Santa Clara University School of Law

Santa Clara Law Review

Volume 56 | Number 3 Article 3

6-17-2016

Shades of Gray: Seeing the Full Spectrum of Practical Data De-identification

Jules Polonetsky

Omer Tene

Kelsey Finch

SHADES OF GRAY: SEEING THE FULL SPECTRUM OF PRACTICAL DATA DE-IDENTIFICATION

Jules Polonetsky, Omer Tene and Kelsey Finch*

TABLE OF CONTENTS

Introduction 594
I. The Current Landscape 596
A. Nomenclature 596
B. Sustainability 599
C. Law 601
1. Federal Trade Commission 601
2. HIPAA 602
3. European Data Protection Directive 603
II. A Spectrum of Personal Data 604
A. The Variables 605
1. Direct Identifiers 605
2. Indirect Identifiers 605
3. Controls and Safeguards on the Use of Data 606
B. Common Categories of Data 607
III. Key Inflection Points on the Data Spectrum 609
A. Explicitly Personal Data 609
B. Potentially Identifiable and Not Readily Identifiabile Data 609
C. Key-Coded 613
D. Pseudonymous and Protected Pseudonymous Data 615
E. De-Identified and Protected De-Identified Data 617
F. Anonymous and Aggregated Anonymous Data 618

* Jules Polonetsky is Executive Director, Omer Tene is Senior Fellow and Kelsey Finch is Policy Counsel at the Future of Privacy Forum. Tene is Vice President of Research and Education at the International Association of Privacy Professionals and Associate Professor at the College of Management School of Law, Rishon Lezion, Israel. The authors would like to thank Peter Lefkowitz, Kim Gray, Khaled El Emam, Simson Garfinkel, Cameron Kerry, Javier Salido and Scott Goss for helpful comments.

IV. Additional Considerations 620
A. Sensitivity 620
B. Safeguards and Controls 620
Conclusion 621
Appendix A 623
A. Educational Programs 623
B. Geolocation and Traffic Services 625
C. Payment Processing 625
D. Medical Devices 626
E. Genetic Research 627
F. Mobile Devices 628

INTRODUCTION

For more than a decade, scholars and policymakers have debated the central notion of identifiability in privacy law.[1] De-identification, the process of removing personally identifiable information from data collected that is stored and used by organizations, was once viewed as a silver bullet allowing organizations to reap data benefits while at the same time avoiding risks and legal requirements. However, the concept of de-identification has come under intense pressure to the point of being discredited by some critics.[2] Computer scientists and mathematicians have come up with a re-identification tit for every de-identification tat.[3] At the same time, organizations around the world necessarily continue to rely on a wide range of technical, administrative and legal measures to reduce the identifiability of personal data to enable critical uses and valuable research while providing protection to individuals' identity and privacy.[4]

The debate around the contours of the term personally

1. For literature review *see* Ira S. Rubinstein & Woodrow Hartzog, *Anonymization and Risk*, 91 WASH. L. REV. ___ (forthcoming 2016).

2. Paul Ohm, *Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization*, 57 UCLA L. REV. 1701 (2010); *but cf.* Jane Yakowitz, *Tragedy of the Data Commons*, 25 HARV. J.L. & TECH.1 (2011); *also see* Felix T. Wu, *Defining Privacy and Utility in Data Sets*, 84 U. COLO. L. REV. 1117 (2013).

3. Arvind Narayanan & Vitaly Shmatikov, *Robust De-anonymization of Large Sparse Datasets*, 2008 PROC. 29TH IEEE SYMP. ON SECURITY & PRIVACY 111; Latanya Sweeney, *Achieving k-Anonymity Privacy Protection Using Generalization and Suppression*, 10 INT'L J. ON UNCERTAINTY, FUZZINESS & KNOWLEDGE-BASED SYSTEMS 571 (2002).

4. Ann Cavoukian & Khaled El Emam, *Dispelling the Myths Surrounding De-identification: Anonymization Remains a Strong Tool for Protecting Privacy*, June 2011, available at https://www.ipc.on.ca/images/Resources/anonymization.pdf.

identifiable information, which triggers a set of legal and regulatory protections, continues to rage, with scientists and regulators frequently referring to certain categories of information as "personal" even as businesses and trade groups define them as "de-identified" or "non-personal." The stakes in the debate are high. While not foolproof, de-identification techniques unlock value by enabling important public and private research, allowing for the maintenance and use—and, in certain cases, sharing and publication—of valuable information, while mitigating privacy risk. Appendix A to the Article provides examples of such use cases, which would be disrupted if policymakers pursued a less practical approach to de-identification.[5]

This Article proposes parameters for calibrating legal rules to data depending on multiple gradations of identifiability, while also assessing other factors such as an organization's safeguards and controls, as well as the data's sensitivity, accessibility and permanence. It builds on emerging scholarship that suggests that rather than treat data as a black or white dichotomy, policymakers should view data in various shades of gray; and provides guidance on where to place important legal and technical boundaries between categories of identifiability.

This Article recognizes that if data protection law defines personally identifiable information broadly, capturing any "singling out" of individuals and including data that any present or future third party could conceivably use to identify an individual, the law must be relaxed and allow for different types of consent; or for use restrictions tethered to the actual state of the data. Alternatively, if data protection law defines personally identifiable information more narrowly, there will be a need to establish or encourage rules for the collection and use of data sets that are not explicitly personal, yet do allow for decisions to be made that can affect individuals.

The Article urges the development of policy that creates incentives for organizations to avoid explicit identification and deploy elaborate safeguards and controls, while at the same time maintaining the utility of data sets.

5. *See* Appendix A.

I. THE CURRENT LANDSCAPE

A. Nomenclature

Despite a broad consensus around the need for and value of de-identification, the debate as to whether and when data can be said to be truly de-identified has appeared interminable. Although academics, regulators, and other stakeholders have sought for years to establish common standards for de-identification, they have so far failed to adopt even a common terminology.

As the National Institute of Standards and Technology (NIST) observed:

> Some authors and publications use the terms 'de-identification' and 'anonymization' interchangeably. Others use 'de-identification' to describe a process and 'anonymization' to denote a specific kind of de-identification that cannot be reversed. In some healthcare contexts the terms 'de-identification' and 'pseudonymization' are treated equivalently, with the term 'anonymization' being used to indicate that the mapping pseudonyms to subject identities has been erased. . .[6]

The definitional ambiguity is reflected in market behavior, industry guidance and even legislative texts and regulatory interpretations. In privacy policies, companies often refer to data as de-identified or non-personal if it does not contain explicit details, such as name or street address, or persistent identifiers such as a social security or credit card number. The privacy policy for the *New York Times* website, for example, defines as "non-personal" various categories of such data, including devices IDs, cookies, log files and reading history, and even location information.[7] Similarly, Volkswagen's German website defines personal data (*personenbezogene daten*) as "information that is directly related to you, including, for example, your name, your address, your telephone number and your e-mail address," noting that "information that is not directly related to you is not personal data."

According to the Digital Advertising Alliance (DAA), data

---

6. Simson L. Garfinkel, NISTIR 8053, De-Identification of Personal Information 2 (October 2015), http://nvlpubs.nist.gov/nistpubs/ir/2015/NIST.IR.8053.pdf.

7. Privacy Policy, NEW YORK TIMES June 10, 2015, http://www.nytimes.com/content/help/rights/privacy/policy/privacy-policy.html.

are de-identified "when an entity has taken reasonable steps to ensure that the data cannot reasonably be re-associated or connected to an individual or connected to or be associated with a particular computer or device."[8] Meanwhile, the Network Advertising Initiative (NAI), an industry self-regulatory framework, distinguishes between personally identifiable information (PII), defined as "data that is used, or intended to be used, to identify a particular *individual*," non-PII, defined as "data that is not linked, or reasonably linkable, to an individual, but is linked or reasonably linkable to a particular *computer or device*," and de-identified data, defined as "data that is not linkable to either an individual or a device."[9]

Regulators too differ in their perception of de-identification. According to the Federal Trade Commission (FTC), data are not "reasonably linkable" to individual identity to the extent that a company: (1) takes reasonable measures to ensure that the data are de-identified; (2) publicly commits not to try to re-identify the data; and (3) contractually prohibits downstream recipients from trying to re-identify the data (the "Three-Part Test").[10] Under the Health Insurance Portability and Accountability Act (HIPAA), de-identification is a term of art recognized under either an expert determination using generally accepted statistical and scientific principles and methods for rendering information not individually identifiable, or a safe harbor method based on removing from the data eighteen enumerated fields.

