Rillor

EXPLAINER

De-identification, pseudonymization and anonymization compared

De-identification vs anonymization vs pseudonymization: what HIPAA, the GDPR and NIST each mean by the terms, how EU law tests anonymity and why removing names is not enough.



11 November 2026 Data services rillor.com/insights/de-identification-pseudonymization-and-anonymization-compared

Rillor

Contents

01 Summary

02 Background

03 Method

04 Findings

05 Implications

06 Limits of this analysis

07 Sources

Summary

De-identification, pseudonymization and anonymization mean different things in HIPAA, the GDPR and NIST guidance, and HIPAA and the GDPR each set their own test for when data stop identifying a person.

- Pseudonymization replaces identifiers with codes that a separately held key or lookup table links back to the person [4, 11]. The GDPR treats pseudonymized data that such additional information can attribute to a person as personal data [3].
- Under HIPAA, health information is de-identified by Expert Determination, a documented finding that the risk of identification is very small, or by Safe Harbor, which removes 18 kinds of identifiers [1]. The U.S. Department of Health and Human Services (HHS) sets no numeric level for "very small" [2].
- The GDPR does not define "anonymisation". Anonymous information, which does not relate to an identifiable person, falls outside it, judged by the means reasonably likely to be used [3]. NIST SP 800-188 recommends avoiding the word anonymization [4].
- In September 2025 the EU Court of Justice held that pseudonymized data must not be regarded as personal data "in all cases and for every person" [12]. European Data Protection Board guidelines adopted for public consultation in July 2026 assess anonymity from each entity's perspective [10].
- Removing names leaves people identifiable. ZIP code, gender and date of birth likely made 87% of the U.S. population unique in 1990 census data [18, 19], and a 2019 study estimated that 15 demographic attributes would re-identify 99.98% of Americans in any dataset [20].
- Free text and trained models hold identifiers that field rules miss. HHS applies Safe Harbor to free text [2], and a 2023 study found that scrubbing reduces but does not prevent leakage of personal details from language models [28].
- A de-identification claim can be checked on five points: the standard and method, who holds any key or lookup table, whether re-identification was tested, what binds recipients and when the risk was last reassessed [1, 2, 4, 10, 13].

Background

Records about people are shared, licensed and used to train models under labels such as de-identified, pseudonymized and anonymized. The UK Information Commissioner's Office (ICO) observes that "It is common to refer to datasets as 'anonymised' when in fact they still contain personal data, just in pseudonymised form" [17].

The labels come from different traditions. U.S. health privacy law speaks of de-identified health information [1], and the GDPR defines "pseudonymisation" and speaks of anonymous information [3]. NIST IR 8053 noted in 2015 that some authors use de-identification and anonymization interchangeably, others reserve anonymization for de-identification that cannot be reversed, and some healthcare contexts treat de-identification and pseudonymization as equivalent [5]. ISO/IEC 20889, published in November 2018, sets out terminology and a classification of de-identification techniques [33].

The labels decide whether a law applies. HIPAA's Privacy Rule does not restrict the use or disclosure of de-identified health information [2], the GDPR does not concern the processing of anonymous information [3], and the California Consumer Privacy Act (CCPA) excludes deidentified consumer information from personal information [13].

This report is for data engineers, model training teams, privacy leads and dataset buyers. It sets out what laws, regulators, courts and standards say, as research, never legal advice.

Method

What was examined. How HIPAA, the GDPR, the CCPA, NIST and EU and UK regulators define the three terms, the test each sets, the EU decisions of 2025 and 2026, re-identification studies, formal privacy models, and methods for free text and trained models.

Sources. Thirty-six sources, each opened on 8 November 2026: U.S. and EU law [1, 3, 13, 31], the Commission's Digital Omnibus proposal [36], HHS, FTC and ICO publications [2, 14, 15, 17], NIST and ISO publications [4, 5, 6, 7, 8, 33], documents of the European Data Protection Board (EDPB), the European Data Protection Supervisor (EDPS) and the Article 29 Working Party [9, 10, 11, 16, 29, 30], an EU Court of Justice judgment [12], published studies [18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28] and pages from ENISA, an open-source detection toolkit and the U.S. Census Bureau [32, 34, 35].

