aggregated audience reports clear both steps. We believe they are most likely anonymous in the hands of the publishing platform as the recipient, while the measurement provider still holds personal data.

Case Summary. *This case was provided by a platform displaying video and news content in support of privacy-preserving audience measurement. Online publishers need reliable statistics on their audience, but those reports are produced from underlying data that may initially relate to individual users. The use case shows how aggregation, one-way data flows, contractual restrictions, and commercial separation ensure that the publisher receives only population-level statistics, with no access to raw data, panel lists, or identification keys. As a result, the publisher has no realistic way to single out, link, or infer information about a specific person from the reports. One limit remains. A thin demographic cell without a minimum cell size leaves membership inference possible for an actor that holds data about a specific person. Treating such reports as non-personal, where these safeguards are effective, can support trusted audience measurement in Europe, which underpins the monetisation of ad inventory and service improvement.*

This use case examines whether aggregated audience measurement reports delivered by a specialised measurement provider to an online publisher constitute personal data in the publisher's hands, even though the measurement provider has processed personal data of individual website visitors and app users to produce those reports.

6.1 How the Measurement Provider Collects and Processes Data

A specialised measurement provider (the Provider) offers a monthly audience measurement service to an online media publisher (the Platform). The Provider collects data, processes it, and delivers statistical reports to the Platform. The reports tell the Platform how many people visited its properties, who those people are in broad demographic terms, and how the Platform's reach compares to the overall internet population of an EU Member State.

The data flow consists of two stages and two data flows. In Stage 1, the Provider collects individual-level data through two separate data flows. People who volunteer to be panellists install metering software on their devices, and consent to the monitoring of their browsing behaviour across computers, mobile phones and tablets (Data Flow 1). In parallel, and only *after* the user, i.e., the website visitor or app user, has given consent, the Provider deploys specialised software across the Platform's websites and mobile applications (the Audience Measurement SDK or Software Development

Kit, a library of code that handles specific functions). The software tracks specific user actions, such as playing, pausing, or stopping a video, and sends this data directly from the user's browser or app to the Provider's servers. Each transmission also includes details about the content being watched, such as the video title and duration, basic information about the user's device, and a record of what they have consented to (Data Flow 2).[30] The data never passes through the Platform's own infrastructure. It goes directly to the Provider.

In Stage 2, the Provider (a) combines both data streams (data flows 1 and 2), (b) corrects for panel biases,[31] and (c) extrapolates the results to represent the full internet population in a European Member State. The output is a set of monthly statistical reports. These aggregated reports are the only thing the Provider delivers to the Platform. The Platform has no access to the underlying data, the panel roster, processing algorithms, or any intermediate datasets.

6.2 Aggregation as the Main Privacy Safeguard

The primary privacy enhancing technique in this use case is aggregation. Aggregation pools individual-level records into group-level counts and statistical summaries. Once expressed at the population level, the individual contribution disappears into the group. There is no record left to single out, link, or reverse.

In practice, the Provider combines signals from large numbers of users into figures such as a coverage rate or a unique visitor count for a given month. No individual behaviour produces any single figure. Each figure reflects the activity of a substantial population. The reports contain no individual-level records, no quasi-identifiers, and no data point from which a natural person can be identified. The reports alone do not let anyone reconstruct any individual from those figures.

Aggregation, however, does not eliminate all the likelihood of re-identification on its own. A membership inference attack,[32] asks a more limited question. Rather than identifying who someone is, it asks whether a specific individual's data contributed to a given aggregate. This remains a realistic threat even where full re-identification is not. In theory, an adversary with additional information about a specific person could attempt to detect their presence in a monthly report by comparing outputs across time or across demographic cells.

In this use case, that risk stays low because of the specific architecture.[33] Whatever first-party data the Platform holds from running its own websites and apps, the monthly reports give it nothing to match those records against. Each report carries population-level counts and demographic splits, with no individual record and no identifier. The SDK sends the measurement signals straight to the Provider, so the Platform never learns which of its visitors the Provider counted. The

30. The data the Provider collects depends on the product the Platform selects. The basic, site-centric product uses both data flows. Data Flow 1 (panellists) and data flow 2 (the SDK) together supply IP address, visited page, visit volume, and device type, enabling the Platform to receive monthly and daily average visit counts per channel, socio-demographic breakdowns, and audience rankings. The video-centric option goes further. A separate tracker alongside Data Flow 2 collects video title, viewing duration, viewing behaviour, IP address, and volume of videos watched. The Platform then receives unique viewer counts, time spent per video, audience split by device type, demographic breakdowns, and rankings.

31. See Glossary.

32. ibid

33. The low reading holds for the Platform, which cannot query the reports. It does not hold in general. Without differential privacy or a minimum cell size, a thin demographic cell, such as a narrow age band in a small region, leaves membership inference possible for an actor that holds additional data about a specific person. In our view, a minimum cell size would remove that risk.

reports are also static monthly deliveries with no querying mechanism. The standard membership inference technique queries a system repeatedly, with and without a specific person, to detect a difference. The Platform has no way to do this. Across millions of internet users, any single person's contribution to a monthly figure is, we believe, practically undetectable. The one exception is a thin demographic cell. The next paragraph deals with that risk.