In Europe, regulators avoid the term de-identification altogether, and instead employ a strict version of anonymization that leaves little room for nuance. In its opinion on the term "personal data," the Article 29 Working Party interpreted the term "anonymized data" as "anonymous data that previously referred to an identifiable person, but where that identification is no longer possible."[11] The concept of pseudonymity, which is

8. DIGITAL ADVERTISING ALLIANCE, SELF REGULATORY PRINCIPLES FOR MULTI-SITE DATA, Nov. 2011, http://www.aboutads.info/resource/download/Multi-Site-Data-Principles.pdf.

9. NETWORK ADVERTISING INITIATIVE, 2015 UPDATE TO THE NAI CODE OF CONDUCT 3, available at https://www.networkadvertising.org/sites/default/files/NAI_Code15encr.pdf.

10. FEDERAL TRADE COMMISSION, PROTECTING CONSUMER PRIVACY IN AN ERA OF RAPID CHANGE 21 (2012), available at https://www.ftc.gov/sites/default/files/documents/reports/federal-trade-commission-report-protecting-consumer-privacy-era-rapid-change-recommendations/120326privacyreport.pdf.

11. WORKING PARTY, ARTICLE 29 WORKING PARTY OPINION 5/2014 ON

defined in an ISO technical specification as a "particular type of anonymization that both removes the association with a data subject and adds an association between a particular set of characteristics relating to the data subject and one or more pseudonyms,"[12] is recognized in German law.[13] It has reemerged under the new text of the General Data Protection Regulation (GDPR), albeit with ambiguous legal consequences.[14]

The force driving these discussions into definitional deadends is the profound legal implication of defining data as personally identifiable or not. As one scholar put it, anonymization is ubiquitous, trusted and rewarded by law.[15] In the current legal frameworks of both the U.S. and the EU, data identifiability—rolled into the term "personal data" in EU law or "PII" in the U.S.—operates as a forceful legal trigger. Once data are viewed as personally identifiable, they become subject to the full panoply of legal obligations and restrictions.[16] Accordingly, organizations around the world have structured their internal and external privacy policies and practices around variations of "PII"—and its converse, de-identified data—locking themselves into a binary that does not accurately reflect how data are treated in practice.

ANONYMIZATION TECHNIQUES 8 (April 10, 2014), http://ec.europa.eu/justice/data-protection/article-29/documentation/opinion-recommendation/files/2014/wp216_en.pdf.

12. International Organization for Standardization, *ISO/TS 25237:2008 Health informatics – Pseudonymization*, http://www.iso.org/iso/catalogue_detail?csnumber=42807.

13. Section 3(6a) of the German Data Protection Act, *Bundesdatenschutzgesetz*, defines "Aliasing" as "replacing a person's name and other identifying characteristics with a label, in order to preclude identification of the data subject or to render such identification substantially difficult."

14. REGULATION (EU) NO 2016/679 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL OF APRIL 27, 2016, on the protection of individuals 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) (Text with EEA relevance), http://ec.europa.eu/justice/data-protection/reform/files/regulation_oj_en.pdf.

15. Ohm, *supra* note 2.

16. Although, note that the "full panoply" in the US is markedly less than in EU. Paul Schwartz & Dan Solove, *The PII Problem: Privacy and a New Concept of Personally Identifiable Information*, 86 NYU L. REV. 1814 (2011); Elosie Gratton, *If Personal Information Is Privacy's Gatekeeper, then Risk of Harm is the Key: A Proposed Method for Determining What Counts as Personal Information*, 24 ALBANY L.J. SCI. & TECH. (2013).

To help ease the bind created by this entrenched nomenclature, leading scholars have called for recognition of a broader spectrum. Solove and Schwartz, for example, proposed a "PII 2.0" continuum, which effectively converts the existing dichotomy into a trichotomy, with data categorized as "identified, identifiable, or non-identifiable."[17] While allowing for more flexibility than current laws, this proposal warrants expansion to account for multiple shades of de-identified data and the legal issues arising in each case.

To steer clear of this terminological fray, this article applies loose headings to various data categories in order to distinguish them from one another rather than to stake a claim as to whether they fit within any given regulatory framework.

B. Sustainability

Notwithstanding disagreements around nomenclature, some critics contend that true de-identification is not possible, or at least is not sustainable. Rather than focus on *how* to de-identify personal information, the discussion has increasingly shifted to *whether* personal information can be (or can be said to be) "de-identified" and thus not personally identifiable.

To prove this point, commentators rely on several well-publicized attacks against purportedly de-identified databases, which led to the successful *re*-identification of certain individuals. Some of the most (in)famous examples of re-identification arose from the public release of AOL search data,[18] a Massachusetts medical database,[19] Netflix recommendations,[20] and an open genomics database.[21] In each of these cases, "even though administrators had removed any data

---

17. Paul Schwartz & Dan Solove, *The PII Problem: Privacy and a New Concept of Personally Identifiable Information*, 86 NYU L. REV. 1814 (2011); Elosie Gratton, *If Personal Information Is Privacy's Gatekeeper, then Risk of Harm is the Key: A Proposed Method for Determining What Counts as Personal Information*, 24 ALBANY L.J. SCI. & TECH. (2013).

18. Michael Barbaro & Tom Zeller, *A Face Is Exposed for AOL Searcher No. 4417749*, NY.
TIMES, Aug. 9, 2006, http://www.nytimes.com/2006/08/09/technology/09aol.html.

19. Latanya Sweeney, *Uniqueness of Simple Demographics in the U.S. Population*,
(Laboratory for International Data Privacy, Working Paper No. 4, 2000).

20. Narayanan & Shmatikov, *supra* note 3.

21. M Gymrek, A.L. McGuire et al, *Identifying personal genomes by surname inference*, 339 SCIENCE 312 (2013).

fields they thought might uniquely identify individuals, researchers . . . unlocked identity by discovering pockets of surprising uniqueness remaining in the data."[22] At the same time, supporters of de-identification pointed out that only a handful of thousands or millions of records were re-identified.[23]

More generally, scientists have demonstrated that leakage of apparently benign information, such as the publication of photos of celebrities boarding NYC cabs,[24] can lead to an unexpected unraveling of de-identification efforts. Repeatedly, researchers have shown that in a big data world, even mundane data points, such as the battery life remaining on an individual's phone, can serve as potent identifiers singling out an individual from the crowd.[25]

Given the sophistication of the data handlers in these cases and the repeated success of re-identification attacks, critics concluded that "de-identification fails to resist inference of sensitive information either in theory or in practice," adding that "attempts to quantify its efficacy are unscientific and promote a false sense of security."[26] In its report, the President's Council of Advisors on Science and Technology (PCAST) concluded that "Anonymization remains somewhat useful as an added safeguard, but it is not robust against near-term future re-identification methods. PCAST does not see it as being a useful basis for policy."[27]

At the same time, defenders of de-identification argued that these attacks have all been on databases that were not

---

22. Ohm, *supra* note 2 at 1723.

23. Kathleen Benitez & Bradley K. Malin, *Evaluating Re-Identification Risks With Respect to the HIPAA Privacy Rule*, 17 J. AMER. MED INFORMATICS ASSOC. 169 (2010); Deborah Lafkey, *The Safe Harbor Method of Deidentification: An Empirical Test* 19, 2009, www.ehcca.com/presentations/HIPAAWest4/lafky_2.pdf; Khaled El Emam et al, *A Systematic Review of Re-Identification Attacks on Health Data*, 6 *PLoS One* 1, December 2011.

