A principled approach to defining anonymization as applied to EU data protection law

[discussion draft, May 10, 2022]

Micah Altman, Aloni Cohen, Francesca Falzon, Evangelia Anna Markatou, Kobbi Nissim, Michel José Reymond, Sidhant Saraogi, and Alexandra Wood[1]

1. Introduction

Privacy and data protection laws often conceive of some process—called anonymization[2] or, alternatively, de-identification[3]—by which regulated data can be transformed into unregulated data by destroying the link between the data and the individuals to which the data relate. The concept of anonymization plays a central role in data protection law, defining a broad category

[1] Authors are listed in alphabetical order. We describe the authors' contributions following a standard taxonomy. See Liz Allen et al., 'Credit Where Credit Is Due' (2014) 508 Nature 312–13. MA, KN, and AW authored the first draft of the manuscript and had overall responsibility for content and revisions. MA, AC, KN, and AW contributed to the conception of the report (including core ideas and statement of research questions), to the methodology, and to the writing through critical review and commentary. All authors contributed to substantiating the paper's arguments, revision, critical review, and commentary.

Micah Altman is Research Scientist at the Center for Research in Open and Equitable Scholarship at MIT. Aloni Cohen is Assistant Professor of Computer Science and Data Science at the University of Chicago. Francesca Falzon is a Ph.D. student at the University of Chicago and a visiting research fellow at Brown University 2021-2022. Evangelia Anna Markatou is a Ph.D. student at Brown University. Kobbi Nissim is Professor and McDevitt Chair in Computer Science at Georgetown University and affiliated with Georgetown Law. Michel José Reymond is a lawyer at the firm Byrne-Sutton Bollen Kern. Sidhant Saraogi is a Ph.D. student at Georgetown University. Alexandra Wood is a Fellow at the Berkman Klein Center for Internet & Society at Harvard University.

The writing of this report was partially supported by a gift to the McCourt School of Public Policy and Georgetown University.

[2] For example, in the EU, the General Data Protection Regulation governs the processing of 'personal data' but not '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'. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation), Recital 26.

[3] For example, in the United States, the California Privacy Rights Act, which goes into effect on 1 January 2023, governs the use of 'personal information' but not 'consumer information that is deidentified', meaning 'information that cannot reasonably be used to infer information about, or otherwise be linked to, a particular consumer'. Cal. Civ. Code §§ 1798.140(v)(3), (m) [effective 1 Jan. 2023]. As another example, the Health Insurance Portability and Accountability Act Privacy Rule governs the use of 'protected health information' but not '[h]ealth 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'. 45 C.F.R. § 164.514(a).

of information that falls outside the scope of regulation,[4] and thereby enabling companies, government agencies, and researchers to carry out a wide range of data processing activities.[5] Yet, despite the significance of the concept, it is undertheorized and poorly articulated in regulatory guidance. Moreover, insights about anonymization from the academic privacy and computer science communities have seen little adoption in regulatory settings, despite the increasing volume and availability of sensitive data.

The essential function of anonymization is to provide protection against others' attempts to learn private information specific to individuals from an analysis or release of data. Anonymization is not a panacea; for instance, it cannot, even in theory, be used to ensure that decisions or algorithms based on personal data will be secure, non-discriminatory, explainable, reasonable, nor immune to dangerous misuse.[6] And, in practice, current legal and technical anonymity safeguards are concerned mostly with the anonymity of individual participants, and do not protect the anonymity of marginalized communities, groups of genetically-related families, or other groups.[7]

This article puts forth principles for the regulation of anonymization, and for data protection regulation more broadly. It also provides model language as a starting point for explicitly incorporating these principles into data protection guidance. These principles are grounded in the past 20+ years of research in data privacy. These principles are not intended as absolutes, but better anonymization techniques and regulations will generally satisfy more of these principles - or are the principles intended to be exhaustive.

It is useful to ground this discussion within a particular regulatory framework. As a foil for the proposed principles, we use the most well-developed treatment of the concept of data anonymization in regulatory guidance available today, namely two opinions from the EU's Article 29 Data Protection Working Party Group.[8] The more recent opinion on anonymization techniques specifically breaks down anonymization into protection from three types of attacks:

[4] Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation), Recital 26.

[5] Article 29 Data Protection Working Party, ‘Opinion 05/2014 on Anonymisation Techniques’ (2014) WP 216, 3.

[6] See eg, Micah Altman, Alexandra Wood, and Effy Vayena, ‘A harm-reduction framework for algorithmic fairness’ (2018) 16(3) IEEE Security & Privacy 34-45 (demonstrating how unfair decisions may readily arise from privacy-protected decision processes). For a general review of fairness detections and methods from a technical viewpoint, see Dana Pessach and Erez Shmueli, ‘A Review on Fairness in Machine Learning’ (2022) 55(3) ACM CSUR 1-44. For a provocative perspective on reasonable and explainable inference, see Sandra Wachter and Brent Mittelstadt, ‘A right to reasonable inferences: re-thinking data protection law in the age of big data and AI’ (2019) 2019(2) Colum. Bus. L. Rev. 494-620. For a discussion of the regulation of scientific research that poses ethical and technical concerns separate from privacy, such as research on diseases involving the collection of information from humans and having a high potential for misuse, see for example, Michael J. Selgelid, ‘Governance of dual-use research: an ethical dilemma’ (2009) 87 Bulletin of the World Health Organization 720-723.

[7] For a broad discussion of issues of group privacy in the context of European privacy and data protection law, see Linnet Taylor, Luciano Floridi, and Bart Van der Sloot (eds), 126 Group privacy: New challenges of data technologies (Springer 2016).

[8] Opinion 4/2007 on the concept of personal data and Opinion 05/2014 on Anonymization Techniques.

singling out (also referenced in Recital 26 of GDPR), linkability, and inference.[9] It also makes specific determinations as to whether a handful of data disclosure limitation techniques sufficiently protect against singling out, linkability, and inference. The Working Party's treatment is uniquely rich and concrete, allowing for its rigorous assessment in light of the accumulated theoretical and practical knowledge in the area of privacy protection.

Hence, while this article presents principles which are intended to be universal, the discussion and model language presented are tailored to the European setting—specifically, the GDPR. And though the prior treatment by the Article 29 Data Protection Working Party has significant room for improvement, it is our foil because it is the best guidance currently available.

1.1. The EU Anonymization Guidance

The most developed exploration of the concept of data anonymization in regulation and guidance comes from the Article 29 Data Protection Working Party's interpretations of the EU's Data Protection Directive, the 1995 predecessor to the GDPR. Specifically, their *Opinion 04/2007 on the Concept of Personal Data* and *Opinion 05/2014 on Anonymisation Techniques* provide interpretive guidance on the definition of personal data, anonymous information, and on methods by which personal data can be effectively "rendered anonymous." The GDPR replaced the Article 29 Data Protection Working Party with the European Data Protection Board.[10] However, the Working Party's Opinions 04/2007 and 05/2014 remain a highly-influential interpretive source informing recent guidance and decisions from the European Data Protection Board and a number of national supervisory authorities with respect to the meaning of these concepts in the context of the GDPR.[11]

In the time since the Working Party's Opinions 04/2007 and 05/2014 were released, the data privacy landscape has changed dramatically. It is now recognized that common approaches to processing personal data that have generally been considered to pose little risk may in fact carry significant risks to individual privacy.[12] Consider the following examples of privacy vulnerabilities across a wide range of systems:

- Product recommendations used by websites such as Amazon can leak the activities (i.e., purchases) of other users on the website.[13]

[9] Opinion 05/2014 (n 5)

[10] European Data Protection Board, `Endorsement 1/2018' (25 May 2018).

[11] See eg, European Data Protection Board, Binding decision 1/2021 on the dispute arisen on the draft decision of the Irish Supervisory Authority regarding WhatsApp Ireland under Article 65(1)(a) GDPR (28 July 2021); CNIL [France], Sheet n°1: Identify personal data (11 June 2020); Data Prot. Comm'n [Ireland], Guidance Note: Guidance on Anonymisation and Pseudonymisation (June 2019).

[12] See eg, Hongsheng Hu et al., 'Membership Inference Attacks on Machine Learning: A Survey' (2021) ACM Computing Surveys; Maria Rigaki and Sebastian Garcia, 'A Survey of Privacy Attacks in Machine Learning' (2021).

[13] Joseph A. Calandrino et al., '"You Might Also Like:" Privacy Risks of Collaborative Filtering' (2011) Proc of IEEE Symposium on Security and Privacy 231-246.