A stronger mathematical protection would involve applying differential privacy noise to the outputs or enforcing minimum cell size thresholds,[34] for the demographic breakdowns. Differential privacy,[35] adds calibrated noise so that no individual's presence or absence can be detected from the output. This goes further than aggregation because it provides a mathematical guarantee rather than a practical barrier. Minimum cell size thresholds suppress any demographic cell with fewer than a defined number of individuals, preventing membership inference through small-group combinations. Where the Provider applies either measure, the protection becomes mathematical rather than architectural. The use case does not depend on these measures to reach its conclusion, but their application would further reduce the residual risk and strengthen the overall anonymisation assessment.

Pseudonymisation is also present, but at an earlier stage. The site-centric collection layer links signals to device identifiers and cookie-based identifiers rather than to names or directly identifying data. The Platform never receives even this pseudonymised layer. Aggregation absorbs it entirely before the reports are produced.

Four convergent barriers reinforce the aggregation. The first is *structural*. The Platform receives only the aggregated reports. The individual-level data from which they were derived stays with the Provider. The SDK fires HTTP requests directly from the user's browser to the Provider's servers, so the raw signals never pass through the Platform's infrastructure. The separation is built into the data architecture, not merely promised on paper.

The second barrier is *technical*. Re-identifying any individual from the reports would require the Platform to obtain four things it does not have. It would need (1) the raw device and behavioural signals transmitted by the SDK, (2) the panel roster, (3) the hybridisation and extrapolation algorithms, and (4) the mapping between device signals and identified panellists. None of these reaches the Platform. The SDK fires HTTP requests directly from the user's browser to the Provider's servers. The data never passes through the Platform's own infrastructure.

The third barrier is *contractual*. The Platform is prohibited from attempting to reverse-engineer the reports or re-identify individuals from them. The contractual agreement prevents it from modifying, merging, transforming or combining the reports with any other data without the Provider's prior written consent. These prohibitions carry significant legal risk and work against any such attempt.

The fourth barrier is *commercial*. The Provider's methodology, panel composition and extrapolation algorithms are its core trade secrets. Disclosing them to a publisher-client would undermine the publisher-neutral integrity of a certified measurement that the entire market depends on, expose panellist data, and create serious competitive risk. Even if a Platform sought to obtain this information, the Provider would have every structural incentive to refuse. The "means reasonably likely to be used" under Recital 26 GDPR must account not only for what a recipient could technically attempt, but for what a third party holding the necessary information would realistically agree to provide. In this case, cooperation is commercially inconceivable. Rankings add a further layer. Many audience rankings circulate publicly across the market, and a Platform ranked fifth already knows which other Platform sits

34. See Glossary.

35. ibid

above and below it. The marginal value of attempting to re-identify individual persons from reports that are partly public is in our view negligible. There is also a self-defeating quality to any such attempt. The Platform commissioned these reports precisely because the market trusts them. Any suspicion of tampering would destroy that trust and waste the investment entirely. The reports also cover a short, defined time period, typically a single month. By the time any re-identification attempt could be conceived and executed, the data would have lost all commercial relevance.

One further point deserves attention. The Provider deletes or anonymises its raw data in accordance with its own retention schedule, but the Platform has no visibility into that schedule. Under an extensive reading of personal data, the reports might qualify as personal data on the day of delivery and as non-personal data months later, once the Provider deletes the underlying records, with the Platform never informed of the change either way. That is an impossible compliance obligation. The Platform would be required to apply GDPR rules to data whose legal status it cannot determine and cannot influence. The SRB ruling resolves the absurdity by anchoring identifiability to the means available to the recipient, rather than to the theoretical capabilities of a third party holding separate, inaccessible data.

6.3 Four Convergent Barriers Place the Reports Outside GDPR Scope for the Platform

The assessment takes the Platform's position, because the Provider holds the panel roster, the raw signals, and the identifier mappings, so the data stay personal in its hands.

Step 1. The SRB ruling sets two cumulative conditions for the recipient, here the Platform. The first asks whether the Platform can lift the measures during its processing. We believe it cannot. Aggregation, panel hybridisation, and extrapolation all run at the Provider, and the Platform receives only the finished monthly reports. It holds no raw signals, no panel roster, no algorithms, and no way to query the underlying data, so it cannot undo an aggregation it never performed.

The second condition asks whether the measures in fact prevent the Platform from attributing a figure to a natural person, including by cross-checking against other factors. We believe they do. Each figure pools the activity of a substantial population, the SDK sends signals straight from the user's browser to the Provider, and the Platform never learns which of its visitors the Provider counted. A monthly count contains no individual record to cross-check, whatever data the Platform holds of its own.

One scenario needs special care. A thin demographic cell, such as a narrow age band in a small region, can stand apart, and a unique combination can make membership inference possible again for anyone who holds data about a specific person. We do not rest the verdict on the absence of such data at the Platform. We rest it on the minimum cell size condition in Step 2. Where the Provider suppresses thin cells, the second condition holds, and in our view the SRB test points to anonymous for this recipient.

Step 2, the 2014 anonymisation opinion precautionary bar. We then test the reports against the stricter criteria of the 2014 anonymisation opinion.