24. Anthony Tockar, *Riding with the Stars: Passenger Privacy in the NYC Taxicab Data Set*, September 15, 2014, http://research.neustar.biz/author/atockar.

25. Lukasz Olejnik, Gunes Acar, Claude Castelluccia & Claudia Diaz, *The Leaking Battery: A Privacy Analysis of the HTML5 Battery Status API*, 2015, http://eprint.iacr.org/2015/616.pdf.

26. Arvind Narayanan & Ed Felten, *No Silver Bullet: De-Identification Still Doesn't Work*, July 9, 2014, http://randomwalker.info/publications/no-silver-bullet-de-identification.pdf.

27. PRESIDENT'S COUNCIL OF ADVISORS ON SCIENCE AND TECHNOLOGY (PCAST), BIG DATA AND PRIVACY: A TECHNOLOGICAL PERSPECTIVE, May 2014, https://www.whitehouse.gov/sites/default/files/microsites/ostp/PCAST/pcast_big_data_and_privacy_-_may_2014.pdf.

credibly de-identified to begin with[28] and have "distort[ed] the de-identification policy debate because they are not representative or have been misrepresented in popular media."[29] Rather, they claimed, in these cases, the datasets were either mislabeled *pseudonymous* or at least inadequately de-identified, and should therefore not be used to undermine respect for more rigorous de-identification measures. Underlying this rebuttal, however, is an even more deep-seated disagreement over what it actually means for there to be re-identification risk and what the implications are of a record being re-identified.[30] Until these goal posts are firmly set, it seems likely that re-identification discussions will continue to spin in circles. In the meantime, businesses continue to employ de-identification techniques to reduce privacy risks, and policymakers must decide how to treat such relatively—if not entirely foolproof—de-identified information.

C. Law

While policymakers have frequently tried to draw bright lines around personally identifiable data, there is little consistency in what is or is not considered legally de-identified under the law.

1. Federal Trade Commission

In the U.S., the FTC has acknowledged the broad consensus that "the traditional distinction between PII and non-PII has blurred and that it is appropriate to more comprehensively examine data to determine the data's privacy implications."[31] The FTC's current de-identification standard hinges on whether there is "a reasonable level of justified confidence that the data cannot reasonably be used to infer information about, or otherwise be linked to, a particular consumer, computer, or

28. Daniel Barth-Jones, *The Antidote for "Anecdata": A Little Science Can Separate Data Privacy Facts from Folklore*, Nov. 21, 2014, https://blogs.law.harvard.edu/infolaw/2014/11/21/the-antidote-for-anecdata-a-little-science-can-separate-data-privacy-facts-from-folklore/.

29. Rubinstein & Hartzog, *supra* note 1.

30. NISTIR, *supra* note 6. Including whether re-identification must actually occur, or whether a probability of re-identification is sufficient; whether any or all records must be identifiable; and what level of confidence is needed to declare a data set either de- or re-identified.

31. FTC, *supra* note 10, at 2.

other device."[32] To determine when data are not "reasonably linkable," the FTC has established the Three-Part Test.[33] The FTC further advises that "the nature of the data at issue and the purposes for which it will be used are also relevant. Thus, for example, whether a company publishes data externally affects whether the steps it has taken to de-identify data are considered reasonable."[34] Confusingly, this standard would treat data linked to a particular device as personal, assuming perhaps that all devices are unique to individuals. Certainly computers, cellphones and tablets will qualify, but in today's "internet of things" environment many unique devices have no association to any particular individual.

2. HIPAA

The most elaborate regulatory scheme for de-identification is set forth by HIPAA, which provides that organizations may deem health data "de-identified" by removing eighteen categories of identifiers from a data file, after which data can be released publicly.[35] Such data can include a special purpose code of identification allowing the organization that created the data to re-identify individuals, as long as the identifier is not related to information about the individual and cannot be used by others to identify the individual.[36] If the data are shared under contractual protections for limited research, public health, or health care operations, the data may include specific dates and other indirect identifiers.[37] But in neither case can an IP address be included.[38]

---

32. FTC, *supra* note 10, at 21.

33. FTC, *supra* note 10, at 21.

34. FTC, *supra* note 10, at 21. The FTC also requires that a company must publicly commit to maintain and use the data in a de-identified fashion, and not to attempt to re-identify the data, and to impose similar restrictions by contract on any third party data recipient.

35. *See* 45 CFR § 164.514(b)(2); *see also* Office of Civil Rights, Health & Human Services, Guidance Regarding Methods for De-Identification of Protected Health Information in Accordance with the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule 7-8 (November 26, 2012), available at http://www.hhs.gov/sites/default/files/ocr/privacy/hipaa/understanding/coveredentities/De-identification/hhs_deid_guidance.pdf

36. *Id.* at 21-22.

37. OFFICE OF CIVIL RIGHTS, HEALTH & HUMAN SERVICES, OCR HIPAA PRIVACY: RESEARCH (June 5, 2013), http://www.hhs.gov/sites/default/files/ocr/privacy/hipaa/understanding/special/research/research.pdf

38. *See* 45 CFR § 164.514(b)(2)(i)(P); 45 CFR § 164.514(e)(2)(xiv).

3. European Data Protection Directive

In Europe, policymakers use the term anonymization, as opposed to de-identification. The European Data Protection Directive defines anonymization negatively, noting that its provisions "shall not apply to data rendered anonymous in such a way that the data subject is no longer identifiable," where "an identifiable person is one who can be identified, directly or indirectly, in particular by reference to an identification number or to one or more factors specific to his physical, physiological, mental, economic, cultural or social identity."[39] The Data Protection Directive does not mention the term pseudonymisation at all.

Among European Union Member States, anonymization and pseudonymization have been a "major area of divergent interpretation," as the European Commission's staff noted in an evaluation of the implementation of the Data Protection Directive.[40] According to that report, several EU Member States (e.g., Austria, Germany, Ireland, Netherlands and UK), "consider encoded or pseudonymised data as identifiable—and thus as personal data—in relation to the actors who have means (the 'key') for re-identifying the data, but not in relation to other persons or entities."[41] In contrast, other Member States (e.g., Denmark, France, Italy, Spain, Sweden), regard any data that can possibly be linked to an individual by any third party as "personal," even in the hands of someone who has no reasonable means for such re-identification.[42] However, regulators "in those Member States apply less demanding obligations with regard to the processing of data that is not immediately identifiable, taking into account the likelihood of re-identification."[43]

According to guidance it issued, the Article 29 Working Party assesses anonymization primarily according to technical

39. Directive 95/46/EC of the European Parliament and of the Council of 24 October 1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such data, OFFICIAL JOURNAL L 281 , 23/11/1995 P. 0031 – 0050, *available at* http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=CELEX:31995L0046:en:HTML ("Data Protection Directive").

40. European Council, Annex 2, Evaluation of the Implementation of the Data Protection Directive 244 (2012), *available at* http://lobbyplag.eu/governments/assets/pdf_all/CD-all.pdf

41. *Id.* at 15.

42. *Id.*

43. *Id.*

measures.[44] While cogently presenting the technical issues and privacy risks inherent in de-identification, the Article 29 Working Party's understanding of acceptable re-identification risk has been understood by some as requiring near-zero probability, an impractical standard.[45]

As demonstrated below, the various drafts of the GDPR reflect an ambivalence towards pseudonymization and anonymization.[46] The final draft defines anonymous information as "information which does not relate to an identified or identifiable natural person" and "to data rendered anonymous in such a way that the data subject is not or no longer identifiable."[47] Importantly—and similar to the Article 29 Working Party position—it calibrates the concept of identifiability to an organization's ability to "single out" an individual based on a piece of information.