- Neural networks and other machine learning models often unintentionally memorize and later leak their training data (e.g., as a suggested completion for the phrase “My social-security number is").[14]
- Publications of statistical tables such as those released from the 2010 US Decennial Census can enable the reconstruction and subsequent re-identification of individual responses.[15]
- Advertising systems of social media platforms can be exploited to infer private information about individual users, such as information that users tag as visible to "Only Me" on Facebook.[16]
- Release of statistics on gene mutation (allele) frequencies in a DNA mixture can enable outsiders to infer whether an individual’s DNA was present in the mixture.[17]

The growing availability of massive datasets and the emergence of new modes of privacy attacks put pressure on traditional approaches to anonymization and data protection. Against this backdrop, scholars and practitioners have argued that the existing anonymization guidance creates uncertainty for practitioners and that further clarity is needed[18] (or alternatively, that the whole concept of anonymization should be abandoned[19]).

This article recognizes that it is critical to ensure clarity and consistency in the practice of anonymization within a rapidly evolving data privacy landscape. We argue that this goal can best be achieved by applying principles from the scientific study of data privacy that have been devised to provide strong, general protection across different contexts. In this article, we propose a collection of principles, as well as specific recommendations and model language for updating the EU anonymization guidance based on such principles.

1.2. The Need for Definitions that Are Legally Clear, Technically Sound, and Generally Applicable

Anonymization is a concept that inherently carries both legal and technical meaning, and ensuring clarity will require harmonizing both its legal and technical understandings. Further, the technical aspects of anonymization must be well defined in order to enable enforcement within software systems. The Article 29 Working Party has made great strides towards putting forward definitions which are legally clear, technically sound, and actionable. Nevertheless, research

[14] Nicholas Carlini, Chang Liu, Úlfar Erlingsson, Jernej Kos and Dawn Song, ‘The secret sharer: evaluating and testing unintended memorization in neural networks’ (2019) Proc of 28th USENIX Security Symposium 267–284; Gavin Brown, Mark Bun, Vitaly Feldman, Adam Smith and Kunal Talwar, ‘When is Memorization of Irrelevant Training Data Necessary for High-Accuracy Learning?’ (2021) STOC 123-132.

[15] Simson Garfinkel, John M. Abowd and Christian Martindale, ‘Understanding Database Reconstruction Attacks on Public Data’ (2018) 16(5) ACMQueue 1-26.

[16] Aleksandra Korolova, ‘Privacy Violations Using Microtargeted Ads: A Case Study’ (2010) Proc of IEEE ICDMW.

[17] Nils Homer et al., ‘Resolving Individuals Contributing Trace Amounts of DNA to Highly Complex Mixtures Using High-Density SNP Genotyping Microarrays’ (2008) 4(8) PLoS Genetics e1000167.

[18] See eg, Sophie Stalla-Bourdillon and Alison Knight, ‘Anonymous Data v. Personal Data — A False Debate: An EU Perspective on Anonymization, Pseudonymization and Personal Data’ (2017) 34 Wis. Int’l L.J. 284, 285-6.

[19] See Paul Ohm, ‘Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization’, 57 UCLA L. REV. 1701, 1703 (2010).

from the last decade demonstrates that the current definitions and guidance, as elaborated in Opinions 04/2007 and 05/2014, may not be technically sound as they arguably endorse the use of anonymization systems that have serious vulnerabilities, as illustrated in Example 1.2.1.

Example 1.2.1. Diffix is a commercial anonymization system developed by a company called Aircloak and was specifically designed to satisfy the WP29's guidance on anonymization. Its design was carefully documented and justified with specific reference to the guidance, and provided perhaps the most thorough public-facing legal analysis for any commercial anonymization product. Aircloak's marketing materials stated that the French data protection authority 'evaluated Aircloak's Diffix framework and determined that it satisfies the three criteria of the opinion 05/2014 of the WP29 on Anonymization Techniques for all use cases'.[20]

Researchers subsequently demonstrated that Diffix was vulnerable to reconstruction attacks.[21] That is, an analyst using Diffix could perfectly infer certain attributes in the underlying data without much difficulty. It was shown that reconstruction attacks were still possible even after Aircloak modified Diffix to prevent these specific attacks.[22]

The case of Diffix demonstrates that existing anonymization guidance is either not legally clear or not technically sound. Were the guidance both technically sound and legally clear, a platform that underwent as much vetting as Diffix would not have allowed for the reconstruction of individuals' data.

Furthermore, disputes have emerged among supervisory authorities regarding whether particular systems render personal data anonymous, as illustrated by Example 1.2.2.

Example 1.2.2. In a recent draft decision, the Irish supervisory authority concluded, in part, that WhatsApp IE's use of a specific technique, referred to as a 'lossy hash', qualified non-users' data as 'anonymised data'.[23] Eight other supervisory authorities objected that the lossy hash process does not constitute effective anonymization because WhatsApp IE could use additional information to identify non-users' data.[24] In response, the Irish supervisory authority acknowledged that, while there is a 'greater-than-zero risk that some non-users could be re-identified by inference, linking or singling out', a 'zero-risk approach is likely to result in very few, if any, processes achieving anonymisation.'[25]

The EDPB issued a binding decision, finding 'that given the means and the data which are

[20] Aircloak, 'Building Trust' (5 July 2017) Medium <https://medium.com/@aircloak/building-trust-d8d341431a8f> [https://perma.cc/3HKE-MVJH].
[21] Aloni Cohen and Kobbi Nissim, 'Linear Program Reconstruction in Practice' (2020) 10(1) Journal of Privacy and Confidentiality.
[22] Aloni Cohen, Sasho Nikolov, Zachary Schutzman and Jonathan Ullman, 'Reconstruction Attacks in Practice' (27 October 2020) DifferentialPrivacy.org Blog <https://differentialprivacy.org/diffix-attack>.
[23] Binding decision 1/2021 (n 9), 7-8.
[24] Ibid. 6, 20.
[25] Ibid. 25.

available to WhatsApp IE and are reasonably likely to be used, its capacity to single out data subjects is too high to consider the dataset anonymous'.[26] Notably, the EDPB maintained that the 'network of connections between users and non-users, and thereby indirectly among users, constitutes a sort of topological signature of lossy hashes which becomes fairly unique as the dimension of the network and the number of connections grows', which can 'substantially increase the re-identification risk of data subjects'.[27]

The WhatsApp IE dispute demonstrates the need for clearer guidance regarding what constitutes effective anonymization in light of the growing availability of massive data sources, evolving privacy-enhancing technologies, new modes of privacy attack, and the fact that any processing of data about individuals necessarily carries some risk to privacy.

Moreover, even when guidance is clear it can be technically unsound, as illustrated by Example 1.2.3.

Example 1.2.3. Recital 26 of the GDPR identifies 'singling out' as one type of privacy attack that anonymization must prevent. In prior work, some of the authors have argued that the existing guidance understands singling out as the ability to 'isolate': to identify a set of attributes that distinguishes an individual from all other individuals in the data underlying a given data release. But singling out as isolation is an unworkable theory. It ignores a high inherent risk of isolation present in all data releases---even those that are unquestionably anonymous.[28]

Working Party guidance states that a technique called k-anonymity (and variants like l-diversity and t-closeness) prevent singling out attacks when properly used.[29] This was recently reiterated by the EDPB.[30] But researchers have demonstrated the possibility of carrying out successful singling-out attacks against data releases that satisfy k-anonymity and its variants.[31]

Harmonizing the treatment of anonymization techniques with respect to the data protection rules will require definitions that are legally and technically sound, actionable, and generally applicable in settings where personal information is processed. In the sections that follow, we introduce design principles and specific model language to guide the development of definitions and guidance that hold these properties.

[26] Ibid. 30.
[27] Ibid. 32 (citations omitted).
[28] Aloni Cohen and Kobbi Nissim, 'Towards formalizing the GDPR's notion of singling out' (2020) 117(15) Proc. Natl. Acad. Sci., 8344-8352; Micah Altman, Aloni Cohen, Kobbi Nissim and Alexandra Wood, 'What a hybrid legal-technical analysis teaches us about privacy regulation: The case of singling out' (2021) 27(1) B.U. J. Sci. & Tech. L. 1.
[29] Opinion 05/2014 (n 2).
[30] Binding decision 1/2021 (n 9), 32.
[31] Cohen and Nissim (n 30); Altman et al. (n 30); Aloni Cohen, 'Attacks on Deidentification's Defenses' (2022) Proc of 31st USENIX Security Symposium; Srivatsava Ranjit Ganta, Shiva Kasiviswanathan, and Adam Smith, 'Composition Attacks and Auxiliary Information in Data Privacy' (2008) Proc of 14th ACM SIGKDD 265–273.