How. Definitions are quoted from the primary text, keeping each source's spelling ("pseudonymisation", "anonymisation"). The HHS, ICO and ISO sites refuse automated clients, so their pages were read from Internet Archive captures of 26 October 2026, of 28 and 29 August 2026 and of 25 October 2026 [2, 17, 33]. GDPR word counts were made on its English text on EUR-Lex [3]. ENISA's report and ISO/IEC 20889 are described from their summary pages [32, 33]. A dataset, model or toolkit named after a company is described by kind.

Findings

1. The three words have no shared vocabulary

Framework	Term	What it means there	Source
HIPAA Privacy Rule	De-identified	No reasonable basis to believe the information can identify the person, met by Expert Determination or Safe Harbor. No longer protected health information	[1, 2]
GDPR	"Pseudonymisation"	Cannot be attributed to a person without additional information kept separately. Still personal data	[3]
GDPR	Anonymous information	Does not relate to an identifiable person, judged by the means reasonably likely to be used. Outside the Regulation	[3]
NIST SP 800-188	De-identification	General term for removing the association between identifying data and the person	[4, 8]
NIST SP 800-188	Pseudonymization	De-identification that adds an association with one or more pseudonyms	[4]
NIST SP 800-188	Anonymization	Reserved for processes that cannot be reversed. The publication recommends avoiding the word	[4]
CCPA	Deidentified	Cannot reasonably be linked to a consumer, with reasonable measures, a public commitment and contracts binding recipients	[13]
UK ICO	Pseudonymous information	Still personal data	[17]

Table 1. Terms by framework, as each source states them.

The GDPR's English text on EUR-Lex uses "pseudonymisation" 14 times and "anonymous" 3 times, all in Recital 26, and never uses the words "anonymisation" or "de-identification" (counted for this report on 8 November 2026) [3].

NIST's usage moved. Its 2010 guide defined anonymized information as de-identified information "for which a code or other association for re-identification no longer exists" [6]. In 2023, SP 800-188 recommended avoiding the word, because "it is not possible to know whether such an auxiliary dataset exists", and stated

that "there is no technical mechanism to ensure that de-identified data cannot be re-identified" [4]. In Europe, it notes, "the term data anonymization is frequently used as a synonym for de-identification" [4].

The frameworks meet on one point. The GDPR and the ICO treat pseudonymized data as personal data [3, 17], the Article 29 Working Party wrote in 2014 that "pseudonymisation is not a method of anonymisation" [16], and NIST counts pseudonymization as one type of de-identification [4].

2. HIPAA: Expert Determination and Safe Harbor

The Privacy Rule's standard reads: "Health information that does not identify an individual and with respect to which there is no reasonable basis to believe that the information can be used to identify an individual is not individually identifiable health information" [1]. Two methods meet it.

Expert Determination. A person with appropriate knowledge of and experience with generally accepted statistical and scientific principles determines that the risk is very small that an anticipated recipient could use the information, alone or with other reasonably available information, to identify a subject, and documents the methods and results [1]. HHS states: "There is no explicit numerical level of identification risk that is deemed to universally meet the 'very small' level indicated by the method" [2]. No expiration date is required, but some practitioners use time-limited certifications, because technology, social conditions and the availability of information change [2]. Under this method, codes derived from protected health information, including cryptographic hash values, may be disclosed if the keys are not [2].

Safe Harbor. Eighteen kinds of identifiers of the individual and of relatives, employers or household members are removed, from names, geographic subdivisions smaller than a state and all elements of dates except the year for dates directly related to the individual, through record, account and device numbers, IP addresses and full-face photographs, to "Any other unique identifying number, characteristic, or code" [1]. The first three ZIP digits may stay only where that area holds more than 20,000 people, and ages over 89 may be grouped as 90 or older [1]. The covered entity must also have no actual knowledge that what remains could identify a subject [1], and knowing about re-identification studies does not by itself amount to such knowledge [2]. HHS's examples show why the fields matter: year of birth, sex and three-digit ZIP code are unique for about 0.04% of U.S. residents, and date of birth, sex and five-digit ZIP code for over 50% [2].

Neither route reaches zero. HHS writes of both methods: "Although the risk is very small, it is not zero" [2]. De-identified data are free of Privacy Rule limits on use and disclosure, and no data use agreement is required, though a covered entity may ask for one that prohibits re-identification [2].