The reports clear both steps for the recipient. We believe the aggregated audience measurement reports are most likely anonymous in the hands of the Platform, within the meaning of Recital 26 GDPR and the SRB ruling, and they remain personal in the hands of the Provider.

Two limits frame that verdict. First, it depends on real aggregation across large groups of people. When a demographic cell is thin and the Provider sets no minimum cell size, and when it also adds no differential privacy noise, membership inference remains possible. The verdict for the demographic breakdowns therefore holds only where the Provider suppresses cells below a set minimum, and can show this on request. Second, the verdict is recipient-specific. It holds for the Platform, whose own data cannot connect to the population-level figures. It does not apply to a recipient that holds means

reasonably likely to identify a contributor. Each new recipient needs its own assessment.

USE CASE 4 — SRB FRAMEWORK ASSESSMENT	Case C-413/23 P
CRITERION	ASSESSMENT
Separation	The Platform receives only the aggregated reports. The panel records, the raw device signals, and the identifier mappings stay with the Provider, and the SDK fires its requests straight from the user's browser to the Provider's servers, so the raw signals never touch the Platform's infrastructure. The data flow creates the separation, not a promise on paper.
Irreversibility	The Platform cannot reverse the aggregation. It lacks the raw inputs, the panel roster, and the hybridisation algorithms that turn panel data and site measurements into population estimates. Once a figure pools millions of contributions, no record survives for the Platform to unpick.
Re-identification	The likelihood of re-identifying a natural person from the reports most likely stays insignificantly low. Each figure pools the activity of a substantial population, so no individual record remains in the output. Re-identifying one person would require the raw signals, the panel roster, the extrapolation algorithms, and the mapping between device signals and panellists. The Platform receives none of them.
Singling out	The reports carry no individual-level records and no quasi-identifiers. It is not plausible to single out a natural person from a population-level count.
Linkability	The Platform holds no panel roster, no raw tag data, and no identifier mapping. A population-level figure contains no record to link to any other dataset.
Inference	The reports describe distributions across large segments, and they say nothing about one person's behaviour. The residual risk sits here rather than anywhere else. Where a demographic breakdown produces a thin cell, and the Provider applies neither differential privacy nor a minimum cell size threshold, membership inference remains possible for an actor that holds additional data about a specific person. A minimum cell size answers this directly. Where the Provider suppresses cells below a set minimum, that risk disappears and the verdict of most likely anonymous holds for the breakdowns. Without that suppression, the thin demographic breakdowns may remain personal data.

USE CASE 5

7 · Use Case 5 — Training Ad Prediction Models Without Individual User Records

ASSESSMENT

Assessed under the framework in Section 2, the synthetic training dataset clears both steps. We believe it is most likely anonymous in the hands of the advertising provider, and the source records stay personal data throughout.

Case Summary. *This case was provided by a European online advertising provider in support of privacy-preserving advertising. To train AI models that predict whether users may click on ads, the organisation does not rely on individual click records. The use case shows how pseudonymisation, aggregation, statistical noise and synthetic data, can be combined so that model training uses group-level patterns rather than identifiable user-level data. As a result, the model can learn useful advertising signals without exposing who clicked, linking records back to individuals, or inferring a specific person's behaviour. One limit remains. The verdict depends on a small, tracked privacy budget, and on group counts large enough to hide any one person. Treating the synthetic training data as non-personal, where these safeguards are effective, can support privacy-preserving AI development in Europe while enabling innovation in advertising and other data-driven sectors.*

This use case walks through how an online advertising provider combined four privacy enhancing technologies to reduce the likelihood of re-identification at each stage of AI model training. Pseudonymisation, aggregation, differential privacy, and synthetic data each target a distinct vulnerability, and together they bring the residual risk profile to a level the Assessment Framework can accommodate.

An online advertiser provider (the Organisation) trains AI models to make predictions about individual users with privacy-by-design in mind. This case focuses on predicting the probability that a user clicks on an online advertisement. To train such a model, the Organisation could use data that links individual characteristics to individual outcomes. For example, someone clicking on sports shoes on a retail website might like sportswear. However, the Organisation does not have access to individual outcome data. It cannot know whether a specific user clicked. It can only access aggregated totals, such as the number of clicks across a (large) group of users.

The use case is relevant not only for online ads. The same challenge applies in other sectors. For example, a medical researcher may want to predict whether a patient has a specific disease but can only base this on the total count of cases in a population (i.e., the entire French population). In both cases, individual-level data is not accessible for privacy reasons.

7.1 Replacing Direct Identifiers with Hashed or Random IDs

The Organisation starts with a dataset of pseudonymised records. The records contain non-personal data such as ad campaign type or ad slot. These also include hashed records. Hashing is a mathematical way of scrambling a piece of information, such as an email address or a telephone number, into a fixed string of letters and numbers that cannot easily be reversed. An example of the hash equivalent of an identifier '13278a5c-3997-4b97-826d-19609eecb975' in a cookie is 'c68e128c46549947416241878e16e9db811da0c99e7431d6982c56481008dc6c'. This type of pseudonymised data still carries three distinct paths of re-identification.