2. Recommendations: General Principles

In this section, we propose general desiderata for anonymization guidance and the data protection rules more broadly. While we do not argue that it is necessary for all aspects of the guidance to meet every one of the following principles, we contend that stronger anonymization mechanisms will satisfy more of these principles.

2.1. Principle 1: Process Protection (Define Data Protection in Terms of the Processing of Personal Data, Not the Results of the Processing)

A key insight from computer science research is that privacy is a property of the informational relationship between the input and output of an analysis, not a property of the output alone. This motivates our first principle: When evaluating whether a data release can be made public one needs to consider the particulars of the computational process used to produce that release. Guidance that considers only the output of computations, rather than the computations themselves, represents a critically incomplete theory of privacy protection and is likely to fail to provide systematic, reliable, and future-proof protection.

We illustrate this principle with a stylized example.[32]

Example 2.1.1. Consider a release of the following statistic: ‘a representative ninth-grade GPA at City High School is 3.5’. One might naturally think that this statistic is unlikely to reveal private information about an individual student. However, one must know how the statistic was computed in order to make that determination. For instance, if the representative ninth-grade GPA was calculated by taking the GPA of the alphabetically first student in the school, then the statistic completely reveals the GPA of that student. Alternatively, a representative statistic could instead be based on the most common features of the ninth graders in the school, such as using the most common first name, the most common last name, the average age, and the average GPA to produce ‘John Smith, a fourteen-year-old in the ninth grade, has a 3.1 GPA’. Suppose that coincidentally a student named John Smith subsequently joins the ninth-grade class. Although a name identical to his appears in the published statistic, one knows with certainty that the statistic does not reveal private information about him because it was not based on his student records in any way.

Further, the principle applies to any mode of data release, including interactive mechanisms (see also Principle 2.2). Consider, for example, a dataset accessed via a query interface that allows an analyst to ask questions and receive answers. One approach to privacy that has been considered is called *query auditing*. Query auditing involves some mechanism that creates a log of queries and their answers. The auditing mechanism examines each new query in light of

[32] Example 2.1.1 is adapted from Alexandra Wood et al., ‘Differential Privacy: A Primer for a Non-Technical Audience’ (2020) 21(1) Vanderbilt Journal of Entertainment and Technology Law 209-276, and Alexandra Wood, Micah Altman, Kobbi Nissim and Salil Vadhan, ‘Designing Access with Differential Privacy’ in Shawn Cole, Iqbal Dhaliwal, Anja Sautmann and Lars Vilhuber (eds), *Handbook on Using Administrative Data for Research and Evidence-based Policy* § 6.2.2 (Abdul Latif Jameel Poverty Action Lab 2021).

previous queries and their answers to determine whether answering the new query would reveal information about any individual in the database. If that is the case, the mechanism denies the query; otherwise, it answers the query accurately. Perhaps surprisingly, researchers have demonstrated that the mechanism's decision to deny a query may itself leak information based on how it determines which queries to deny.[33] This is illustrated in Example 2.1.2,

Example 2.1.2. Suppose an analyst knows the names and gender of individuals and seeks to learn their ages from the personal data in a database by interacting with a query-auditing system. The analyst first asks for the maximum age among the three individuals in the following table.

Name	Krzysiek	Gabriel	Frida
Gender	M	M	F
Age	44	21	75

The answer (75) reveals the age of somebody in the table, but does not associate the age with a specific person. Hence, the query-auditing system answers the analysts question: 75.

Now suppose that the analyst next asks for the maximum age among the males in the table. On its own, the answer (44) would be allowed following the same reasoning as before. However, because the answer differs from the previous answer of 75, the two answers together reveal that Frida, the lone female, must be 75 years old. As such, a query-auditing mechanism would deny the second question after already answering the first.

But from this denial itself, the analyst can already conclude that Frida's age is 75. If instead one of the men was 75 years old, then the answer to the second question would also have been 75. In this case, the two answers together would not reveal any specific individual's age and hence would have been answered. The second query was denied only because Frida is the 75 year old, rather than one of the men.

Model Language

We recommend that, where a regulation or guideline defines a requirement for anonymization, it should explicitly incorporate this principle, as in Model Language 1.

Model Language 1. Effective anonymization is determined by the process or mechanism used to produce outputs. *Anonymization may be considered effective only*

[33] See Krishnaram Kenthapadi, Nina Mishra and Kobbi Nissim, 'Denials leak information: Simulatable auditing' (2013) 79(8) J. Comput. Syst. Sci., 1322-1340.

when the data processing itself is analyzed to guarantee that the informational relationships the processing creates between personal data and the outcome results in limited risks of singling out, linkability, and inference.

2.2. Principle 2: Format Neutrality (Define Data Protection Generally Regardless of Data Release Format)

To be effective, data protection definitions and mechanisms should be generally applicable and interpretable for any type of data release—regardless of whether that release is in the form of microdata, a summary table, an information visualization, statistical model coefficients, a trained model output by a machine-learning algorithm, a textual summary, or any other form. Disclosure risks are less obvious—but no less serious—when derivatives of some data are released, but not the data themselves. Data derivatives (i.e., any information derived from underlying data that does not necessarily have an apparent one-to-one correspondence with the underlying data) can be simple tables of aggregate statistics, like demographic data published by statistical agencies, statistics on allele frequencies, or results of query-response systems. They can also be highly complex, such as black-box AI models, including language models like GPT-3, recommender systems like Amazon's “You might also like" feature, and synthetic data.[34][35]

The alternative—i.e., limiting the scope of data protection rules to certain data types, such as releases of individual records—is short-sighted. The rapidly-evolving data landscape continuously yields new types of data analyses and releases, and new privacy attacks reveal vulnerabilities with respect to new data formats. Furthermore, exempting certain types of data formats is likely to incentivize actors to merely change the form of the data use or release without meaningfully mitigating attendant data protection risks.

There is already some recognition among policymakers that the format of the data is less important than its contents. For instance, the Working Party's Opinion 04/2007 clarifies that ‘the concept of personal data includes information available in whatever form, be it alphabetical, numerical, graphical, photographical or acoustic, for example’.[36]

However, some traditional data protection techniques, as well as many existing laws and guidance documents,[37] assumed, explicitly or implicitly, that aggregation and summarization can

[34] Synthetic data is data generated algorithmically to match the characteristics of a population (so as to allow for the application of statistical and machine learning analyses) while preserving its individual members’ privacy.

[35] See eg, the membership inference attack on the CTGAN synthetic data generation algorithm, which is used for example by the MIT Synthetic Data Vault Project. Theresa Stadler, Bristena Oprisanu, and Carmela Troncoso, ‘Synthetic Data – Anonymisation Groundhog Day’ (2022) Proc 31st USENIX Security Symposium (preprint available on arXiv <https://arxiv.org/abs/2011.07018>); CTGAN GitHub repository <https://github.com/sdv-dev/CTGAN>; Laboratory for Information and Decision Systems, ‘The real promise of synthetic data’ (16 October 2020) MIT News.

[36] Opinion 04/2007 (n 6).

[37] See eg, GDPR, Recital 162; Opinion 05/2014 (n 2); California Consumer Privacy Act of 2018, Cal. Civ. Code § 1798.140(o)(3); Cable Communications Policy Act of 1984, 47 U.S. Code § 551(a)(2)(A); Privacy of Consumer Financial Information Rule Under the Gramm-Leach-Bliley Act, 16 C.F.R. § 313.3(o)(2)(ii).

effectively reduce or even eliminate risk. This assumption has since been shattered both theoretically and practically. Theoretically, all computations, unless totally useless, come at some price to privacy. Given a sufficient number of informative summary results, the data underlying these summary results may be reconstructed completely or partially, regardless of the protection mechanism used.[38] Practically, the ability to glean sensitive individual information from aggregate data has been demonstrated in a large number of examples, including every one of the aforementioned types of data derivatives, as illustrated above in Section 1.1, and in the following detailed example (Example 2.2.1).

Example 2.2.1. In 2018, an internal study by the US Census Bureau found that the published statistical tables from the 2010 US Decennial Census could be used to narrow down the possible values of individual-level records and thereby reconstruct the underlying confidential data from respondents.[39] They found that the reported sex, age, race, ethnicity, and fine-grained geographic location could be reconstructed for 46% of the US population (or 71% of the population when allowing age to vary by up to one year).[40] They were also able to assign personally identifiable information to individual records using commercial databases, with confirmation that these re-identifications were accurate for 52 million people, or 17% of the US population.[41] This illustrates that even publications of aggregate statistical tables protected using statistical disclosure limitation techniques may be vulnerable to serious privacy attacks.