Re-identification codes and limited data sets. Section 164.514(c) lets a covered entity keep a code to re-identify records, if the code is "not derived from or related to information about the individual" and the mechanism is not disclosed [1]. NIST notes that healthcare settings call pseudonymized data coded [4]. A limited data set excludes 16 direct identifiers but may keep town or city, state and ZIP code. It remains protected health information, usable only for research, public health or health care operations under a data use agreement whose recipient agrees to "Not identify the information or contact the individuals" [1].

3. The GDPR: pseudonymized data stay personal, anonymous data fall outside

Article 4(1) defines personal data as information relating to an identified or identifiable natural person, identifiable directly or indirectly through an identifier such as a name, an identification number, location data or an online identifier [3]. Article 4(5) defines "pseudonymisation" as processing "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 protected by "technical and organisational measures" [3].

Recital 26 draws the line. Pseudonymized data "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" [3]. Identifiability counts "all the means reasonably likely to be used, such as singling out, either by the controller or by another person", weighing cost, time and available technology [3]. The Regulation "does not therefore concern the processing of such anonymous information, including for statistical or research purposes" [3].

Pseudonymization appears as a safeguard in Articles 6(4)(e), 25(1), 32(1)(a) and 89(1), and Recital 28 says it can reduce risks without precluding other measures [3]. The EDPB's 2025 guidelines state that the GDPR "does not impose a general obligation to use pseudonymisation" [11].

The Article 29 Working Party wrote in 2014 that "anonymisation results from processing personal data in order to irreversibly prevent identification" [16]. It named three risks: singling out (isolating records that identify a person), linkability (linking two records about the same person) and inference (deducing an attribute from others) [16]. It added that "removing directly identifying elements in itself is not enough" [16].

The EU AI Act's Article 10(5) allows special categories of personal data for bias detection only where other data, "including synthetic or anonymised data", cannot do the job, under measures that include pseudonymization [31].

The ICO's guidance, published on 28 March 2025, carried a banner in August 2026 saying it is under review after the Data (Use and Access) Act [17]. It states that "Pseudonymous information is still personal data and the law applies to it" [17]. It notes that re-identifying de-identified personal data is a criminal offence under section 171 of the Data Protection Act 2018 [17].

4. Anonymous for whom: the 2025 ruling and the 2026 guidelines

Three EU documents from January 2025 to July 2026 address whether data that one party can identify are personal data for another [10, 11, 12].

January 2025. The EDPB's Guidelines 01/2025 on Pseudonymisation, adopted on 16 January 2025 for public consultation, state that pseudonymized data are personal, and that this "also holds true if pseudonymised data and additional information are not in the hands of the same person" [11]. Even after the pseudonymizing controller erases all additional information, the data count as anonymous only if the conditions for anonymity are met [11].

September 2025. In Case C-413/23 P, EDPS v SRB, decided on 4 September 2025, the Court of Justice held that "pseudonymised data must not be regarded as constituting, in all cases and for every person, personal data", in so far as pseudonymization may effectively prevent persons other than the controller from identifying the data subject [12]. The comments at issue were shared in pseudonymized form with an outside valuation firm. For the controller holding the additional information they were "still personal in nature". For the firm they could be non-personal only if it was "not in a position to lift those measures", and only if the measures also prevented it from attributing the comments by other means, such as cross-checking with other factors [12]. For the duty to inform, identifiability is assessed at the time of collection and from the controller's point of view [12]. The Court set aside the General Court's judgment of 26 April 2023 and referred the case back [12]. The case applied Regulation 2018/1725, which covers EU institutions [12].

July 2026. The EDPB adopted Guidelines 02/2026 on Anonymisation, version 1.0, on 7 July 2026 for public consultation, taking the ruling into account [9, 10]. Anonymity "should be assessed from each relevant entity's perspective", and data may be anonymous for some entities and not for others [10]. Three criteria test the result (Table 2) [10].

Criterion	Met when	Source
No Record Isolation	The data contain no unique combination of attribute values relating to one individual	[10]
No Linkage	No record can be linked to a record about the same individual from a different dataset	[10]
No Inference	No specific and meaningful inference can be drawn from the data	[10]

Table 2. The criteria of EDPB Guidelines 02/2026, version 1.0, adopted for public consultation.

Data meeting all three "can be safely considered anonymous", and "further analysis should be done" when any is not met [9]. An entity's means include those of a third party it could hire, and the guidelines caution against relying on a lack of motivation [10]. Anonymization is itself processing and needs a legal basis under Article 6 [10]. Periodic reassessment of re-identification risk is called good practice [10]. The consultation ran until 30 October 2026, and on 8 November 2026 the EDPB's page showed it closed, with version 1.0 only [9, 10].