II. A SPECTRUM OF PERSONAL DATA

While the rhetorical and policy debates surrounding de-identification fulminate, organizations continue to employ a range of techniques and controls to de-identify, obscure, and protect their data. These methods offer widely varying levels of protection and obscurity, depending on the context of their use. Too often, they have become square pegs forced into the round, all-or-nothing holes of the current PII framework. In order to help advance the discussions around practical de-identification, this paper examines the range of practices and proposes a reclassification of data on a spectrum according to differing categories of identifiability. A more nuanced understanding of how organizations are protecting their data on the ground will help the de-identification community better assess and respond to both data opportunities and privacy risks.

---

44. Article 29, supra note 11.

45. *See, e.g.*, Khaled El Emam & Cecilia Alvarez, *A Critical Appraisal of the Article 29 Working Party Opinion 05/2014 on Data Anonymization Techniques*, Int'l Data Privacy Law (Dec. 13, 2014), http://idpl.oxfordjournals.org/content/early/2014/12/12/idpl.ipu033.full.pdf?keytype=ref&ijkey=K8xdZaj1rw3EzDx.

46. *See infra* notes 79-86 and accompanying text.

47. GDPR, Rec. 23.

A. The Variables

In order to redraw the de-identification spectrum, it is important to first understand the difference between direct and indirect identifiers, as well as how de-identified data sets are commonly shared or made public. Different combinations of these parameters account for the different categories of data described below.

1. Direct Identifiers

Data about an individual can identify that individual either directly or indirectly. In de-identification literature, *direct identifiers* are "data that can be used to identify a person without additional information or with cross-linking through other information that is in the public domain."[48] These include names, social security numbers, or basic contact information. Clearly, in order to render data non-personal, these obvious identity fragments must be removed or altered. Some common methods of addressing direct identifiers include their suppression or replacement with symbols, generic names, or random values.[49] If a direct identifier is consistently replaced with another specific value it becomes a pseudonym, which allows linking information belonging to an individual across multiple data records or information systems, provided that similar direct identifiers are systematically pseudonymized.[50]

2. Indirect Identifiers

Data that identifies an individual *indirectly* helps connect pieces of information until a particular individual can be singled out. Some of the most common indirect identifiers (also known as quasi-identifiers) include date of birth, age, gender, ZIP code, and other basic demographic information. No single individual can be identified based on any one of these data points. Yet, as additional indirect identifiers compound, an identity can emerge. As leading researchers have pointed out, "Whereas direct identifiers can be removed from the dataset, quasi-identifiers generally convey some sort of information that might be important for a later analysis and removing them may damage the utility of the dataset."[51] Common ways

48. ISO/TS 25237:2008(E), *supra* note 12, at 3.
49. NISTIR, *supra* note 6.
50. NISTIR, *supra* note 6.
51. NISTIR, *supra* note 6 at 20.

to de-identify indirect identifiers include: suppressing or removal; generalizing values as sets or ranges; swapping data between individual records; and perturbing or adding noise.[52]

3. Controls and Safeguards on the Use of Data

In addition to the nature of identifiers, a framework for practical de-identification also takes into account the *safeguards and controls* placed on the way data are obtained, used or disseminated.[53] There are a several typical models for releasing de-identified data, including: "Release and Forget model," where data are published publicly or made available on the internet; "Data Use Agreements model," where data are provided under legally binding contracts detailing how data may and may not be used (typically either in a negotiated agreement with a "qualified investigator" or via "click-through" license agreements); and the "Enclave model," where data are "kept in some kind of segregated enclave that accepts queries from qualified researchers, runs the queries on the de-identified data, and responds with results."[54]

Non-technical safeguards and controls include two broad categories: 1) internal administrative and physical controls (internal controls)[55]; and 2) external contractual and legal protections (external controls).[56] *Internal controls* encompass security policies, access limits, employee training, data segregation guidelines, and data deletion practices that aim to stop confidential information from being exploited or leaked to the public. *External controls* involve contractual terms that restrict how partners use and share information, and the corresponding remedies and auditing rights to ensure compliance.[57] By implementing administrative safeguards, organizations provide important privacy protections independent of technical de-identification. Policymakers in the U.S. and Europe have recognized the value of such safeguards and controls, setting forth de-identification standards that factor in various types of safeguards to meet legal tests such as reasonableness (FTC) or

52. NISTIR, *supra* note 6 at 20.
53. NISTIR, *supra* note 6.
54. NISTIR, *supra* note 6 at 14.
55. *See* Privacy Act of 1974, 5 U.S.C. § 552a(e)(10) (2011).
56. *See discussion in* Yianni Lagos & Jules Polonetsky, *Public vs. Nonpublic Data: The Benefits of Administrative Control*, 66 STAN. L. REV. ONLINE 103 (2013).
57. *Id.* at 106.

"likely reasonably to be used" (EU).

B. Common Categories of Data

In the table on the following page, "*A Visual Guide to Practical Data De-Identification*," we delineate common categories of data as proposed sign posts on the de-identification spectrum, based on the parameters identified above—the existence of direct or indirect identifiers as well as the safeguards and controls over dissemination.[58]

58. For a similar framing *see* Khaled El Emam, Eloïse Gratton & Jules Polonetsky, *The Seven States of Data* (unpublished manuscript, on file with authors).

A VISUAL GUIDE TO PRACTICAL DATA DE-IDENTIFICATION

Produced by FUTURE OF PRIVACY FORUM FPF.ORG

What do scientists, regulators and lawyers mean when they talk about de-identification? How does anonymous data differ from pseudonymous or de-identified information? Data identifiability is not binary. Data lies on a spectrum with multiple shades of identifiability.

This is a primer on how to distinguish different categories of data.

DEGREES OF IDENTIFIABILITY

Information containing direct and indirect identifiers.

PSEUDONYMOUS DATA

Information from which direct identifiers have been eliminated or transformed, but indirect identifiers remain intact.

DE-IDENTIFIED DATA

Direct and known indirect identifiers have been removed or manipulated to break the linkage to real world identities.

ANONYMOUS DATA

Direct and indirect identifiers have been removed or manipulated together with mathematical and technical guarantees to prevent re-identification.

	EXPLICITLY PERSONAL	POTENTIALLY IDENTIFIABLE	NOT READILY IDENTIFIABLE	KEY CODED	PSEUDONYMOUS	PROTECTED PSEUDONYMOUS	DE-IDENTIFIED	PROTECTED DE-IDENTIFIED	ANONYMOUS	AGGREGATED ANONYMOUS
DIRECT IDENTIFIERS Data that identifies a person without additional information or by linking to information in the public domain (e.g., name, SSN)	INTACT	PARTIALLY MASKED	PARTIALLY MASKED	ELIMINATED or TRANSFORMED	ELIMINATED or TRANSFORMED	ELIMINATED or TRANSFORMED	ELIMINATED or TRANSFORMED	ELIMINATED or TRANSFORMED	ELIMINATED or TRANSFORMED	ELIMINATED or TRANSFORMED
INDIRECT IDENTIFIERS Data that identifies an individual indirectly. Helps connect pieces of information until an individual can be singled out (e.g., DOB, gender)	INTACT	INTACT	INTACT	INTACT	INTACT	INTACT	ELIMINATED or TRANSFORMED	ELIMINATED or TRANSFORMED	ELIMINATED or TRANSFORMED	ELIMINATED or TRANSFORMED
SAFEGUARDS and CONTROLS Technical, organizational and legal controls preventing employees, researchers or other third parties from re-identifying individuals	NOT RELEVANT due to nature of data	LIMITED or NONE IN PLACE	CONTROLS IN PLACE	CONTROLS IN PLACE	LIMITED or NONE IN PLACE	CONTROLS IN PLACE	LIMITED or NONE IN PLACE	CONTROLS IN PLACE	NOT RELEVANT due to nature of data	NOT RELEVANT due to high degree of data aggregation
SELECTED EXAMPLES	Name, address, phone number, SSN, government-issued ID (e.g., Jane Smith, 123 Main Street, 555-555-5555)	Unique device ID, license plate, medical record number, cookie, IP address (e.g., MAC address 68:A8:6D:35:65:03)	Same as Potentially Identifiable except data are also protected by safeguards and controls (e.g., hashed MAC addresses & legal representations)	Clinical or research datasets where only curator retains key (e.g., Jane Smith, diabetes, HgB 15.1 g/dl = Csrk123)	Unique, artificial pseudonyms replace direct identifiers (e.g., HIPAA Limited Datasets, John Doe = SL7T LX619Z) (unique sequence not used anywhere else)	Same as Pseudonymous, except data are also protected by safeguards and controls	Data are suppressed, generalized, perturbed, swapped, etc. (e.g., GPA: 3.2 > 3.0-3.5, gender: female > gender: male)	Same as De-Identified, except data are also protected by safeguards and controls	For example, noise is calibrated to a data set to hide whether an individual is present or not (differential privacy)	Very highly aggregated data (e.g., statistical data, census data, or population data that 52.6% of Washington, DC residents are women)