First, an adversary can single out an individual by combining quasi-identifiers, even when they are hashed. A quasi-identifier is a piece of information that does not identify someone on its own but can do so when combined with other pieces of information. A postal code alone does not identify you. Your date of birth alone does not identify you. But the combination of a postal code and a partial date of birth might narrow a dataset to a single person. Second, an adversary can link pseudonymised records to external datasets that contain directly identifiable information and recover a person's identity. Third, an adversary can infer the characteristics or behaviours of a specific individual from patterns in the data, even without direct identification.

Pseudonymised data alone does not eliminate these risks. The Organisation therefore applies additional PETs to the training data and uses a third during model training.

7.2 Pooling Records into Group Counts Removes Individual Data

The Organisation pools single pseudonymised records across many users into group-level counts. These counts record how many users share a given characteristic and how many had the outcome of interest, such as a click on an advertisement. The pseudonymised record disappears. An adversary who wants to single out or link a specific person must reverse the aggregation. This requires a brute-force attack across all possible combinations. At scale, this is computationally unreasonable. The attacker also needs to know the aggregation logic, which the Organisation does not disclose.

7.3 Adding Statistical Noise Protects Each Person's Contribution

Think of statistical noise as deliberate imprecision. Instead of reporting that exactly 47 people in a postal code are between 30 and 40 years old, the system might report 43 or 51. The true count shifts by a small random amount every time. That randomness changes the calculus for an attacker entirely. The Organisation adds statistical noise above a defined threshold to the aggregated counts. The noise level is governed by a configurable privacy parameter (epsilon),[36] that determines how much any individual's data can influence a given output. The Organisation sets this parameter at a level that provides meaningful plausible deniability for each individual contribution. In the demographic age-bracket example, the reported count for any group may shift by more than two years of age range. Any individual's contribution to a given aggregate becomes undetectable. No one can reasonably confirm whether a specific person's data contributed to a given aggregate. This blocks inference attacks that target an individual's presence or characteristics in the dataset.

7.4 Reconstructing Training Signals from Group Counts Avoids User-Level Data

The Organisation generates synthetic data by reconstructing user-level training signals from group counts, without using pseudonymised records. The idea is to generate fake-but-plausible data that behaves like

36. ibid

real data, without exposing anyone in it. When the outcome of interest is not available at the level of a single user, the Organisation cannot train the model directly. For example, it knows that 1 in 1,000 users clicked on an advertisement. It does not know which user clicked. The Organisation uses a probabilistic model to generate synthetic individual records that are consistent with the observed group counts. This mimics how individual user data could have been generated. The Organisation then trains the AI model on this synthetic dataset. It never uses individual records.

7.5 The Synthetic Dataset Under the Two-Step Assessment

The four-layer architecture resists classical re-identification techniques, including quasi-identifier combination and external dataset linkage. Synthetic data ensures that model training never requires access to individual records. After applying all four PETs, the residual risk profile differs substantially from that of the original pseudonymised dataset.

This use case illustrates a layered PET architecture. Each PET targets a specific re-identification threat. Assessed against the Assessment Framework, the layered architecture directly addresses the three core anonymisation criteria. It also accounts for means reasonably likely to be used.

Step 1. The synthetic dataset is a different object. Use Case 3 kept a secret salt that reverses every hash on demand. This pipeline runs one way, and it keeps no such key, so the Organisation cannot link a synthetic record back to a person. Recital 26 GDPR treats data that no one can attribute to a person, the controller included, as anonymous. For the synthetic dataset Step 1 points to anonymous, while the source records stay personal.

Step 2. A single controller that anonymises its own output is exactly the case where precaution earns its keep, so we test the synthetic dataset against the stricter criteria before we accept the result.

The four PETs work in sequence, and each closes a gap left open by the previous layer. Pseudonymisation removes direct identifiers, aggregation removes the individual record, noise protects each person's contribution, and synthetic data severs the link between training inputs and real users. The combination of four stacked PETs leaves no individual record at any stage of model training. The residual risk falls within the threshold the Assessment Framework treats as anonymous, accounting in our view for means reasonably likely to be used.

One objection needs a direct answer. The 2014 anonymisation opinion states that differential privacy does not change the original data. As long as the controller keeps those data, it can identify people in the noisy results, and the results stay personal data.[37] The Organisation keeps its source records, so at first sight that position points the other way. The warning, however, covers noisy views of real records, where each output still describes real people. The synthetic records describe no one. The generator builds them from group counts alone, never from a pseudonymised record, so

37. 2014 anonymisation opinion, Section 3.1.3. "It has however to be clarified that differential privacy techniques will not change the original data and thus, as long as the original data remains, the data controller is able to identify individuals in results of differential privacy queries taking into account all the means likely reasonably to be used. Such results have also to be considered as personal data". See also section 3.1.3.3 on failures of differential privacy.

the source records open no route back. The tracked privacy budget then caps what the counts can leak about any one person. We read the opinion as a warning against shortcuts, not as a block on a pipeline that removes the record before training begins.

We believe the synthetic training dataset is most likely anonymous in the hands of the Organisation, within the meaning of Recital 26 GDPR, while the source records remain personal in those same hands. The conclusion rests on the transformation being genuinely one way, not on a promise to keep a key locked away.

The verdict holds for the synthetic dataset under a tight privacy budget and minimum cell sizes,[38] and it does not extend to the source records, which stay personal throughout.