The Digital Omnibus proposal. The European Commission's proposal of 19 November 2025, COM(2025) 837, would add to Article 4(1) GDPR that information "shall not be personal for a given entity where that entity cannot identify the natural person to whom the information relates, taking into account the means reasonably likely to be used by that entity" [36]. In a joint opinion adopted on 10 February 2026, the EDPB and the EDPS "strongly urge the co-legislators to not adopt the proposed changes to the definition of personal data" [30].

5. NIST's terms and techniques

SP 800-188 separates direct identifiers, such as names and phone numbers, from quasi-identifiers, which usually match several people but "can be used to triangulate a specific individual" in a linkage attack [4]. Redaction "alone is insufficient to provide formal privacy guarantees" [4]. Table 3 lists its techniques for direct identifiers.

Technique	Example in SP 800-188	Named as pseudonymization
Removal	Field deleted or replaced with NULL	No
Masking	Characters replaced with XXXXXX	No
Encryption	Value encrypted	Yes
Keyed hash	SHA-256 HMAC with a 256-bit randomly generated key	Yes
Keywords	George Washington replaced with PATIENT	No
Realistic surrogates	A name replaced with a plausible substitute	Yes

Table 3. Techniques for direct identifiers. Source: [4].

Keywords can still identify, as in "transforming George Washington to PRESIDENT" [4]. Hashing without a key "generally does not confer security because an attacker can brute force all possible values to be hashed" [4], and NIST's 2010 guide noted that hashing "may not be appropriate for de-identifying information covered by HIPAA" [6]. ENISA's 2019 report analyzes attacks on pseudonymization such as brute force, dictionary search and guesswork [32].

Keys carry the risk. Releasing the lookup table or the secret behind a repeatable transformation compromises identities, and "There is no way to verify that a key has not been released" [4]. Organizations that share one scheme can match records on the pseudonyms, an approach sometimes called privacy-preserving record linkage [4]. SP 800-188 also states that "there is no known method to reliably de-identify genetic sequences", and describes a motivated intruder test, in which competent outsiders try re-identification with public, commercial or private datasets [4]. NIST IR 8053 notes that more aggressive techniques improve privacy while leaving less utility [5].

6. U.S. consumer privacy: the CCPA and the FTC

The CCPA, in its text effective 1 January 2025, defines deidentified as "information that cannot reasonably be used to infer information about, or otherwise be linked to, a particular consumer", provided the business holding it meets three conditions [13]:

1. It takes reasonable measures so that the information cannot be associated with a consumer or household.
2. It publicly commits to keep the information deidentified and not to attempt re-identification, except to test its own processes.
3. It contractually obliges recipients to comply.

Deidentified and aggregate consumer information is not personal information [13]. The CCPA's definition of pseudonymize follows GDPR Article 4(5) closely, with additional information kept separately and protected [3, 13].

The FTC set out the same three protections in March 2012: reasonable measures to de-identify, a public commitment not to re-identify and contracts that prohibit recipients from re-identifying [15]. "The standard is not an absolute one", it wrote, and its techniques included deleting or modifying fields, adding noise, statistical sampling and aggregate or synthetic data [15]. In July 2024 its Office of Technology staff wrote that "data is only anonymous when it can never be associated back to a person", that a hash "still creates a unique signature that can track a person or device over time", and that "Companies should not act or claim as if hashing personal information renders it anonymized" [14].

7. Re-identification evidence

The published studies measure how few attributes single out a person (Table 4).

Study	Data	Attributes	Unique or re-identified	Source
Sweeney, 2000	U.S. population, 1990 census data	5-digit ZIP, gender, date of birth	87% (216 of 248 million)	[18, 19]
Sweeney, 2000	Same	Place, gender, date of birth	53%	[18]
Golle, 2006	U.S. population, 2000 census data	5-digit ZIP, gender, date of birth	63.3%	[19]
Golle, 2006	Same	5-digit ZIP, gender, year and month of birth	4.2%	[19]

Study	Data	Attributes	Unique or re-identified	Source
Rocher and others, 2019	Generative model, tested on 210 populations	15 demographic attributes	99.98% of Americans	[20]
de Montjoye and others, 2013	Mobility of 1.5 million people over 15 months	4 points in time and place	95%	[22]

Table 4. Re-identification and uniqueness figures as each source reports them.