III. KEY INFLECTION POINTS ON THE DATA SPECTRUM

A. Explicitly Personal Data

At one end of the spectrum, a dataset contains direct identifiers thus making the data *Explicitly Personal*. Data fall into this category when no attempt has been made to obscure either direct or indirect identifiers. This information is typically considered PII under current legal regimes whether or not safeguards around sharing or use controls are present.[59] Even at this extreme, however, not all personal data are created equal.[60] For example, although an email address may be just as quick to identify an individual as a social security number, the latter is still considered more sensitive given its persistence, resistance to change, and common use as a key to additional sets of PII. Similarly, not all names are created equal. Some names are unique enough to identify an individual in a small crowd; others, like John Smith, are so prevalent that standing on their own they do not constitute personally identifiable information.

B. Potentially Identifiable and Not Readily Identifiabile Data

When an attempt has been made to obscure or avoid collecting explicit direct identifiers, data slides down the identifiability spectrum. This is the case, for example, when a company decides to use a customer's static device ID in lieu of their name or social security number, or to hash a customer's email address. Such partially masked data may still be identifiable since it could allow the company to single out specific individuals; but it is less explicit or risky than data stored with a user's name and real-world identity attached. The *Potentially Identifiable* category captures data that may not be *explicitly* personal, but that has been only partially masked as to be *potentially identifiable*. Like the Explicitly Personal category, Potentially Identifiable data leaves indirect identifiers intact and

59. See, e.g., GDPR Art. 4(1); Video Privacy Protection Act of 1988, 18 U.S.C. § 2710(a)(3) (2006).

60. For example, one website indicates that there are 46,576 people named John Smith in the U.S., but only four people named Kelsey Finch and only one Jules Polonetsky and one Omer Tene. *How Many of Me*, howmanyofme.com, last visited Apr. 11, 2016.

applies limited or no safeguards and controls on sharing, publication or use of information.[61]

Potentially Identifiable data are widely used in a variety of online and mobile ecosystems, and include cookies with unique IDs, device identifiers, MAC and IP addresses, advertising identifiers, and common hashes of such data. One of the most contentious discussions in the de-identification debate revolves around the role of such device identifiers, which are direct yet not explicitly tied to real-world identities

On the one hand, businesses have relied on device identifiers for a range of tracking and targeting activities, while referring to them as "anonymous," "non-personal," "de-identified" or similar terms. Advertising industry standards today consider such data as non-PII.[62]

On the other hand, many regulators and technologists consider such identifiers to be personally identifiable. For example, the Article 29 Working Party asserts, "If pseudonymization is based on the substitution of an identity by another unique code, the presumption that this constitutes a robust de-identification is naïf and does not take into account the complexity of identification methodologies and the multifarious contexts where they might be applied."[63] In similar vein, the text of the GDPR grounds the very notion of identifiability on a company's ability to "single out" an individual, regardless of whether such an identity is linked to a real-world identifier such as name or address.[64]

The FTC has clarified its view that device-linked information is personally identifiable.[65] Where it has rulemaking authority (under COPPA), the agency expressly defined such identifiers as PII.[66] In a non-COPPA context, the FTC's posi-

---

61. In the Not Readily Identifiable category, similar data are subject to more controls.

62. *See, e.g.*, privacy policies *supra* notes 7-8.

63. Article 29 WP at 31, *supra* note 11.

64. General Data Protection Regulation, rec. 23.

65. *See recently*, FTC Press Release, T*wo App Developers Settle FTC Charges They Violated Children's Online Privacy Protection Act*, Dec. 17, 2015, https://www.ftc.gov/news-events/press-releases/2015/12/two-app-developers-settle-ftc-charges-they-violated-childrens?utm_source=govdelivery.

66. FTC, *Complying with COPPA: Frequently Asked Questions, A Guide for Business and Parents and Small Entity Compliance Guide*, March 2015, https://www.ftc.gov/tips-advice/business-center/guidance/complying-coppa-frequently-asked-questions#General Questions.

tion on whether such data is personally identifiable will depend on satisfaction of the Three-Part Test, including the existence of technological and legal controls and user choices.[67]

Importantly, if policymakers completely disregarded such means of obfuscation, companies would lose any incentive to deploy them. Consequently, in practice, policymakers recognize the need to use and exchange such device identifiers for various purposes, as well as their being less explicit than direct identifiers such as name and address. Some regulators have informally sanctioned the sharing of such device identifiers for limited purposes or subject to use restrictions.

For example, in 2009, the German data protection authority in Hamburg passed a resolution that made the analysis of user behavior tied to a full IP address permissible only with a user's explicit consent.[68] In apparent violation of these new rules, most web analytics services, including market leader Google Analytics, which gather such information as a matter of course, did not have the practices in place to gather such consent. Rather than oust these services entirely, the Hamburg DPA entered into a binding resolution with Google in 2011 implementing certain—but not all—of data protection's legal obligations.[69] These measures included allowing users to opt out, enabling website operators to request that any IP addresses collected be "anonymized" by deleting the last digits, and requiring data processing agreements to be executed between Google and website operators using Google Analytics. Website operators were also required to inform users about the use of Google Analytics in their privacy policies, including notice of the opt out, and to delete data collected using previous, non-compliant analytics profiles.[70]

A complete discussion of the factors that turn an item of data into a direct identifier is beyond the scope of this paper. Clearly, the existence and prevalence of a look-up database render an identifier direct. For purposes of the discussion here, suffice it to note that not all identifiers are created equal. For

67. FTC, *supra* note 10 at 21.

68. *Hamburg Data Protection Authority: Data protection-conforming use of Google Analytics*, IITR.us (Oct. 29, 2011), http://www.iitr.us/publications/35-hamburg-data-protection-authority-data-protection-conforming-use-of-google-analytics.html.

69. *Id.*

70. *Id.*

example, some identifiers are reversible only by the organization holding the data, while others have public look-up databases. One identifier might be easily cleared by users, for example, by deleting their cookie file, while another could be hard-coded or only clearable by resetting a user's device. One identifier might be stored only locally, while another is shared globally. A persistent unique identifier, which can be used over time to compile an increasingly detailed profile of an individual or device, entails different privacy risk than a special identifier that is dynamically reassigned to multiple individuals and regularly rotated.

Thus emerges the next category, *Not Readily Identifiable* data, which recognizes that adding significant safeguards and controls to partially masked identifiers can make such data less readily identifiable. Critics will likely challenge the characterization of an identifier, such as a unique cookie ID or device ID, as "non personal," if it is shared widely, cannot be deleted by users, or is in fact commonly linked by companies to personal information. But if an identifier can be cleared by a user, its dissemination and retention controlled, and strong technical and legal constraints prevent it from being linked to personal information—it should slide an additional step down the identifiability spectrum and warrant a more flexible legal regime.[71]

Regulators could take advantage of these gradations of identifiability to impose more nuanced use restrictions, similar to self-regulatory frameworks in the U.S. The NAI Code of Conduct, for example, applies obligations for notice, choice, opt-out, and non-discrimination to datasets defined as "non-personal"—that is, neither anonymous nor obviously personally identifiable.[72] The DAA Self-Regulatory Principles also set forth protections for pseudonymous identifiers, determining that "data is not considered PII under the Principles if the data is not used in an identifiable manner."[73] Here, collection in isolation of an IP address, for example, is not considered pro-

71. Given adequate perturbation of indirect identifiers and the presence sufficient controls, including requirements to not re-identify data and to require any downstream recipients to make the same commitment, such data may be characterized under as De-Identified data (see below) and satisfy the FTC de-identification standard.