The idea that data derivatives generally protect privacy is simply untenable. While data derivatives may obscure the relationship with the underlying data, they do not destroy it. Powerful inference and AI techniques penetrate the veil in unanticipated ways.

Example 2.2.4. Synthetic data generation is often touted as a way to release statistically accurate datasets while preserving the privacy of the individuals in the dataset.[42] However, as with any other overly accurate query response algorithm, they are susceptible to database reconstruction attacks which can lead to re-identification. Moreover, as demonstrated by

[38] This counterintuitive rule has come to be known as the fundamental law of information recovery, as coined by Cynthia Dwork and Aaron Roth, ‘The Algorithmic Foundations of Differential Privacy’ in 9(3-4) Foundations and Trends in Theoretical Computer Science (2014). This fundamental theory of database reconstruction was established in Irit Dinur and Kobbi Nissim, ‘Revealing Information while Preserving Privacy’ (2003) Proc of ACM PODS 202 and refined in Cynthia Dwork, Frank McSherry and Kunal Talwar, ‘The Price of Privacy and the Limits of LP Decoding’ (2007) STOC; Shiva Prasad Kasiviswanathan, Mark Rudelson, Adam Smith and Jonathan Ullman, ‘The price of privately releasing contingency tables and the spectra of random matrices with correlated rows’ (2010) STOC 775-784; Shiva Prasad Kasiviswanathan, Mark Rudelson and Adam Smith, ‘The Power of Linear Reconstruction Attacks’ (2013) ACM-SIAM SODA; S. Muthukrishnan and Aleksandar Nikolov, ‘Optimal Privacy Halfspace Counting via Discrepancy’ (2012) STOC 1285-1292; Cynthia Dwork, Adam Smith, Thomas Steinke, Jonathan Ullman and Salil Vadhan, ‘Robust traceability from trace amounts’ (2015) Proc of IEEE FOCS.

[39] Garfinkel et al. (n 14).

[40] John Abowd, ‘Stepping-up: The Census Bureau tries to be a good data steward in the 21st century’ (4 March 2019) Presentation at the Simons Institute for the Theory of Computing.

[41] Ibid.

[42] Ibid. 33..

Stadler et al.,[43] commonly used synthetic data generation algorithms such as the CTGAN algorithm[44] are susceptible to membership inference attacks. Given black-box access to the generation algorithm, an adversary may be able to identify a single data point used to train the algorithm. Moreover, this can lead to re-identification of individuals corresponding to the data point. The authors note that outliers in the dataset are especially vulnerable to such attacks.

Example 2.2.5. A social network can be represented as nodes (corresponding to individuals) and edges between pairs of nodes (corresponding to social links between the respective individuals). One naive way of anonymizing such a social network is by replacing the personal data labeling the nodes (e.g. actual names, email addresses, etc.) with a random identifier. This would preserve the underlying structure of the social links, while removing personal identifiable information. However, Backstrom et al.[45]—which was recently cited by the EDPB in its binding decision discussed in Example 1.2.2 as a mode of re-identification to be taken into consideration when assessing whether a dataset is personal data—demonstrates that this anonymization technique is insufficient by describing three attacks that can successfully recover the identity of a set of users in the network. For example, they show that, by creating as few as seven new accounts and creating social links to a small number of targeted individuals (e.g, by sending messages), the identity of those targeted individuals can be recovered within the social network.

The principle of format neutrality applies even when there is no single data release as such. Many times, data is only accessible through an interactive query system. A user or analyst may issue queries about some sensitive underlying data, and the interactive system will respond with an answer. It is easy to believe that such systems are much more protective than if the data were simply published, but this is not always the case. **Example 2.1.2** already illustrated one way that such interactive systems can fail. The Diffix system described in **Example 1.2.1** was also an interactive query system that allowed individual-level attributes to be discovered by an analyst making queries---despite being specifically designed to provide GDPR-level anonymization guarantees.

Model Language

We propose the following model language based on this principle.

Model Language 2. Effective anonymization applies generally regardless of data format. *Any informative data output carries risk and this risk accumulates with each analysis or publication. It is insufficient to limit data protection to individual-level records and datasets alone. Interactive mechanisms, aggregated data, statistical summaries, data derivatives and*

[43] Ibid. 34.
[44] Ibid. 34.
[45] Lars Backstrom et al. 'Wherefore art thou r3579x?: anonymized social networks, hidden patterns, and structural steganography' (2007) Proc of 16th International Conference on World Wide Web.

algorithmic uses of data equally carry risk to anonymization and require protection. To be considered protective, the effectiveness of anonymization measures should be demonstrated for all types of data releases, regardless of the release format and of whether information is exposed intentionally (eg, when publishing aggregate statistics) or unintentionally (eg, when the protocol inadvertently leaks information), and regardless of whether the release mechanism is static or interactive.

2.3. Principle 3: Composition Awareness (Control Composition Risks)

Any analysis or release of information carries disclosure risk, and this risk accumulates with each analysis or release.[46] As a consequence, one cannot describe privacy as being preserved or not preserved; rather, it falls along a continuum.

The simulated reconstruction of the 2010 US Decennial Census data (Example 2.2.1) and the learning of an individual's personal attributes by combining multiple queries to a database (Example 2.1.2) are examples of attacks leveraging composition effects. As another illustration, Example 2.3.1 below highlights the fragility of k-anonymity with respect to post-processing and demonstrates how the improper composition of privacy techniques can increase the potential harm associated with a data release.

Example 2.3.1. A recent attack revealed vulnerabilities in a k-anonymized research dataset created by the online learning platform edX.[47] The edX dataset was de-identified in order to comply with FERPA. The Harvard and MIT research team that performed the deidentification—including experts in statistics and privacy guided by the general counsel of their respective institutions—applied k-anonymity with k=5 to students' data[48]. Trying to protect the student records from different types of attackers, the team effectively applied k-anonymity in multiple different, overlapping ways. Cohen argues that doing so ignored composition risks and undermined the k-anonymity guarantees. As a result, 245 students are unique in the dataset—i.e, they do not enjoy any protection from k-anonymity. Moreover, 1.7% (120) of all edX students who posted on course discussion forums are unambiguously identifiable by some of their classmates. Without using composition, Cohen also re-identified 3 edX students by matching them with their public LinkedIn profiles. This re-identification led to attribute disclosure: each of these 3 students had also failed to complete at least one edX courses

[46] Aaron Fluitt, et al., 'Data Protection's Composition Problem' (2019) 5 Eur. Data Prot. L. Rev. 285.

[47] Cohen (n 33).

[48] A data release is k-anonymized if the information provided for each individual data is identical to the data released for k-1 or more other individuals in the data release. k-anonymity is achieved by means of data suppression and generalization. See Pierangela Samarati and Latanya Sweeney, "Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression". *Harvard Data Privacy Lab*, 1998; Latanya Sweeney, Achieving k-Anonymity Privacy Protection Using Generalization and Suppression. Int. J. Uncertain. Fuzziness Knowl. Based Syst. 10(5): 571-588 (2002).

Composability is a property of technical privacy concepts and mechanisms that enables one to reason about—and thereby manage and control—composition effects on privacy in a modular way.[49] If an approach to preserving privacy is composable, then the composition effects of multiple data uses employing the approach can be understood by analyzing each data use in isolation. Differential privacy is an example of an anonymization technique that is composable, as the combination of multiple differentially private analyses is also differentially private.

Absent the use of a composable mechanism, two independent releases of data that each carry minimal privacy risks can together create catastrophic privacy failures. As the volume and complexity of data uses and publications grow rapidly across a broad range of contexts, it has become impossible to monitor all past data releases and anticipate all future attacks. Therefore, we argue that composability is a necessary requirement for any proposed anonymization mechanism to ensure it remains protective against future data releases and attacks.

Model Language

We recommend that anonymization guidance should explicitly require the composition of risk to be controlled, as provided in the following model language.

Model Language 3. Effective anonymization requires controlling composition risks. *To be considered protective, an anonymization technique must guarantee that, when outputs are produced by applying such technique to the data, the risk of releasing protected output contributes minimally to future risks to the anonymity of any included individual, even those future risks involving unknown future processing of, or data releases based on the same data.*

2.4. Principle 4: Assumption Minimization (Limit Assumptions Regarding Downstream Users and Uses)

Assumptions regarding downstream uses or users of anonymized data should be minimized in order to enhance clarity and ensure strong protection of personal data. A key insight from the scientific literature on privacy is that adopting a wide conception of potential attackers and attack modes is essential to ensuring strong privacy protection that can withstand current privacy attacks as well as those that will emerge in the future. As an increasingly expansive body of privacy attacks has demonstrated, attacks have revealed surprising, often difficult to foresee, vulnerabilities that can be exploited with readily available technology.