Golle found 0.2% unique with year of birth in place of a full date [19]. Rocher, Hendrickx and de Montjoye's model reached AUC scores of 0.84 to 0.97 for predicting uniqueness, and they concluded that their results "seriously challenge the technical and legal adequacy of the de-identification release-and-forget model" [20]. Mobility uniqueness decays approximately as the 1/10 power of resolution, so "even coarse datasets provide little anonymity" [22].

Linkage needs little. In a dataset of movie ratings from 500,000 subscribers, released for a public prize, an adversary who knows a little about one subscriber can identify that subscriber's record, using a public movie database as background knowledge [21]. NIST IR 8053 recounts how Sweeney re-identified the medical records of the governor of Massachusetts from a release that kept date of birth, ZIP code and sex, matched to the Cambridge voter list she bought for $20, and reports the critique that only 55% of Cambridge's population was registered to vote in 1996-97 [5].

8. Formal models: k-anonymity, differential privacy and synthetic data

k-anonymity. A release provides k-anonymity if each person's information "cannot be distinguished from at least k-1 individuals whose information also appears in the release" [23]. Sweeney's paper also examines re-identification attacks on such releases when accompanying policies are not respected [23]. NIST contrasts the two models: k-anonymity rests on the content of the published data, while differential privacy bounds what can be learned about the confidential data [4].

Differential privacy. NIST SP 800-226, published in March 2025, calls it "a mathematical framework that quantifies privacy loss to entities when their data appears in a dataset" [7]. The loss is bounded for every individual in the data, it is generally achieved by adding random noise, and more noise gives better privacy and less useful results [7]. The ε parameter is commonly treated as a privacy budget, an upper bound on total privacy loss across analyses [7]. The U.S. Census Bureau describes differential privacy, first developed in 2006, as a framework for measuring the disclosure risk of each release, and uses it for 2020 Census respondent data [35].

Synthetic data. The FTC listed synthetic data among de-identification techniques in 2012 [15]. The EDPB's 2026 guidelines say inference concerns synthetic data and models in particular, since querying them can yield specific inferences [10].

9. Free text and trained models

Text. HHS states that the standard "makes no distinction between data entered into standardized fields and information entered as free text", so a listed identifier must be removed wherever it is recognizable [2]. Ambiguous strings need judgment, as in its example "is 5/97 a date or a ratio?" [2]. NIST IR 8053 adds that identifiers in text may not be marked, medical information may be mistaken for personal information and removed, and identifying information may remain after the 18 Safe Harbor elements are gone [5]. It reports a 2013 University of Utah study that found no tool good enough to de-identify VHA records for public release, and a 2013 National Library of Medicine report in which its own tool removed 99.2% of the identifiers matching Safe Harbor on its test set [5].

In the 2014 i2b2/UTHealth shared task on clinical narratives, ten teams submitted 22 system outputs [24]. Three of the ten systems scored F1 above 0.90 and seven above 0.75, and the organizers concluded that automated systems can be very effective while de-identification remains unsolved [24]. The Text Anonymization Benchmark, 1,268 English-language cases of the European Court of Human Rights, marks which spans to mask beyond predefined categories of identifiers [25]. A 2021 review argued for explicit measures of disclosure risk beyond sequence labeling [26]. One open-source toolkit for detecting personally identifiable information (PII) states that its automated detection carries no guarantee of finding all sensitive information [34].

Models. A study first posted in December 2020 extracted hundreds of verbatim training sequences, including public names, phone numbers and email addresses, from a public language model released in 2019, some of them from a single document of its training data [27]. A 2023 study on case law, health care and email data found that "Scrubbing techniques reduce but do not prevent the risk of PII leakage", and that sentence-level differential privacy still leaked about 3% of PII sequences [28]. The EDPB's Opinion 28/2024 of 17 December 2024 considers that "AI models trained with personal data cannot, in all cases, be considered anonymous" [29]. A model counts as anonymous only if the likelihood of extracting personal data, directly or through queries, is insignificant [29].

Implications

For teams preparing data. Name the standard before the method, since HIPAA's two routes, the GDPR's test and the CCPA's three conditions ask different things [1, 3, 13]. Treat keyed hashes, codes and surrogates as pseudonymization, with keys and tables guarded as personal data [4, 11]. Check free text as closely as fields [2].