USE CASE 5 — SRB FRAMEWORK ASSESSMENT	Case C-413/23 P
CRITERION	ASSESSMENT
Separation	The synthetic dataset holds no pseudonymised record. Aggregation strips the individual record before generation begins, and the synthetic records carry no identifier from the source. The separation is built into the pipeline, not bolted on afterwards.
Irreversibility	This criterion decides the cross-check, and here it passes. The transformation runs forward only. Aggregation destroys the record, differential privacy adds noise the Organisation cannot subtract, and synthetic data severs the tie to real users, so the Organisation cannot re-link a synthetic record to a person. The differential privacy step carries this finding, and it carries it only while the privacy budget (epsilon) stays small and tracked across every release.
Re-identification	The likelihood of re-identifying a natural person most likely stays insignificantly low. An attacker would need to reverse the aggregation, subtract the noise, and defeat the synthetic generation all at once. The records themselves describe no real person to start from.
Singling out	The synthetic records describe no real person. An attacker who isolates one synthetic record isolates an invented one, so the operation reaches no natural person.
Linkability	The dataset holds no real record to match against an external source. A synthetic record links to nothing outside the model.
Inference	The residual risk sits here rather than anywhere else. Statistical noise gives every individual plausible deniability about a contribution to any count, and membership inference against a real person stays blocked while the budget holds. A thin demographic cell still deserves watching, because the protection is statistical, not absolute.

38. Synthetic data inherits the statistical properties of the counts it learns from. A cell that covers one person or only a few does not describe a group. It describes that person. The generator can then reproduce that rare combination of attributes in a synthetic record. This residual risk is attribute disclosure rather than membership inference, and the privacy budget alone does not remove it. Minimum cell sizes and the suppression of small counts must close that gap before generation begins.

SECTION 08

8 · Conclusions — What the Use Cases Show About Pseudonymous Data in Practice

The five use cases sit in very different industries. One sends prompts to a generative AI model. Another turns road sensor data into a live traffic map. A third trains anomaly detection on relational business data, but the provider keeps the key. The last two follow the same logic in their own settings, a publisher that receives only monthly audience reports, and an advertiser that trains on noisy synthetic records.

8.1 What the Five Use Cases Share

In every case, anonymisation rests on more than a single technique. The legal intelligence provider does not just scrub prompts. It also signs a zero-retention agreement and relies on the AI provider's inability to retain anything beyond the moment of inference. The vehicle manufacturer does not just strip the hashed VIN. It also delays arrival times, deletes its own copy, and forwards only road segment observations. The cloud business software provider does not just hash the customer identifiers. It also locks the salt in a separate environment, beyond the reach of the development team. The audience measurement provider does not just aggregate. It also withholds the panel roster, the hybridisation algorithms, and the mapping between devices and panellists, and it backs all of that with a contract. The ad prediction stack layers pseudonymisation, aggregation, noise, and synthetic data, each plugging a gap left by the previous step.

The conclusions for each use case do not all run one way. We believe that four of the five reach anonymity, but only for the recipient and only on the facts. The scrubbed prompts are most likely anonymous for the AI provider, the traffic maps for the third-party provider once a minimum contributor threshold holds, the audience reports for the publishing platform, and the synthetic training data for the advertiser that builds them. In each of those four, the data stay personal for the party that holds the source records. The anomaly detection case is the exception. That single counter-example carries as much weight as the four that pass since it helps us sharpen the discussion on clear criteria for determining when pseudonymised data become anonymous data.

The CJEU's relative approach comes down to a simple question. Does the recipient have reasonable means to identify individuals? Not whether someone, somewhere, with unlimited resources and bad intent, could in theory rebuild an identity. The use cases show why that distinction matters. From Deloitte's perspective as the recipient, the data qualifies as anonymous. From the SRB's perspective as data controller, the data were personal because the SRB held all information needed to identify the authors. At the same time, the use cases show that the likelihood of re-identification should not be dismissed entirely.

8.2 Looking Ahead

The use cases offer a concrete starting point for the consistent application of the legal requirements, and for greater clarity on whether data in the hands of an organisation qualify as personal or anonymous data within the meaning of Article 4 GDPR. Criteria to standardise the assessment should, in our view,

1) reward architectural separation;

2) recognise the convergence of multiple barriers;

3) take the organisation's actual capabilities as the reference point, and

4) treat contractual and commercial constraints as reinforcing factors rather than standalone safeguards;

5) leave room for mitigating risks arising from the facts and circumstances of a use case. For example, attachments containing personal data still require care, even when the prompts around them do not. And sparse traffic on rural roads still warrants a minimum contributor threshold to prevent singling out individual contributors;

6) ensure the assessment is contextual and recipient-specific, because the same dataset can be personal data in one set of hands and anonymous in another;

7) layered protection deserves explicit recognition, because the strength in these cases comes from how the layers interact, not from any single method.

We state the limits of those seven criteria plainly. They rest on five use cases. Each came from a single European company with an interest in the outcome. Five is a small sample. It also favours success, since four of the five reach anonymity under our framework, and no company volunteers an architecture it expects to fail. We therefore offer the criteria as working hypotheses rather than settled conclusions. They are precise enough as a starting point. Independent attack testing, supervisory scrutiny, or a certification pilot under Article 42 GDPR could each confirm or correct them. We would welcome all three.