Though it applies to other data protection contexts as well, this principle is particularly important in the context of data anonymization. Typically, the purpose of data anonymization is to create data that is free from further regulation. Thus, one cannot assume that the initial intended use of anonymous data will be the only use of the data---once data is 'rendered anonymous' it may be freely used. It may survive the processor that created it and be used for unforseen purposes: It

[49] Fluitt et al. (n 44).

may be sold to third parties;it may get published online, where it could remain for years and years, etc. As such, it would be short sighted to consider privacy risks stemming from today's intended use in light of today's reasonable attacks and today's external information. Protecting the data subject requires considering tomorrow's unintended uses in light of tomorrow's unforseen attacks and tomorrow's unknowable external information.

The examples in Section 2.2, such as how the use of recommendation systems and publication of statistical tables have been shown to leak underlying private data about individuals, often in surprising ways. Moreover, as illustrated by Example 2.4.1, where current guidance does not adhere to this principle, it contributes to a lack of clarity regarding the concept of personal data.

Example 2.4.1. The Working Party's Opinion 4/2007, in clarifying the meaning of the 'relating to' element of the concept of personal data, refers to an example of the value of a particular house. The opinion explains this piece of information may be considered information relating to an individual under certain circumstances depending on how the information will be used.[50] Where 'this information will be used solely to illustrate the level of real estate prices in a certain district', the data protection rules 'will clearly not apply'; however, where it will be used to determine an individual's property tax liability, it 'should be considered as personal data'.[51]

Defining whether information relates to an individual based on how it will be used in the future arguably creates uncertainty for practitioners evaluating whether information should be considered as personal data. In practice, it is often difficult—if not impossible—to anticipate and constrain future users and uses of a piece of information after it has been disclosed. Moreover, incorporating limiting assumptions into the definition of personal data, such as that the information to be released will be used only as expected, is likely to result in a weak standard that fails to recognize the expansive risks to data protection in the modern data ecosystem.

Such assumptions also play a role in approaches to assessing re-identification risk recommended in guidance from supervisory authorities, such as the 'motivated intruder' test described by the UK Information Commissioner's Office, as explained in Example 2.4.2.

Example 2.4.2. In 2012 guidance, the UK Information Commissioner's Office (ICO) describes the 'motivated intruder' test it applies—and recommends other organizations apply—when carrying out an assessment of the risk of re-identification.[52] The ICO characterizes this test as one that 'sets the bar for the risk of identification higher than considering whether a "relatively inexpert" member of the public can achieve re-identification, but lower than considering whether someone with access to a great deal of specialist expertise, analytical power or prior

[50] Ibid. 9 (Example No. 5).
[51] Ibid.
[52] Information Commissioner's Office, *Anonymisation: Managing Data Protection Risk, Code of Practice* (November 2012).

knowledge could do so'.[53] Components of this test rely on assumptions about the motivated intruder—i.e., that the motivated intruder 'is reasonably competent, has access to resources such as the internet, libraries, and all public documents, and would employ investigative techniques such as making enquiries of people who may have additional knowledge of the identity of the data subject or advertising for anyone with information to come forward' but 'is not assumed to have any specialist knowledge such as computer hacking skills'.[54] By relying on such limiting assumptions about the attacker, this results in a weak standard for anonymization.

Model Language

We propose the following model language based on this principle.

Model Language 4. Effective anonymization requires minimizing assumptions about downstream users and uses of information. *Because personal data can be learned from a data release in unanticipated ways, to be considered protective, the effectiveness of measures for ensuring anonymization should depend minimally on any assumption regarding the potential users and uses of the data.*

2.5. Principle 5: Inclusion-based Protection (Define Anonymization Standards Based on How Participants' Information Affects the Data Release)

The result of data processing should be considered to relate to an individual only when it reveals information about that specific individual *as a result of the inclusion of their information* in the processing.[55]

Conversely, the result of data processing does not relate to an individual if the result of processing does not reveal information about the individual, or if information is revealed only as a byproduct of revealing information about an entire population. For example, publishing the average weight of a large population, or a general statistical relation that holds across the population (such as smoking increases the risk of cancer), does not reveal information that relates to an individual. For a full illustration of this principle, consider Example 2.5.1.

Example 2.5.1. Attacks have demonstrated that personal data can potentially be inferred from the output of machine learning algorithms.[56] Researchers have shown that, given black-box access to certain facial recognition algorithms, one can reconstruct blurry but recognizable

[53] Ibid. 23.
[54] Ibid. 22-23.
[55] The idea of inclusion-based protection is informed by the study of formal privacy models and in particular differential privacy, see Wood et al. (n 31).
[56] Carlini et al. (n 11); Michael Veale, Reuben Binns and Lilian Edwards, 'Algorithms that Remember: Model Inversion Attacks and Data Protection Law' (2018) 376 Philosophical Transactions of the Royal Society A 20180083.

versions of the face shots of an individual in the training dataset using only the individual's name (where the individual's name was used as a label prediction).[57]

If images of Bob were used to train a facial recognition algorithm and an attack produces a reconstruction of a recognizable image of his face, this is a violation of Bob's privacy because it could only be achieved due to the inclusion of his images in the training data. Alternatively, suppose images of Bob were not used to train the algorithm. If an attacker learns from the model that people who have gray hair are more likely to wear glasses—and therefore Bob, who has gray hair, is more likely to wear glasses—this information does not violate Bob's privacy because it was learned without his data being used in the analysis. Further, if adding Bob's images to the training dataset does not change the attacker's inference from the model that people who have gray hair are more likely to wear glasses, this inference also does not violate Bob's privacy because the inclusion of his information did not change the result of the data processing.

When evaluating whether the result of data processing should be considered to relate to an individual, one useful test is whether the result may be used to determine whether the individual's information was included, or not included, in the data processing. If the individual's inclusion or non-inclusion can be inferred, the result of the processing should not be considered anonymized. Note, it is not the individual's inclusion or non-inclusion itself that matters. Rather, it is whether the results of the processing are changed noticeably by the individual's inclusion in the processing.

Statistical information about large populations does not generally reveal that each individual's information was included in the data processing. Moreover, there are techniques for releasing such statistics while provably hiding any single individual's contribution (i.e., differential privacy). Similarly, the result of data processing relates to a group when it may be used to determine whether the group's information was included, or not included, in the data processing.

It also follows from this principle that the information to be protected is the underlying information about the participants in the analysis, not the released dataset itself. In other words, anonymization is concerned with what is revealed about the information serving as input to the processing. Accordingly, anonymization techniques such as k-anonymity which restrict only the form of the output—but not its informational relationship to the input—are vulnerable to serious privacy attacks.

The principle of inclusion-based protection takes into account the fact that learning general statistical relationships about populations is unavoidable in the era of big data and artificial intelligence. In fact, this type of learning is the *raison d'être* of data processing—driving much of modern research, policymaking, and innovation. We argue that such analyses are consistent with, for instance, Recital 157 of the GDPR, which recognizes the value of research data

[57] Matt Fredrikson, Somesh Jha and Thomas Ristenpart, 'Model Inversion Attacks that Exploit Confidence Information and Basic Countermeasures' (2015) Proc of 22nd ACM SIGSAC CCS.

registries and aims to support researchers' use of personal data 'to obtain essential knowledge about the long-term correlation of a number of social conditions such as unemployment and education with other life conditions', subject to appropriate conditions and safeguards.[58]

An alternative to inclusion-based protection is to consider statistical knowledge about a greater population as information relating to every individual included in said population, whether included in the processing or not. The result would be self-defeating. An individual's data protection rights would allow them to enjoin other individuals from participating in research studies, public record releases, and other types of processing that enable learning information about large populations and general statistical relationships. This would render anonymization itself an empty concept.

Inclusion-based protection can be used to provide further guidance on the requirement that data must 'relate to' a person in order to fall within the meaning of personal data under the GDPR, ie, 'any information relating to an identified or identifiable natural person', as illustrated by Example 2.5.2.[59]

Example 2.5.2. Working Party Opinion 4/2007 offers three alternative ways to consider that data can 'relate to' an individual—in content, in purpose, or in result.[60] The 'content' element regards whether the information is about the person, for instance 'the results of medical analysis clearly relate to the patient, or the information contained in a company's folder under the name of a certain client clearly relates to him'. The 'purpose' element relates to whether the data 'are used or are likely to be used . . . with the purpose to evaluate, treat in a certain way or influence the status or behaviour of an individual'. The 'result' element applies when the use of the information 'is likely to have an impact on a certain person's rights and interests'.[61]

We argue that, without further clarification, plausible interpretations of personal data based on the three elements described in Opinion 4/2007 could lead to an overbroad definition of personal data that encompasses types of data processing that do not depend on an individual's participation.[62] For example, publishing a finding such as 'smoking increases the risk of cancer' could be considered a release of personal data with respect to the 'purpose' element because the information may be used to influence of the behavior of an individual, as in the case of cigarette labels that carry warnings about the health risks of smoking with the aim of reducing cigarette usage. Similarly, this publication could also be considered a release of personal data with respect to the 'result' element because, if, for example, the finding is

[58] GDPR, Recital 157.