For buyers and recipients. Ask which standard a claim meets, by which method and judged by whom [1, 10], who holds any key or lookup table [11], whether a motivated intruder test was run [4], what the contract says about re-identification and onward sharing [13, 15], and when the assessment was last repeated [2, 10].

For privacy leads. Three EU documents in eighteen months addressed who must be unable to identify a person [10, 11, 12], and a Commission proposal of November 2025 would change the definition of personal data, a change the EDPB and the EDPS oppose [30, 36]. A record of who held which additional information, and when, bears on each [10, 11, 12, 30].

For teams training models. The EDPB does not treat models trained with personal data as anonymous by default [29], and scrubbing reduces leakage without removing it [28]. Cleaning the training text and testing the trained model are separate checks [27, 28, 29].

Data cleaning. Rillor's training data cleaning removes or replaces personal details to the rules a client sets. Checks run on every version of the cleaned data, and their results ship with it.

Limits of this analysis

- This report is research on what laws, regulators, courts and standards say, never legal advice. It reaches no conclusion on whether any dataset is de-identified or anonymous.
- It covers HIPAA, the CCPA, the FTC, EU law and guidance and UK guidance, and no other jurisdiction.
- EDPB Guidelines 02/2026 are cited as version 1.0 for public consultation [10]. No later version was found on 8 November 2026.
- For the Digital Omnibus, only the proposal's amendment to Article 4(1) GDPR and the joint opinion were read [30, 36]. The legislative outcome was not established.
- The ICO guidance was marked under review in August 2026 [17], and the Data (Use and Access) Act itself was not read.
- HHS, ICO and ISO pages were read from archived captures, and the HHS page shows only a review date [2, 17]. ISO standards and ENISA's report were described from summaries and NIST's quotations [4, 5, 32, 33].
- No source reviewed sets a numeric risk threshold, and HHS states that HIPAA has none [2].
- Uniqueness figures describe U.S. census populations of 1990 and 2000 and particular datasets [18, 19, 20, 21, 22], and text figures come from 2013 and 2014 evaluations [5, 24].

Sources

1. 45 CFR 164.514, Other requirements relating to uses and disclosures of protected health information. eCFR, Office of the Federal Register, current text as amended through 78 FR 34266 (7 June 2013), viewed 8 November 2026. https://www.ecfr.gov/current/title-45/subtitle-A/subchapter-C/part-164/subpart-E/section-164.514

2. Guidance Regarding Methods for De-identification of Protected Health Information in Accordance with the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule. U.S. Department of Health and Human Services, Office for Civil Rights, content last reviewed 3 February 2025, read from the Internet Archive capture of 26 October 2026, viewed 8 November 2026. https://www.hhs.gov/hipaa/for-professionals/special-topics/de-identification/index.html https://web.archive.org/web/20261026152833/https://www.hhs.gov/hipaa/for-professionals/special-topics/de-identification/index.html

3. Regulation (EU) 2016/679 (General Data Protection Regulation). EUR-Lex, Publications Office of the European Union, 27 April 2016, viewed 8 November 2026. https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:32016R0679

4. De-Identifying Government Datasets: Techniques and Governance, NIST SP 800-188. Garfinkel, Near, Dajani, Singer and Guttman, National Institute of Standards and Technology, September 2023, viewed 8 November 2026. https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.800-188.pdf

5. De-Identification of Personal Information, NISTIR 8053. Garfinkel, National Institute of Standards and Technology, October 2015, viewed 8 November 2026. https://nvlpubs.nist.gov/nistpubs/ir/2015/NIST.IR.8053.pdf

6. Guide to Protecting the Confidentiality of Personally Identifiable Information (PII), NIST SP 800-122. McCallister, Grance and Scarfone, National Institute of Standards and Technology, April 2010, viewed 8 November 2026. https://nvlpubs.nist.gov/nistpubs/Legacy/SP/nistspecialpublication800-122.pdf

7. Guidelines for Evaluating Differential Privacy Guarantees, NIST SP 800-226. Near, Darais, Lefkovitz and Howarth, National Institute of Standards and Technology, March 2025, viewed 8 November 2026. https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.800-226.pdf

8. de-identification, glossary term. NIST Computer Security Resource Center, undated, viewed 8 November 2026. https://csrc.nist.gov/glossary/term/de_identification