72. NAI, *supra* note 9.

73. DAA, *supra* note 8.

cessing of PII, and thus does not require consent or transparency even if used for online behavioral advertising, but is considered PII subject to the full set of Principles when it is "in fact linked to an individual in its collection and use."[74] Meanwhile, under the Mobile Location Analytics Code of Conduct, organizations are required to provide in-store notice, hash mobile device ID MAC addresses and set discrimination and retention limits around non-personal but not de-identified sets of "de-personalized" data.[75]

To be sure, at the end of the day the devil is in the detail. Much turns on where the borders are drawn between Explicitly Personal, Potentially Identifiable, and Not Readily Identifiable data, as well as on the safeguards and controls that apply to the various categories of data. But clearly, a more nuanced approach than categorizing data in all of these categories as personally identifiable would provide organizations with an incentive to enhance privacy protection by pushing data down the identifiablity spectrum.

C. Key-Coded

Key-coded data could reside in several categories. Key-coded data are personally identifiable information that have been stripped of direct identifiers, which have been replaced by a key to avoid unwanted or unintended re-identification. Because the data are so readily re-identifiable by the key holder, they must be considered personal data and fall under Explicitly Personal or Potentially Identifiable when they are *in that party's hands*. To any other party, however, key-coded data would fall within the Pseudonymous or De-Identified categories and be considered *non*-personally identifiable, depending on the treatment of any *indirect* identifiers in the dataset. In other words, a key-coded dataset could be viewed as a Pseudonymous dataset, with particularly strong controls allowing dissemination of the key to only a restricted subset of players (*e.g.*, researchers). The strength of the key should play a factor in the analysis, distinguishing between encryption-based key allocation that could conceivably be reversed by an adversary and randomly mapped keys that can survive an attack.

---

74. DAA, *supra* note 8.

75. Future of Privacy Forum, *Mobile Location Analytics Code of Conduct*, October 2013, http://www.futureofprivacy.org/wp-content/uploads/10.22.13-FINAL-MLA-Code.pdf.

Key-coded data are used extensively in a range of circumstances where limited re-identification is necessary or desirable, including pharmaceutical research, scientific and historical research, marketing analysis, and online and mobile services. For example, in clinical trials, health institutions typically must maintain an ability to link research data back to specific patients, in order to alert them of a treatable condition they discover or contain the spread of an infectious disease.[76] In online advertising, in contrast, intermediaries compare hashed data from different sources to identify matches without having to reveal an individual user's identity to any of the transacting parties.[77]

It is important for policymakers to recognize the difference between key-coded data in the hands of the curator who also holds the key, and similar data in the hands of a researcher or other third party who cannot reasonably "unlock" it. Under the European Data Protection Directive, "to determine whether a person is identifiable, account should be taken of all the means likely reasonably to be used either by the controller *or by any other person*."[78] The same formula remains in Recital 26 of the GDPR. A strict reading of this language would impute re-identification to third parties who do not hold a key, based on the capabilities of the party who first coded the data.

In practice, access to key-coded research data is highly restricted, with administrative safeguards and legal controls as well as reputational barriers limiting access to just verified users. If regulators fail to offer credit to such controls and instead choose to treat all key-coded data as if it were readily identifiable, they would inevitably impair critical scientific research. Researchers would be forced to sacrifice useful data to meet more cumbersome de-identification standards, even though they would be no more or less capable of re-identifying the data than before.

---

76. Adrian Thorogood et al., *An implementation framework for the feedback of individual research results and incidental findings in research*, BMC Med Ethics. 2014; 15: 88, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4391472/.

77. ZE ZOOK & PR SMITH, MARKETING COMMUNICATIONS: OFFLINE AND ONLINE INTEGRATION, ENGAGEMENT AND ANALYTICS 363 (2016).

78. Data Protection Directive, Recital 26 (*emphasis* added).

D. Pseudonymous and Protected Pseudonymous Data

Data in one sector of this category are vanilla *Pseudonymous* data, meaning information in which direct identifiers have been eliminated or transformed, but indirect identifiers remain intact without safeguards or controls over their release. When organizations overlay such data with safeguards and controls, the data move further down the identifiability spectrum to *Protected Pseudonymous* data. Limited data sets under the HIPAA are an example of data in this category. They comprise Protected Health Information (PHI) that excludes direct identifiers and various categories of indirect identifiers, but explicitly includes other indirect identifiers that must be scrubbed under the HIPAA de-identification Safe Harbor standard, including dates, city, state, zip code, and age. They may be used or disclosed subject to strict use agreements, for purposes of research, public health, or health care operations. We distinguish Pseudonymous data from Potentially Identifiable or Not Readily Identifiable data in that Pseudonymous data does not contain any direct identifiers that can be used to link data across contexts. In the Pseudonymous category, data can be linked strictly to an *ad hoc* identifier that has no life outside of the specific context in which it is used.

"Pseudonymous" is, admittedly, a highly contentious term in the de-identification literature. Technologists regard pseudonymization as a process for removing direct identifiers and replacing them with pseudonyms, that is, a "particular type of anonymization."[79] In contrast, the Article 29 Working Party stated that "pseudonymisation is *not* a method of anonymisation," but rather merely reduces the linkability of a dataset to the original identity of a data subject, and is therefore merely a "useful security measure."[80]

The recently finalized GDPR split the difference, defining pseudonymization as "the processing of personal data in such a way that the data can no longer be attributed to a specific data subject without the use of additional information, as long as such additional information is kept separately and subject

79. Pseudonymization is a "particular type of anonymization that both removes the association with a data subject and adds an association between a particular set of characteristics relating to the data subject and one or more pseudonyms." ISO/TS 25237:2008(E), *supra* note 12, at 3.

80. Article 29, *supra* note 11, at 3 (*emphasis* added).

to technical and organisational measures to ensure non-attribution to an identified or identifiable person."[81] The GDPR recognizes that "[t]he application of pseudonymisation to personal data can reduce the risks for the data subjects concerned and help controllers and processors meet their data protection obligations."[82] At the same time, the GDPR continues to regard pseudonymous information as personally identifiable, providing in Recital 23, "Data which has undergone pseudonymisation, which could be attributed to a natural person by the use of additional information, should be considered as information on an identifiable natural person." It is important to note that data that the GDPR denotes pseudonymous may in certain circumstances fall under this paper's Potentially Identifiable or Not Readily Identifiable categories, as opposed to its Pseudonymous category, which lacks direct identifiers.

The legal rules around pseudonymous data are equally inconsistent. In the U.S., pseudonymous data could in certain circumstances be considered de-identified under the FTC's Three Part Test, depending on the controls in place and on the nature of a particular pseudonym.[83] For example, regulators may check whether an identifier could be cleared or reassigned; whether there is an easily accessible look-up database; or if the data are directly derived from PII. The negotiating drafts of the GDPR reveal a dispute among European policymakers regarding the scope and implications of pseudonymization. For example, the 2014 European Parliament draft would have assigned legal import to "pseudonymous" data by creating a presumption that "profiling based solely on the processing of pseudonymous data" does not "significantly affect the interests, rights or freedoms of the data subject."[84] Later drafts set

81. GDPR, ART. 4(3b).

82. GDPR, Recital 28.

83. FTC, *supra* note 10 at 21.

84. European Parliament legislative resolution of 12 March 2014 on the proposal for a regulation of the European Parliament and of the Council on the protection of individuals with regard to the processing of personal data and on the free movement of such data (General Data Protection Regulation) (COM(2012)0011 – C7-0025/2012 – 2012/0011(COD)), http://www.europarl.europa.eu/sides/getDoc.do?pubRef=-//EP//TEXT+TA+P7-TA-2014-0212+0+DOC+XML+V0//EN (Parliament Draft); *cf.* European Commission, *Proposal for a Regulation of the European Parliament and of the Council on the protection of individuals with regard to the processing of personal data and on the free movement of such data* (General Data Protection Regulation) COM(2012) 11 final, January 25, 2012, http://ec.europa.eu/justice/data-protection/document/review2012/com_2012_11_en.pdf (Commission Draft).

forth by the European Council, however, have removed this presumption and locked pseudonymous data in a state of legal limbo, defined in the law but carrying no apparent legal consequence.[85]

The final draft of the GDPR assigns certain advantages to pseudonymous data when compared to personal data. Article 6(3a) of the GDPR permits companies to repurpose data for another compatible use, taking into account "the existence of appropriate safeguards, which may include encryption or pseudonymisation." Similarly, Article 83, which sets "safeguards and derogations for the processing of personal data for archiving purposes in the public interest or scientific and historical research purposes or statistical purposes," suggests pseudonymization as the primary method to effect data minimization, helping balance the rights of individuals against compelling public interests.[86] Article 23 puts pseudonymization forth as the sole example of data protection by design and by default; and Article 30 suggests "pseudonymisation and encryption" as the primary measures of data security.