[59] GDPR, art. 4(1).

[60] Ibid. 10.

[61] Ibid. 10-11.

[62] For a similar argument that the three ways to consider that data can 'relate to' an individual can be subject to overbroad interpretations and, specifically, that 'any information can relate to a person by reason of purpose, and all information relates to a person by reason of impact', see Nadezhda Purtova, 'The law of everything. Broad concept of personal data and the future of EU data protection law' (2017) 10 Law, Innovation and Technology 40-81.

used to calculate life insurance premiums, a smoker could be assessed a higher premium as a result of the publication of this information.

We argue that the requirement that whether data ‘relates to’ an individual should depend on whether the data reveal information about an individual based on their inclusion in the data processing. 'Relates to' should not be interpreted so broadly that it considers a release of information about an entire population (such as the average weight) or about general statistical relationships (such as smoking increases the risk of cancer) as personal data.

For an additional illustration applying the inclusion-based protection principle to evaluating whether the result of data processing relates to an individual, consider Example 2.5.3.

Example 2.5.3. Personalized medicine models. Consider a statistical model of the correlation between certain genetic markers in a patient and the appropriate dosing for an anticoagulant drug Warfarin.[63] Suppose William is a patient who takes Warfarin but whose data was not used in the creation of the statistical model. In a 2014 paper, researchers showed that it is possible to use such a statistical model to infer the genetic markers of a patient like William from his Warfarin dosage.[64]

Applying the principle of inclusion-based protection to this example, the statistical model itself should not be considered as personal data relating to William.[65] The statistical model explains the population-level correlation between dosage and genetic markers, i.e., a type of learning that is not based on a given individual’s participation in the analysis.[66] Considered alone, the model reveals no information about William as a result of his inclusion in the processing that created the model. It cannot, as William was not included in that processing. Similarly, even if William’s information had been included in the processing that created the model, it reveals no information about William, provided that the result of the processing has not been noticeably affected by the inclusion of his information.

The inferences drawn about William's genetic markers---the result of applying the statistical model to William's Warfarin dosage---do relate to William. They reveal information about William as a result of the inclusion of William's dosage (i.e., his personal data) in the application of the model to his personal data---an instance of processing distinct from the processing that created the model itself.

Model Language

[63] Matthew Fredrikson et al., ‘Privacy in Pharmacogenetics: An End-to-End Case Study of Personalized Warfarin Dosing’ (2014) Proc. of 23rd USENIX Security Symposium.
[64] Ibid.
[65] For a detailed explanation of why this example should not be considered a privacy violation, see Frank McSherry, ‘Statistical inference considered harmful’ (14 June 2016), Blog post <https://github.com/frankmcsherry/blog/blob/master/posts/2016-06-14.md>.
[66] Ibid.

Definitions used in data protection standards should explicitly incorporate the principle of inclusion-based protection (hence, that privacy is not violated by merely learning about large populations), as provided in Model Language 5.

Model Language 5. Effective anonymization requires defining the anonymization standards based on how participants' information affects the result of data processing. *When evaluating whether the result of data processing should be considered to relate to an individual, the result should be considered to relate to an individual only when it reveals information about that specific individual as a result of the inclusion of their information in the processing. Conversely, the result of data processing should not be considered as relating to an individual if the result of processing does not reveal information about the individual, or if information is revealed only as a byproduct of revealing information about an entire population. For example, publishing the average weight of a large population, or a general statistical relation that holds across the population (such as smoking increases the risk of cancer), does not reveal information that relates to an individual.*

The principle of inclusion-based protection can also be applied to model language regarding group privacy, as provided in Model Language 6.

Model Language 6. Anonymization standards in the context of group privacy should be defined based on how group members' information affects the result of data processing. *The result of data processing relates to a group when it may be used to determine whether information specific to members of the group was included, or not included, in the data processing.*

2.6. Principle 6: Transparency (Use Public Protection Mechanisms and Reject Privacy by Obscurity)

A basic tenet of computer security is that the security of a system should not rely on the secrecy of the mechanisms that protect it.[67] The alternative—'security through obscurity'—is rejected as shortsighted, misguided, and generally less secure.[68] Accordingly, it is considered best security practice that computer systems should use public algorithms and protocols built and vetted by the greater security community.[69] Following the same principle, the protection of personal data must not depend on the secrecy of the mechanisms used. Ideally, the processing of personal

[67] This is the well-known Kerckhoffs' principle, i.e., 'the security of a cryptosystem must lie in the choice of its keys only; everything else (including the algorithm itself) should be considered public knowledge'. Auguste Kerckhoffs, 'La cryptographie militaire' [Military cryptography] (February 1883) IX Journal des sciences militaires [Military Science Journal] 161–191 (in French). This principle was reformulated or independently stated by Claude Shannon as Shannon's Maxim for systems design: 'one ought to design systems under the assumption that the enemy will immediately gain full familiarity with them'. Claude Shannon, 'Communication Theory of Secrecy Systems' (4 October 1949) 28(4) Bell System Technical Journal 662. See also Niels Ferguson, Bruce Schneier, and Tadayoshi Kohno, Cryptography Engineering: Design Principles and Practical Applications (2015) Sec. 2.2.1.

[68] Steven Bellovin and Randy Bush, 'Security Through Obscurity Considered Dangerous' (February 2002) Internet Engineering Task Force (IETF) <https://www.cs.columbia.edu/~smb/papers/draft-ymbk-obscurity-00.txt>.

[69] Jonathan Katz and Yehuda Lindell. Introduction to Modern Cryptography (3rd ed.) (2021) 5-6.

data that carries disclosure risk should use public protocols and algorithms that were scrutinized and vetted by the privacy community. At the very least, the data controller should have a justified belief that privacy is not undermined if the details of protocols and algorithms used to process the data are exposed.

Using public protection mechanisms has many benefits, including improvements in both protection and utility. As scholars Steven Bellovin and Randy Bush have observed, '[t]he long history of cryptography and cryptanalysis has shown time and time again that open discussion and analysis of algorithms exposes weaknesses not thought of by the original authors', 'allows the users to protect themselves', and 'encourages general protection and repair strategies'.[70] Analogously, privacy research has revealed numerous examples of severe weaknesses found in mechanisms designed to provide strong protection. For example, the US Census Bureau's 2010 disclosure avoidance system whose details are carefully guarded---see Example 2.6.2 further below---proved vulnerable to large-scale reconstruction as described in Example 2.2.1 above. Scholars and practitioners are increasingly rejecting the historical practice of concealing the precise details of protection mechanisms—either because the protection mechanisms had known weaknesses that enable adversaries to infer information about the data from the configuration parameters of the protection mechanism or as a general attempt to create 'privacy through obscurity'.[71]