In closing, this report does not constitute legal advice and we are not a supervisory authority. We are privacy practitioners applying a framework with precautionary intent. Where risk arises or increases under specific circumstances, the right response is clear mitigation or documented residual risk, not a blanket assumption of compliance.

We encourage the Commission, regulators, and other relevant stakeholders to use these use cases as a foundation for defining clear, harmonised criteria enabling organisations to determine, with greater legal certainty, when data falls within or outside the scope of the GDPR. In parallel, this work should be complemented by the development, in close cooperation with industry, of practical instruments such as codes of conduct and certification frameworks, which can operationalise these criteria, promote consistent application across sectors, and reward robust anonymisation practices. Such an approach would preserve a high level of protection for individuals while fostering innovation and enabling responsible data use for innovation at scale.

We thank every reader who took the time to get this far into the report, and we hope the examples prove useful to those working to shape harmonised criteria and clearer guidance.

REFERENCE

Glossary of Technical Terms

Agentic AI

Agentic AI refers to systems that autonomously plan and execute multi-step tasks without requiring human input at each stage. Unlike a chatbot that answers one question at a time, an agentic system chains operations together, accesses multiple data sources, uses specialised tools if needed, and acts on the results. This matters for re-identification risk because scale and automation change what qualifies as a means reasonably likely to be used under Recital 26 GDPR.

Differential Privacy

Differential privacy adds calibrated random noise to statistical outputs. The result looks accurate at the population level but makes it impossible to tell whether any specific person's data contributed to a given figure.

Epsilon

Epsilon is the parameter that controls how much noise differential privacy injects into an output. A lower value means stronger protection and more distortion of the underlying figures.

k-Anonymity

k-Anonymity hides each person in a crowd of lookalikes. It groups records by their *quasi-identifiers*, the details that single someone out when combined, such as postcode, age, and job title, so that every combination appears at least k times. To get there it coarsens the data, turning an exact age into a age group.

Membership Inference Attack

A membership inference attack asks a narrower question than re-identification. Rather than asking who someone is, it asks whether a specific person's data contributed to a given aggregate. The distinction matters because this attack works even against well-anonymised outputs.

Minimum Cell Size Threshold

A minimum cell size threshold suppresses any statistical output that draws on fewer than a defined number of people. Without it, a report about a very small demographic group can effectively point to specific individuals. For example, a small group wearing red shirts in a crowd wearing blue shirts. Regulators have begun treating the likelihood of singling out as a baseline expectation.

Panel Bias and Extrapolation

A measurement panel is a recruited group of people who agree to have their behaviour tracked. Because such panels never perfectly mirror the broader population, a systematic gap emerges between the sample and the population it represents. Extrapolation corrects for that gap through statistical modelling, allowing a provider to produce population-level estimates from a limited sample.

Quasi-Identifier

A quasi-identifier is a piece of data that does not identify anyone on its own but can do so when combined with other pieces of information. A postal code alone tells you nothing. Combine it with a date of birth, and the combination may narrow a dataset to one person. The likelihood of identification is not in any single field but in what an adversary can assemble from several.

Swarm Aggregation

Swarm aggregation occurs when so many devices simultaneously report observations for the same location that no single device's contribution stands out. The volume of overlapping data makes attribution practically impossible. Unique patterns, such as a quiet rural road at night, weaken the effect and warrants additional safeguards.

Synthetic Data

Synthetic data consists of artificially generated records that mimic the statistical properties of real data. No record in a synthetic dataset corresponds to an actual person.

REFERENCE

Appendices

The cases below are the EU court judgments behind the Assessment Framework. We look at the General Court and the CJEU, including the appeal cases. National court rulings and DPA decisions are out of scope. Part A is the CJEU SRB litigation. Part B goes through the CJEU case law on identifiability that came before it, oldest first.

A: CJEU SRB LITIGATION

A1: SINGLE RESOLUTION BOARD V EUROPEAN DATA PROTECTION SUPERVISOR

Case T-557/20 Single Resolution Board v European Data Protection Supervisor EU:T:2023:219.

The General Court moved toward a relative approach to identifiability. Sharing pseudonymous data with a third party, such as Deloitte, without also transferring the re-identification information, does not automatically render the data anonymous from that third party's perspective. Whether the data remain personal depends on whether the recipient has legal means which could in practice enable it to access the additional information needed for re-identification. The Court also held that pseudonymised data do not automatically constitute personal data in every case and for every person.

The CJEU set this judgment aside on appeal in C-413/23 P, sent the case back to the General Court.

A2: EUROPEAN DATA PROTECTION SUPERVISOR V SINGLE RESOLUTION BOARD

Case C-413/23 P European Data Protection Supervisor v Single Resolution Board EU:C:2025:645.

Where information consists of personal opinions or views, their nature as expressions of a person's thinking makes them necessarily closely linked to that person. No separate analysis of content, purpose or effect is required to establish that such information relates to an identifiable person.