E. De-Identified and Protected De-Identified Data

Data in the next category are termed *De-Identified*, if lacking additional safeguards or controls, or *Protected De-Identified*, where additional safeguards or controls are present. Here, direct *and* known indirect identifiers have been either removed or manipulated in a fashion that breaks the linkage between the information and the data subject. Organizations often put in place safeguards on publication and use controls to complement and buttress their technical de-identification measures. This means that the stricter the safeguards and controls, the less perturbed data must be to achieve a sufficiently low risk of re-identification; when data are more perturbed, fewer controls are needed. A significant portion of medical and pharmaceutical research operates on the basis of De-Identified data, which includes information satisfying the HIPAA de-identification standard.

---

85. Council of the European Union, Prop*osal for a Regulation of the European Parliament and of the Council on the protection of individuals with regard to the processing of personal data and on the free movement of such data* (General Data Protection Regulation), June 11, 2015, http://data.consilium.europa.eu/doc/document/ST-9788-2015-INIT/en/pdf ("Council Draft").

86. *See also* GDPR Recital 125.

De-Identified data is the focal point for much controversy within the technical de-identification community, based in large part around different views as to whether certain information in a database is likely to be useful—now or at some point in the future—as an *indirect* identifier. Under a strictly formalist view, any conceivable risk merits treating even remotely-potential indirect identifiers similarly to explicit PII. In determining whether such data are personally identifiable or not, regulators are confronted by difficult policy tradeoffs, including whether and to what extent to limit research in the name of privacy. Experts who distrust any administrative safeguards or legal controls frequently question the merits of de-identification, joined by data scientists skeptical that risk-based de-identification is sustainable or sufficient.[87]

By definition, however, and in contrast to data in the Anonymous category, our category of De-Identified data is intended to accommodate a risk-based approach to de-identification. Rather than focusing on every possible attack vector, pragmatic organizations can calibrate the protections they afford to attacks that are truly feasible or likely to be available to a probable attacker.

F. Anonymous and Aggregated Anonymous Data

Finally, data in the final category may be described as *Anonymous*, or *Aggregated Anonymous*, depending on the level of data aggregation and attendant controls. As with the previous category, both direct and indirect identifiers have been removed or transformed so that they cannot link back to any individual. Unlike De-Identified data, however, Anonymous data features mathematical and technical guarantees that are sufficient on their own to distort the data so as to prevent re-identification. In the Aggregated Anonymous category, the data are so highly aggregated that additional safeguards or controls are no longer relevant (*e.g.*, published census statistics).

Data in the Anonymous category are at the forefront of de-identification science, and must continually confront uncertainties over what it means to re-identify data. Open questions within the de-identification community include: must a record actually be re-identified in order to discredit de-identification,

87. Felten & Narayanan, *supra* note 26.

or is some probability of future re-identification sufficient? How many records need to be re-identified for a database as a whole to be considered broken? Do attackers need to uncover a person's *name* to consider a record re-identified or is it sufficient for them to single out an individual from the database? What level of confidence is needed to declare a dataset either de-identified or re-identified? What consideration should be given to allowing a lower level of perturbation of this data if it is subject to additional controls?

The abundance of big data is believed to undermine de-identification efforts through a combination of powerful new computing capabilities and ever-growing databases of peripheral information. Although some indirect identifiers have repeatedly figured in re-identification attacks (*e.g.*, dates, geolocation, transaction codes), it may be impossible to predict which items of data will *in the future* become linkable indirect identifiers. Already today, lists of movie ratings[88] or the battery life of a mobile device[89] have proven to be susceptible to linkage attacks. At the same time, data points that can be used to pierce through de-identification often carry important social benefits in areas like healthcare, education, and science, rendering their suppression undesirable.

While researchers continue to debate these points, policymakers should be careful not to prove too much with the re-identification risk argument. If, for example, only two individuals can potentially be re-identified from a database of 15,000 patient records, for a match rate of 0.013%,[90] policymakers must weigh those privacy risks against the societal benefit locked into those records. They should account for the fact that overly strict de-identification rules that are geared at eliminating remote privacy risks may jeopardize valuable data uses in return for small privacy gains.

---

88. Narayanan & Shmatikov, *supra* note 3.

89. Lukasz Olejnik et al, *supra* note 25; *see also* Alex Hern, *How Your Smartphone's Battery Life Can Be Used to Invade Your Privacy*, THE GUARDIAN, August 4, 2015, http://www.theguardian.com/technology/2015/aug/03/privacy-smartphones-battery-life.

90. Deborah Lafkey, *The Safe Harbor Method of De-Identification: An Empirical Test*, ONC Presentation, October 8, 2009, www.ehcca.com/presentations/HIPAAWest4/lafky_2.pdf; *see also* discussion in Cavoukian & El Emam, *supra* note 4, at 6.

IV. ADDITIONAL CONSIDERATIONS

A. Sensitivity

For the purposes of this paper, it is important to separate between the sensitivity of a data item and its degree of identifiability. Some risk-based models of regulation factor sensitivity into the overall determination of whether data is classified as "personal" or what level of risk it entails.[91] In contrast, this paper, while recognizing that both identifiability and sensitivity affect an overall risk calculus, seeks to isolate identifiability as an independent characteristic of data. Data do not become any more or less identifiable because of their sensitivity, although there is some overlap between the set of direct identifiers and the set of sensitive data. Information could be highly sensitive yet not readily identifiable (*e.g.*, rare medical condition), or entirely mundane but easily linkable to a specific individual (*e.g.*, IP address). Data that may appear to be non-sensitive, such as a zip code, could ultimately lead to sensitive inferences, such as an individual's race. And certain information may be deemed sensitive by one culture, but not another.

This Article suggests that in assessing risk, sensitivity should be taken into account only after a determination has been made as to identifiability; at which point, policymakers can overlay more nuanced legal protections based on sensitivity, whether the data are identifiable or not. Ultimately, the end results of the proposed analysis may converge with those of a risk-based framework that takes sensitivity into account. For example, data may be deemed de-identified but restricted due to sensitivity concerns.

B. Safeguards and Controls

The success or failure of a de-identification framework ultimately rests on the efficacy of the underlying combination of safeguards and controls used to balance intended uses of data against the risk of re-identification. Determining which controls are required to assure stakeholders, including regulators

---

91. El Emam et al, *supra* note 58; *see* Eloïse Gratton, *If Personal Information is Privacy's Gatekeeper, Then Risk of Harm is the Key: A Proposed Method for Determining What Counts as Personal Information*, 24 ALB. L.J. SCI. & TECH. 105 (2014).

and individuals, that datasets of different degrees of identifiability can be used with limited privacy risks, is context specific.