9. EDPB sheds light on anonymisation and web scraping for generative AI and adopts final version of guidelines on blockchain. European Data Protection Board, 8 July 2026, viewed 8 November 2026. https://www.edpb.europa.eu/news/edpb-sheds-light-on-anonymisation-and-web-scraping-for-generative-ai-and-adopts-final-version_en

10. Guidelines 02/2026 on Anonymisation, Version 1.0, adopted for public consultation. European Data Protection Board, 7 July 2026, viewed 8 November 2026. https://www.edpb.europa.eu/system/files/2026-09/edpb_guidelines_202602_anonymisation_v1_en_0.pdf https://www.edpb.europa.eu/documents/guideline/guidelines-022026-on-anonymisation_en

11. Guidelines 01/2025 on Pseudonymisation, version for public consultation. European Data Protection Board, 16 January 2025, viewed 8 November 2026. https://www.edpb.europa.eu/system/files/2025-

01/edpb_guidelines_202501_pseudonymisation_en.pdf

12. Judgment of the Court (First Chamber), Case C-413/23 P, EDPS v SRB. Court of Justice of the European Union, EUR-Lex, 4 September 2025, viewed 8 November 2026. https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:62023CJ0413

13. California Consumer Privacy Act of 2018, statute text effective 1 January 2025. California Privacy Protection Agency, January 2025, viewed 8 November 2026. https://cppa.ca.gov/regulations/pdf/ccpa_statute.pdf

14. No, hashing still doesn't make your data anonymous. U.S. Federal Trade Commission, Office of Technology staff, Technology Blog, 24 July 2024, viewed 8 November 2026. https://www.ftc.gov/policy/advocacy-research/tech-at-ftc/2024/07/no-hashing-still-doesnt-make-your-data-anonymous

15. Protecting Consumer Privacy in an Era of Rapid Change: Recommendations for Businesses and Policymakers. U.S. Federal Trade Commission, March 2012, viewed 8 November 2026. https://www.ftc.gov/sites/default/files/documents/reports/federal-trade-commission-report-protecting-consumer-privacy-era-rapid-change-recommendations/120326privacyreport.pdf

16. Opinion 05/2014 on Anonymisation Techniques (WP216). Article 29 Data Protection Working Party, 10 April 2014, viewed 8 November 2026. https://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2014/wp216_en.pdf

17. Anonymisation, UK GDPR guidance (About this guidance, and Introduction to anonymisation). Information Commissioner's Office, 28 March 2025, read from Internet Archive captures of 28 and 29 August 2026, viewed 8 November 2026. https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/data-sharing/anonymisation/ https://web.archive.org/web/20260828221210/https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/data-sharing/anonymisation/ https://web.archive.org/web/20260828221221/https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/data-sharing/anonymisation/about-this-guidance/ https://web.archive.org/web/20260829220305/https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/data-sharing/anonymisation/introduction-to-anonymisation/

18. Simple Demographics Often Identify People Uniquely, Data Privacy Working Paper 3. Sweeney, Carnegie Mellon University, 2000, viewed 8 November 2026. https://dataprivacylab.org/projects/identifiability/paper1.pdf

19. Revisiting the Uniqueness of Simple Demographics in the US Population. Golle, ACM Workshop on Privacy in the Electronic Society (WPES '06), author copy, 30 October 2006, viewed 8 November 2026. https://crypto.stanford.edu/~pgolle/papers/census.pdf

20. Estimating the success of re-identifications in incomplete datasets using generative models. Rocher, Hendrickx and de Montjoye, Nature Communications 10, 3069, 23 July 2019, viewed 8 November 2026. https://www.nature.com/articles/s41467-019-10933-3

21. How To Break Anonymity of the Netflix Prize Dataset. Narayanan and Shmatikov, arXiv:cs/0610105, version 2, 22 November 2007 (first version 18 October 2006), viewed 8 November 2026. https://arxiv.org/abs/cs/0610105

22. Unique in the Crowd: The privacy bounds of human mobility. de Montjoye, Hidalgo, Verleysen and Blondel, Scientific Reports 3, 1376, 25 March 2013, viewed 8 November 2026. https://www.nature.com/articles/srep01376

23. k-anonymity: a model for protecting privacy. Sweeney, International Journal on Uncertainty, Fuzziness and Knowledge-based Systems 10(5), 557-570, 2002, author copy at the Data Privacy Lab, viewed 8 November 2026. https://dataprivacylab.org/dataprivacy/projects/kanonymity/kanonymity.pdf