Pseudonymised data are not automatically personal data for every person in every circumstance. The same data can be personal for the controller, who holds the re-identification key, and anonymous for a third party who has no key and no reasonable means to obtain one. The relevant question is whether the specific recipient has means reasonably likely to be used to identify the data subject. The obligation to inform data subjects under Article 15(1)(d) of Regulation 2018/1725 attaches at the time of data collection and runs from the controller's perspective. The SRB therefore had to inform data subjects before transferring pseudonymised comments to Deloitte, regardless of whether those data were personal from Deloitte's perspective.

A3: AG OPINION

Case C-413/23 P, Opinion of AG Spielmann, European Data Protection Supervisor v Single Resolution Board EU:C:2025:59.

The Advocate General advised the Court to set aside the General Court judgment. He confirmed that the relative approach to identifiability applies and that identifiability turns on the means reasonably available to the specific recipient, not on the theoretical capabilities of any actor with unlimited resources.

B: CJEU CASE LAW

B1: SCARLET EXTENDED SA V SOCIÉTÉ BELGE DES AUTEURS, COMPOSITEURS ET ÉDITEURS SCRL (SABAM)

Case C-70/10 Scarlet Extended SA v Société Belge des Auteurs, Compositeurs et Éditeurs SCRL (SABAM) EU:C:2011:771.

IP addresses collected by internet service providers are personal data because they enable the identification of users. The ruling was an early point in EU case law establishing that network-level technical identifiers fall within the Article 4 definition.

B2: PATRICK BREYER V BUNDESREPUBLIK DEUTSCHLAND

Case C-582/14 Patrick Breyer v Bundesrepublik Deutschland EU:C:2016:779.

An online media services provider holds a dynamic IP address as personal data where it has legal means to obtain additional identifying information from the internet service provider. The same reasoning extends further. Data that are otherwise impersonal can still be personal for a controller who has legal means to access identifying information held by a third party. Where that information sits with another party, that alone does not put the data subject beyond the reach of the GDPR.

B3: PETER NOWAK V DATA PROTECTION COMMISSIONER

Case C-434/16 Peter Nowak v Data Protection Commissioner EU:C:2017:994.

The phrase "any information" in Article 4 GDPR reflects a deliberate legislative choice to assign the concept of personal data a broad scope. Information relates to an identified or identifiable person where its content, purpose or effect links it to that person. Personal opinions, as expressions of a person's thinking, necessarily point back to their author. An analysis of content alone can therefore suffice. The three criteria are alternatives connected by "or," not a cumulative test.

B4: GESAMTVERBAND AUTOTEILE-HANDEL E.V. V SCANIA CV AB

Case C-319/22 Gesamtverband Autoteile-Handel eV v Scania CV AB EU:C:2023:837.

A Vehicle Identification Number alone is not personal data. It becomes personal data when the receiving party has the means and information to identify a vehicle owner who is a natural person. Data that are otherwise impersonal can become personal. The trigger is whether the controller passes them to persons who have the means reasonably likely to enable identification. Where that happens, the data are personal both for those persons and, indirectly, for the original controller.

B5: OC V EUROPEAN COMMISSION

Case C-479/22 P OC v European Commission EU:C:2024:215.

Identification is not reasonably likely where law prohibits it or where the effort in time, cost and labour would be disproportionate. Which perspective governs the assessment of identifiability depends on the circumstances of the specific processing. The court does not stop at asking whether the controller holds identifying information. It must examine whether the relevant audience could reasonably identify someone from the statements made, including by combining them with information from other sources.

B6: IAB EUROPE V GEGEVENSBESCHERMINGSAUTORITEIT

Case C-604/22 IAB Europe v Gegevensbeschermingsautoriteit EU:C:2024:214.

A Transparency and Consent String constitutes personal data within the meaning of the GDPR, particularly when it accompanies an IP address. The ruling confirms the "content, purpose or effect" criterion from Nowak and extends the Breyer reasoning to this context. The core point remains the same. Where the controller has legal means to obtain identifying information from another party, data that are otherwise impersonal still connect to an identifiable person. It makes no difference whether the controller holds that information directly or obtains it through a third party.

C: RELEVANT GDPR DEFINITIONS AND RECITALS

C1: RECITAL 26 GDPR

(26) The principles of data protection should apply to any information concerning an identified or identifiable natural person. Personal data which have undergone pseudonymisation, which could be attributed to a natural person by the use of additional information should be considered to be information on an identifiable natural person. To determine whether a natural person is identifiable, account should be taken of 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. To ascertain whether means are reasonably likely to be used to identify the natural person, account should be taken of all objective factors, such as the costs of and the amount of time required for identification, taking into consideration the available technology at the time of the processing and technological developments. The principles of data protection should therefore not apply to anonymous information, namely information which does not relate to an identified or identifiable natural person or to personal data rendered anonymous in such a manner that the data subject is not or no longer identifiable. This Regulation does not therefore concern the processing of such anonymous information, including for statistical or research purposes.