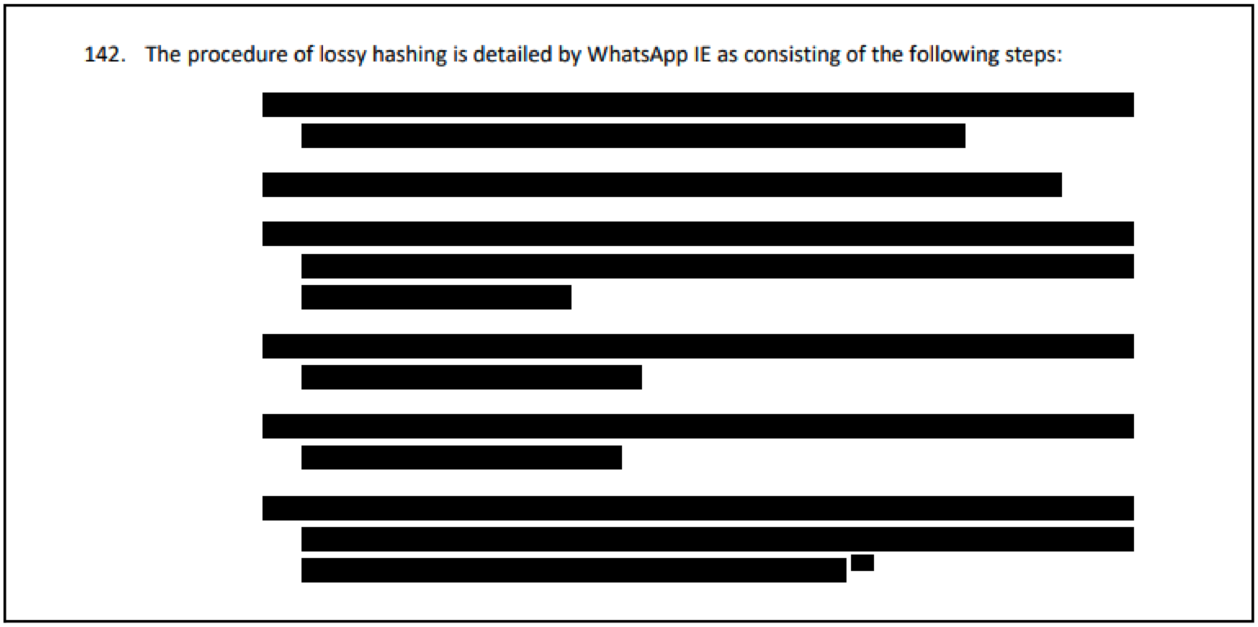
Example 2.6.1. At the center of the WhatsApp IE dispute, discussed above in Example 1.2.2, was the use of a lossy hash algorithm as an anonymization technique.[72] In its binding decision, the EDPB referred to the procedure but redacted the details of the algorithm in full, as illustrated in the excerpt below.[73] This practice prevents third parties from evaluating the effectiveness of the lossy hash algorithm and likely contributes to greater uncertainty for practitioners as well as weaker protection of privacy and personal data generally.

[70] Bellovin and Bush (n 48).

[71] John M. Abowd and Ian M. Schmutte, 'Economic analysis and statistical disclosure limitation' (2016) 2015.1 Brookings Papers on Economic Activity 221-293.

[72] Binding decision 1/2021 (n 9).

[73] Ibid. 29.



142. The procedure of lossy hashing is detailed by WhatsApp IE as consisting of the following steps:

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

Making anonymization protocols and algorithms public also makes it possible for analysts and data consumers to take into account facts such as the magnitude of noise added for privacy protection, enabling them to make better recommendations, as illustrated in Example 2.6.2.

Example 2.6.2. One of the disclosure avoidance methods used by the US Census Bureau in the 1990, 2000, and 2010 Decennial Censuses was data swapping, ie, ‘the practice of switching the values of a selected set of attributes for one data record with the values reported in another record’ with the goal of ‘protect[ing] the confidentiality of sensitive values while maintaining the validity of the data for specific analyses’.[74] Given a record about a 6-member household in Boston and a record of a 7-member household in Cambridge, the published data, after swapping, may show that there is a 7-member household in Boston and a 6-member household in Cambridge. Because publicly revealing the swap ratet—as well as the details of many other traditional statistical disclosure limitation (SDL) techniques—increases disclosure risk, statistical agencies ‘do not publish them or release more than a few details of their swapping procedures’.[75]

Additionally, this has effects on the results of data analyses using swapped data, and concealing the details of the swapping algorithm makes it difficult to measure those effects and take them into account.[76] In fact, the Census Bureau has analyzed the effects of swapping on the quality of its published statistics but ‘has not published its evaluation results due to concerns that they might compromise the SDL procedures themselves’.[77]

[74] Laura McKenna, ‘Disclosure avoidance techniques used for the 1970 through 2010 decennial censuses of population and housing’ (2018).
[75] Abowd and Schmutte (n 52), 231, 233.
[76] Ibid. 233.
[77] Ibid. 237.

Model Language

We recommend the following model language based on this principle.

Model Language 7. Effective anonymization requires ensuring the mechanisms in use provide sufficient protection whether or not their design details are exposed. *The processing of personal data that carries disclosure risk should use public protocols and algorithms which were scrutinized and vetted by the privacy community. At the very least, the data controller should have a justified belief that privacy is not undermined if the details of protocols and algorithms used to process the data are exposed. A best (but not sufficient) practice is to make specification documents, design documents, and well-documented code available for public inspection. Because the analysis of anonymized data by parties other than the data controller generally requires details of how those data were created, the publication of the design details of the anonymization mechanism serves the additional purpose of making the data more usable.*

3. Recommendations: Revised Definitions

This section is planned to include an analysis of the three criteria for effective anonymization identified in the Article 29 Working Party Opinion 05/2014, namely linkability, singling out, and inference[78] and provide model language for each. As an illustration of this analysis, the current draft includes an application of the principles identified in Section 2 above to the concept of inference.

3.1 Defining Inference

Opinion 05/2014 defines inference broadly as 'the possibility to deduce, with significant probability, the value of an attribute from the values of a set of other attributes'.[79] This definition of inference does not explicitly distinguish between inferences of information about a specific individual as a result of the inclusion of the individual's information in the data processing, and other types of inferences. As argued in Principle 5, only the former type of inferences should be considered as relevant to anonymization, as the latter type may non-privacy harming inferences such as 'smoking causes cancer'. Taken at the extreme, the Working Party's definition of inference could be read to include inferences such as 'people who are over 50 years old are also over 40 years old' within the definition of personal data. As explained in Section 2.5, updated guidance should clearly exclude this type of overbroad interpretation of inference.

While no formal technical definition of inference is provided, Opinion 05/2014's reference to the deduction of the value of one attribute from the values of other attributes echoes the concept known both as 'inferential disclosure' and 'probabilistic attribute disclosure' within the classical

[78] Opinion 05/2014 (n 2), 11-12.
[79] Opinion 05/2014 (n 2), 11-12.

disclosure control literature.[80] By Duncan and Lambert, such disclosure happens when an intruder reasons about the release made available by a statistical agency, together with other information, to learn (maybe without certainty) the value of a respondent-reported attribute which the statistical agency attempted to remove from the release.[81] That the Working Party view of inference is related to inferential disclosure or probabilistic attribute disclosure is supported by the Opinion's conclusion that '[t]he main improvement of l-diversity and t-closeness over k-anonymity is that it is no longer possible to set up inference attacks against a "l-diverse" or "t-close" database with a 100% confidence'.[82]

While still regularly used, inferential disclosure protection as currently practiced violates the principles presented in Section 2. Moreover, the core concept of inferential disclosure (ie, not merely its practice) is incompatible with the principles of defining anonymity based on how participants' data affects a release, and of managing composition coherently.

The common interpretation of inferential disclosure—and the way protections against inferential disclosure are implemented in practice—violates Principle 1: Process Protection. For example, Opinion 05/2014 states that l-diversity and t-closeness prevent inference attacks.[83] The intuition underlying this conclusion is based on the assumption that a privacy attacker would only attempt to pair between an individual and the sensitive information they are matched with, which is l-diverse or t-close to the underlying distribution. But this intuition only makes sense if one focuses on the output alone and not the mechanism itself (nor the informational relationship between inputs and outputs that the mechanism provides). For example, datasets are often made l-diverse or t-close using algorithms that generalize and suppress attributes that may be used to identify individual data subjects. Many of these algorithms are potentially vulnerable to downcoding, a type of attack that undoes generalization and suppression (see Example 3.3.5).[84]

While the common interpretation of inferential disclosure violates the Process Protection Principle, the technical definition of Duncan & Lambert (1989) itself does not. The same is true for some of the other principles -- they are violated in common interpretation but not by Duncan & Lambert. In application, disclosure control techniques to protect against inference have often relied on keeping implementation parameters secret, violating Principle 6: Transparency (see eg, the secrecy of Census swap rates in Example 2.4.2 above).

[80] Federal Committee on Statistical Methodology, 'Statistical Policy Working Paper 2' (1978) uses the terms 'disclosure' and 'D-disclosure' to refer to this concept, employing language that parallels Opinion 05/2014: 'If the release of the statistics S makes it possible to determine the value Dx more accurately than i., possible without access to S, a disclosure has taken place. More exactly, a D-disclosure has taken place'. Ibid. 10. They attribute this concept to Tore Dalenius, 'Towards a methodology for statistical disclosure control' (1977) 15 Statistik Tidskrift 15, 222–429. The concept was later refined, most notably by Duncan and Lambert (1989) providing a notational framework, and by Chris J. Skinner, 'On identification disclosure and prediction disclosure for microdata' (1992) 46(1) Statistica Neerlandica 21-32, who adopts an absolute threshold for prediction accuracy and refers to the modified concept as 'predictive disclosure'.
[81] George Duncan and Diane Lambert, 'The risk of disclosure for microdata' (1989) 7(2) J. of Bus. & Econ. Stat. 207-217.
[82] Opinion 05/2014 (n 2), 18.
[83] Opinion 05/2014 (n 2), 18.
[84] Aloni Cohen, "Attacks on Deidentification's Defenses", USENIX 2022. https://arxiv.org/abs/2202.13470

Principles 3: Composition Awareness and 4: Assumption Minimization are violated in practice. The Composition awareness Principle is violated by use of k-anonymity and its variants (as above) to protect against inference, as these and other protections discussed in Opinion 05/2014 do not control composition risks. It has not been established whether or not inferencial disclosure as defined by Duncan & Lambert (1989) is theoretically compatible with composition awareness.