24. Automated systems for the de-identification of longitudinal clinical narratives: Overview of 2014 i2b2/UTHealth shared task Track 1. Stubbs, Kotfila and Uzuner, Journal of Biomedical Informatics 58 (Suppl), S11-S19, author manuscript in PubMed Central, 28 July 2015, viewed 8 November 2026. https://pmc.ncbi.nlm.nih.gov/articles/PMC4989908/

25. The Text Anonymization Benchmark (TAB): A Dedicated Corpus and Evaluation Framework for Text Anonymization. Pilán, Lison, Øvrelid, Papadopoulou, Sánchez and Batet, arXiv:2202.00443, version of 1 July 2022 (first version 25 January 2022), viewed 8 November 2026. https://arxiv.org/abs/2202.00443

26. Anonymisation Models for Text Data: State of the art, Challenges and Future Directions. Lison, Pilán, Sanchez, Batet and Øvrelid, ACL-IJCNLP 2021, pages 4188-4203, Association for Computational Linguistics, August 2021, viewed 8 November 2026. https://aclanthology.org/2021.acl-long.323/

27. Extracting Training Data from Large Language Models. Carlini and 11 others, arXiv:2012.07805, version of 15 June 2021 (first version 14 December 2020), viewed 8 November 2026. https://arxiv.org/abs/2012.07805

28. Analyzing Leakage of Personally Identifiable Information in Language Models. Lukas, Salem, Sim, Tople, Wutschitz and Zanella-Béguelin, IEEE S&P 2023, arXiv:2302.00539, version of 23 April 2023 (first version 1 February 2023), viewed 8 November 2026. https://arxiv.org/abs/2302.00539

29. Opinion 28/2024 on certain data protection aspects related to the processing of personal data in the context of AI models. European Data Protection Board, 17 December 2024, viewed 8 November 2026. https://www.edpb.europa.eu/system/files/documents/2024-12/edpb_opinion_202428_ai-models_en.pdf

30. EDPB-EDPS Joint Opinion 2/2026 on the Proposal for a Regulation as regards the simplification of the digital legislative framework (Digital Omnibus). European Data Protection Board and European Data Protection Supervisor, 10 February 2026, viewed 8 November 2026. https://www.edpb.europa.eu/system/files/documents/2026-02/edpb_edps_jointopinion_202602_digitalomnibus_en.pdf

31. Regulation (EU) 2024/1689 (Artificial Intelligence Act), Article 10. EUR-Lex, Publications Office of the European Union, 13 June 2024 (Official Journal 12 July 2024), viewed 8 November 2026. https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:32024R1689

32. Pseudonymisation techniques and best practices. ENISA, 3 December 2019, publication page, viewed 8 November 2026. https://www.enisa.europa.eu/publications/pseudonymisation-techniques-and-best-practices

33. ISO/IEC 20889:2018, Privacy enhancing data de-identification terminology and classification of techniques. International Organization for Standardization, November 2018, standard page, read from the Internet Archive capture of 25 October 2026, viewed 8 November 2026. https://www.iso.org/standard/69373.html https://web.archive.org/web/20261025033624/https://www.iso.org/standard/69373.html

34. Presidio, documentation home page. Presidio project (open source), undated, viewed 8 November 2026. https://presidio.dataprivacystack.org/

35. Understanding Differential Privacy. U.S. Census Bureau, 2020 Census disclosure avoidance, undated, viewed 8 November 2026. https://www.census.gov/programs-surveys/decennial-census/decade/2020/planning-management/process/disclosure-avoidance/differential-privacy.html

36. Proposal for a Regulation (Digital Omnibus), COM(2025) 837 final, procedure 2025/0360(COD). European Commission, 19 November 2025, viewed 8 November 2026. https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:52025PC0837

About this report

Published by Rillor on 11 November 2026. Superintelligence (SI) tools assist research and drafting, and every figure cites its source. Rillor builds and secures SI systems, supplies the data they learn from and delivers the infrastructure they run on. We apply SI to financial and compute markets.

Cite as: Rillor. De-identification, pseudonymization and anonymization compared. 11 November 2026. https://rillor.com/insights/de-identification-pseudonymization-and-anonymization-compared

Work with Rillor

Send the question the data should answer, the sources you know of, the coverage you need and how often it should update. Contact us at rillor.com/contact?type=custom-data or contact@rillor.com. More on data services: rillor.com/data