C2: ARTICLE 4(1) GDPR - PERSONAL DATA

(1) ‘personal data’ means any information relating to an identified or identifiable natural person (‘data subject’); an identifiable natural person is one who can be identified, directly or indirectly, in particular by reference to an identifier such as a name, an identification number, location data, an online identifier or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person;

C3: ARTICLE 4(5) GDPR - PSEUDONYMISATION

(5) ‘pseudonymisation’ means 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;

C4: ARTICLE 25 GDPR - DATA PROTECTION BY DESIGN AND BY DEFAULT

(1) Taking into account the state of the art, the cost of implementation and the nature, scope, context and purposes of processing as well as the risks of varying likelihood and severity for rights and freedoms of natural persons posed by the processing, the controller shall, both at the time of the determination of the means for processing and at the time of the processing itself, implement appropriate technical and organisational measures, such as pseudonymisation, which are

designed to implement data-protection principles, such as data minimisation, in an effective manner and to integrate the necessary safeguards into the processing in order to meet the requirements of this Regulation and protect the rights of data subjects.

(2) The controller shall implement appropriate technical and organisational measures for ensuring that, by default, only personal data which are necessary for each specific purpose of the processing are processed. 2That obligation applies to the amount of personal data collected, the extent of their processing, the period of their storage and their accessibility. 3In particular, such measures shall ensure that by default personal data are not made accessible without the individual's intervention to an indefinite number of natural persons.

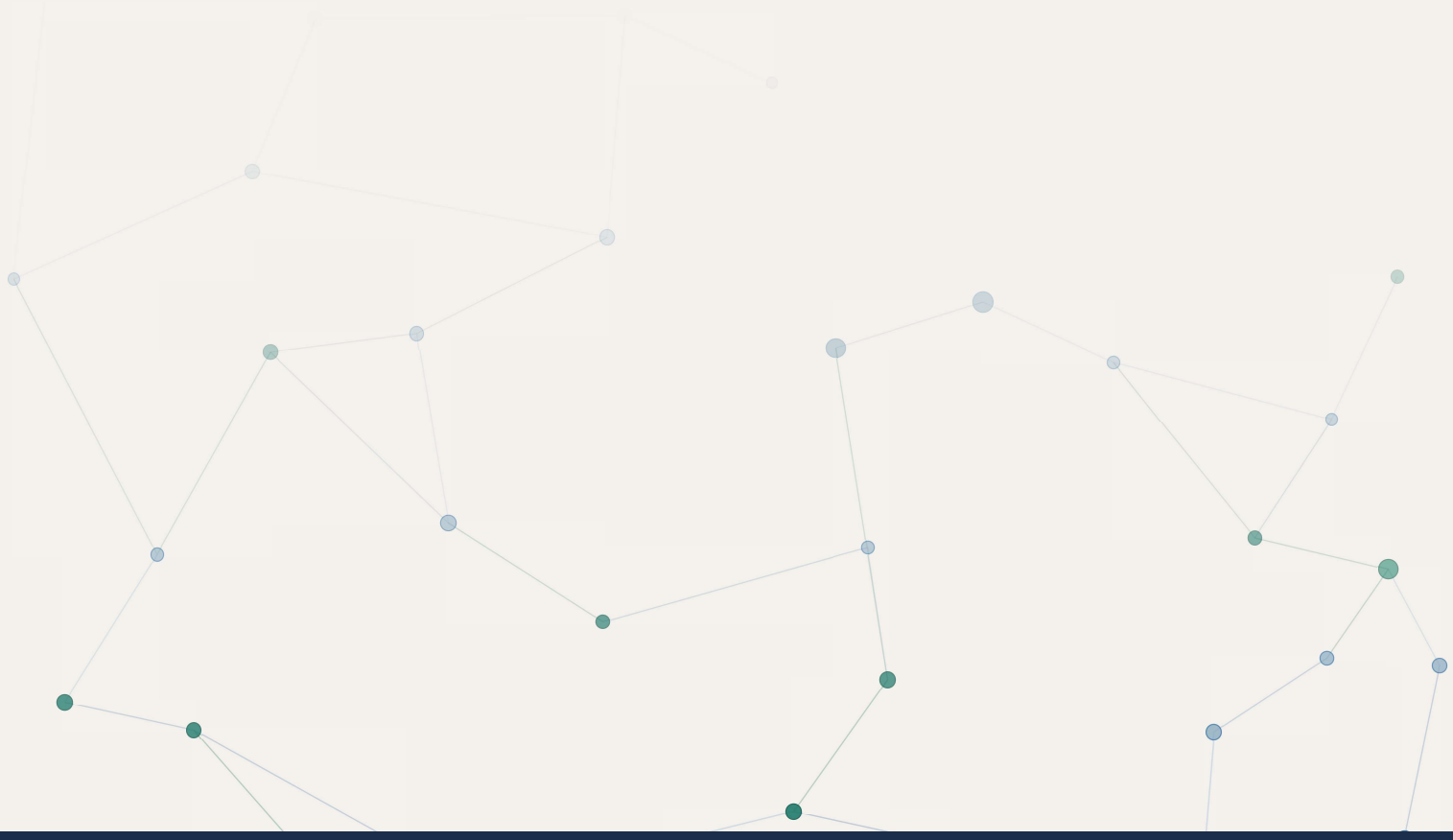
(3) An approved certification mechanism pursuant to Article 42 may be used as an element to demonstrate compliance with the requirements set out in paragraphs 1 and 2 of this Article.

SAMMAN

LAW & CORPORATE AFFAIRS

privacy response team

Blaeu



© 2026 Blaeu Privacy Response Team B.V. and Cabinet Samman. Licensed under CC BY 4.0.