Finally, inferential disclosure fails the Assumption Minimization Principle even in theory, and in two ways. First, it can be underprotective, as one can learn harmful information about participants because of their inclusion in the sample, even if these attributes never appear as measured in data. Inferential disclosure can also be overprotective, as a computation can be labeled an inferential disclosure about an individual, even when that information would have been learned whether or not the individual was included in the data.

Model Language

Most of the aforementioned violations stem from practice, not concept, and the use of the model language in Section 2 should be sufficient to address these. Violations of principles of controlling composition and defining protection based on individuals' participation are deeper problems, and require a modification of the definition of disclosive inference itself.

We suggest the following model language for inference.

Model Language 8. *A (privacy-violating) inference occurs when an information release enables one to learn substantially about characteristics of an individual person or group as a result of the inclusion of that person or members of that group in the data collection process underpinning the information release.* To be considered protective, a disclosure limitation technique must guarantee that when outputs are produced by applying such technique to the data, the protected output produced by the technique cannot be used to significantly increase the likelihood of a privacy-violating inference.

Principle 5: Inclusion-based Protection is directly satisfied by Model Language 8 which can be used in combination with the model language developed for the principles presented in Section 2 above.

This model language can be integrated with the languages from the principles above:

Model Language 9 - self-contained model language for protection from privacy-violating inference.

A (privacy-violating) inference occurs when an information release enables one to learn substantially about characteristics of an individual person or group as a result of the inclusion of that person or members of that group in the data collection process underpinning the *information release.* To be considered protective, a disclosure limitation technique must guarantee that when outputs are produced by applying such technique to the data, the protected output produced by the technique cannot be used to significantly increase the likelihood of a privacy-violating inference.

Effective protection from privacy-violating inference is determined by the process or mechanism used to produce outputs. Protection from privacy-violating inference *may be considered effective only when the data processing itself is analyzed to guarantee that the informational relationships the processing creates between personal data and the outcome results in limited inference risks.*

Effective protection from privacy-violating inference applies generally regardless of data format. *Any informative data output carries inference risk and this risk accumulates with each analysis or publication. It is insufficient to limit protection from privacy-violating inference to individual-level records and datasets alone. Interactive mechanisms, aggregated data, statistical summaries, data derivatives and algorithmic uses of data equally carry inference risk and require protection. To be considered protective, the effectiveness of protection from privacy-violating inference measures should be demonstrated for all types of data releases, regardless of the release format and of whether information is exposed intentionally (eg, when publishing aggregate statistics) or unintentionally (eg, when the protocol inadvertently leaks information), and regardless of whether the release mechanism is static or interactive.*

Effective protection from privacy-violating inference requires controlling composition risks. *To be considered protective, a technique protecting from privacy-violating inferences must guarantee that, when outputs are produced by applying such technique to the data, the risk of releasing protected output contributes minimally to future inference risks of any included individual, even those future risks involving unknown future processing of, or data releases based on the same data.*

Effective protection from privacy-violating inference requires minimizing assumptions about downstream users and uses of information. *Because personal data can be learned from a data release in unanticipated ways, to be considered protective, the effectiveness of measures for ensuring protection from inference risks should depend minimally on any assumption regarding the potential users and uses of the data.*

Effective protection from privacy-violating inference requires defining the anonymization standards based on how participants' information affects the result of data processing. *When evaluating whether the result of data processing should be considered to relate to an individual, the result should be considered to relate to an individual only when it reveals information about that specific individual as a result of the inclusion of their information in the processing. Conversely, the result of data processing should not be considered as relating to an individual if the result of processing does not reveal information about the individual, or if information is revealed only as a byproduct of revealing information* *about an entire population. For example, publishing the average weight of a large population, or a general statistical relation that holds across the population (such as smoking increases the risk of cancer), does not reveal information that relates to an individual.*

Effective protection from privacy-violating inference requires ensuring the mechanisms in use provide sufficient protection whether or not their design details are exposed. *The processing of personal data that carries disclosure risk should use public protocols and algorithms which were scrutinized and vetted by the privacy community. At the very least, the data controller should have a justified belief that privacy is not undermined if the details of protocols and algorithms used to process the data are exposed. A best (but not sufficient) practice is to make specification documents, design documents, and well-documented code available for public inspection. Because the analysis of anonymized data by parties other than the data controller generally requires details of how those data were created, the publication of the design details of the anonymization mechanism serves the additional purpose of making the data more usable.*

Note that differential privacy definition is consistent with Model Language 9 insofar as it is applied to individuals or small groups.[85] Application of differential privacy—using the individual as the protected unit, using a small value of the privacy budget parameter epsilon, and controlling for composition—is sufficient to protect against inferential disclosure.

Examples 2.5.1 and 2.5.3 illustrate cases in which building statistical models do not violate the principle of inclusion, and thus do not constitute inference. The following examples provided additional clarification on how Model Language 8 should be interpreted.

Example 3.3.1. Inferences from recommendation systems as an example of a privacy-violating inference. Product recommendation systems used by websites such as Amazon can leak the private data on which they were trained, as referenced above in Section 1.1.[86] By mimicking the behavior of a target individual, such as purchasing items known to have been purchased in the past by that individual, and then monitoring temporal changes in the recommendation system's public outputs, the simulated attackers could infer additional information from the system about that specific individual's behavior.[87]

Example 3.3.2. Inferences from downcoding attacks as an example of a privacy-violating inference.

Deidentification approaches---including k-anonymity, l-diversity, and t-closeness---are widely used in the practice of data anonymization. These approaches classify certain data attributes as *quasi-identifiers* which may in principle be available to an attacker, and they aim to limit the

[85] Wood et al. (n 31).
[86] Calandrino et al. (n 11).
[87] Ibid.

inferences that can be made by an attacker that knows all of the quasi-identifiers of a target individual. Typically, this is done by generalizing or suppressing the quasi-identifiers until certain syntactic properties are satisfied (e.g., every combination of quasi-identifiers that appears in the dataset appears in at least k records).

Recently, Cohen showed that some algorithms for deidentification are vulnerable to an attack called *downcoding*, which can partially undo generalization and suppression. The result is that an attacker who knows details of the deidentification algorithm (cf. Principle X -- Transparency) and the overall population---but not the individuals---can learn about the individual data subjects' quasi-identifiers. Deidentification algorithms usually aim to generalize and suppress as few quasi-identifiers as possible so that the resulting data is as rich and useful as possible. Cohen's downcoding attacks leverage this fact to draw inferences about individual data subjects.

4. Conclusion: The Need to Systematize Data Protection Concepts

The rise of big data and artificial intelligence has led to the creation of new massive sources of data, complex analytical techniques, and sophisticated privacy attacks. These rapid developments put pressure on the understanding of longstanding data protection concepts and the effectiveness of traditional anonymization techniques. In light of these developments, we argue that a conceptual shift is needed for privacy and data protection, and such a shift should be informed by principles that have emerged from the scientific study of privacy.

We present in this paper a (perhaps non exhaustive) collection of principles, justify them, and apply them in the context of a particular regulatory framework - the EU's GDPR. We argue that these principles can be used to evolve the systematization of guidance on anonymization techniques and the definitions of criteria for effective anonymization, and we provide model language based on these principles towards more robust anonymization guidance and practices.

The data landscape is rapidly changing. This change is likely to accelerate, and poses a huge risk to regulations -- which risk being outdated the day they are made public. The principles above constitute a call for approaches that can make regulation more "future proof" in a way that is designed to address the challenges of a data ecosystem that is likely to evolve surprising new technologies and data uses. For example, the development of differential privacy demonstrates an approach to creating privacy protection technology that is indeed "future proof" -- its provable guarantees would continue to hold no matter how the data ecosystem evolves.

Such protections are made possible by building on the foundations of well-established formal principles and mathematical rigor. In this paper we demonstrate how this principled rigorous approach can be extended beyond a specific methodology such as differential privacy to development of privacy regulation as a whole.