As a general rule, the more identifiable the information in a dataset, the more circumspect an organization should be in its application of security safeguards and privacy controls. Administrative privacy and security controls are foundational components of data risk analysis. Such controls include policies, contracts, training, and data classification. Yet administrative controls are only effective to a point. They must be complemented with rigorous technical safeguards, such as automated data logging, data retention restrictions, consent management, data analytics restrictions, encryption, access management, and automated data validation.

Clearly, there is no cookie-cutter approach to the use of safeguards and controls. It is therefore imperative that organizations conduct a risk assessment of intended de-identification techniques and document the rationale behind their decision making as part of their de-identification due-diligence

CONCLUSION

A legal system that is closely attuned to different placements of data along an identifiability spectrum will enable organizations to maximize data utility while minimizing privacy risks. Unfortunately, the legacy legal structures in place on both sides of the Atlantic continue to straightjacket policy in this area by insisting on a binary categorization of data and an all-or-nothing approach to privacy protection. Scholars who suggested a layered approach, recognizing the existence of a middle category, have not elaborated a system to categorize and tailor treatment for different flavors of such information.

This paper reframes the debate, proposing a policy model that reflects a reality where data lies along an identifiability spectrum. It posits that for policy purposes, *data should be categorized based on the interplay of three main variables: direct identifiers, indirect identifiers and safeguards and controls on access and use*. It suggests that on the one hand, industry should refrain from referring to information that is tied to unique identifiers as anonymous or non-personal, if such identifiers are shared broadly or are in fact linked to personally identifiable information (Potentially Identifiable data). If companies set tighter controls on access to such data and provide consumers meaningful controls, the same information can

slide down the identifiability scale to merit more liberal legal treatment (Not Readily Identifiable data).

Regulators, on the other hand, should recognize that pseudonymous data affords privacy protections, even if the extent of de-identification is not foolproof (Pseudonymous and Protected Pseudonymous data). Similarly, recognizing the role of a combination of technological, administrative and legal controls, regulators should enable researchers and scientists to work with de-identified data without satisfying privacy requirements, such as obtaining individuals' consent, so as not to impede valuable scientific and technological progress (De-Identified and Protected De-Identified data). Finally, anonymous and aggregated data should be unrestricted, with risk and utility factored into an assessment as to whether a certain data point could become a quasi identifier sometime in the future in remote circumstances. (Anonymous and Aggregated Anonymous data).

A benefit-risk assessment can provide the legal impetus to enable data uses depending on the category of data, intended uses and spectrum of re-identification risk.[92] Indeed, such analysis is couched in existing law and implementation by policymakers and regulators on both sides of the Atlantic. The FTC weighs benefits to consumers when evaluating the unfairness of business practices under Section 5 of the Federal Trade Commission Act.[93] Similarly, the European Article 29 Data Protection Working Party applied a balancing test in its opinion interpreting the legitimate interest clause of the European Data Protection Directive.[94] The GDPR puts forth pseudonymization as a method to protect individuals' rights while at the same time permitting the repurposing of data or its use for research, scientific or statistical endeavors.[95] Finally, the White House Office of Science and Technology Policy, which has recently studied the social and technical ramifications of big

92. Jules Polonetsky, Omer Tene & Joseph Jerome, *Benefit-Risk Analysis for Big Data Projects*, FUTURE OF PRIVACY FORUM WHITE PAPER, Sept. 2014, http://www.futureofprivacy.org/wp-content/uploads/FPF_DataBenefitAnalysis_FINAL.pdf.

93. FTC Act, Section 5(n).

94. Article 29 Data Protection Working Party, Opinion 06/2014 on the notion of legitimate interests of the data controller under Article 7 of Directive 95/46/EC (Apr. 9, 2014), http://ec.europa.eu/justice/data-protection/article-29/documentation/opinion-recommendation/files/2014/wp217_en.pdf.

95. GDPR, Art. 6(3a), 83, rec. 125.

data, recognized the need to strike an appropriate balance between new opportunities and individual values.[96]

New uses of data and technology have the potential to bring humanity a wide range of benefits, but at the same time to generate new and serious harms. Setting the appropriate controls on data based on the potential benefits and the created risks requires understanding the different states of data and establishing appropriate rules for collection, use and controls. This Article advances an approach that supports benefit and deters risk by providing a practical framework for policymakers to analyze various data sets based on their degree of identifiability.

APPENDIX A[97]

In a wide range of industries and research fields, striking an appropriate balance between data utility and individual privacy requires applying practical de-identification measures or maintaining data in varying states along the identifiability spectrum. In areas such as healthcare or human subject research, for example, significant ethical and legal obligations may necessitate an ability to re-link information to a data subject, for example in order to administer medication for a diagnosed disease. In other fields, striving for complete anonymization can degrade the quality of data, skewing results and impeding the utility of important services. The following are illustrative examples of real world circumstances in which practical de-identification plays a critical role:

A. Educational Programs

With the growing popularity of massive open online courses (MOOCs) and their promise of educational attainment for all, it is important for educators, policymakers, and researchers to study and assess the efficacy and development of such programs. In May 2014, for example:

> a team of researchers from Harvard and MIT released an

96. WHITE HOUSE, EXECUTIVE OFFICE OF THE PRESIDENT, BIG DATA: SEIZING OPPORTUNITIES, PRESERVING VALUES (May 2014), http://www.whitehouse.gov/sites/default/files/docs/big_data_privacy_report_may_1_2014.pdf.

97. The information in Appendix A has been gathered from a series of interviews with business leaders who have chosen to remain anonymous for this publication.

> open data set containing student records from 16 courses conducted in the first year of the edX platform. . . The data set is a de-identified version of that used to publish *HarvardX and MITx: The First Year of Open Online Courses*, a report revealing findings about student demographics, course-taking patterns, certification rates, and other measures of student behavior.[98]

The goal of releasing this data was to "allow other researchers to replicate the results of the analysis and to allow researchers to conduct novel analyses beyond the original work, adding to the body of literature about open online courses."[99]

Voluntarily attempting to comply with FERPA's de-identification standards for the public release of student records,[100] the Harvard and MIT researchers used k-anonymity to remove both direct and indirect identifiers from the records. K-anonymity is a process whereby the original dataset containing personally identifiable information is transformed so that it is difficult for an adversary to determine the identity of the individuals in that data set.[101] Technically, a k-anonymized dataset has the property that each record is similar to at least another k-1 other records on the potentially identifying variables.[102] Manipulating their data to meet strict technical de-identification standards, the researchers faced fundamental tradeoffs between privacy on the one hand, and "[their] responsibility to release data for replication and downstream analyses, on the other. For example, the original analysis found that approximately five percent of course registrants earned certificates. Some methods of de-identification cut that percentage in half."[103] The researchers concluded that:

> It is possible to quantify the difference between replications from the de-identified data and original findings; however, it is difficult to fully anticipate whether findings from novel

98. John Daries et al., *Privacy, Anonymity, and Big Data in the Social Sciences: Quality Social Science Research and the Privacy of Human Subjects Requires Trust*, 12 QUEUE 7 (2014), https://queue.acm.org/detail.cfm?id=2661641.

99. *Id.*

100. 34 CFR §99.30.

101. Latanya Sweeney, *K-Anonymity: A Model For Protecting Privacy*, 10 (5) INT'L J. UNCERTAINTY,
FUZZINESS & KNOWLEDGE-BASED SYSTEMS 557 (2002).

102. Khaled El Emam & Fida Kamal Dankar, *Protecting Privacy Using k-Anonymity*, 15(5) *J. Am. Med. Inform. Assoc.* 637 (2008).

103. *Id.*

> analyses will result in valid insights or artifacts of de-identification. Higher standards for de-identification can lead to lower-value de-identified data. This could have a chilling effect on the motivations of social science researchers. If findings are likely to be biased by the de-identification process, why should researchers spend their scarce time on de-identified data?[104]

B. Geolocation and Traffic Services

Traffic applications are a type of location-based service that compiles geolocation data to provide mobile device users with real-time information about their surroundings. This data can be used in an aggregate form to study traffic flows, improve urban planning, and reduce traffic congestion. Typically, when location data collected from a user's mobile device is sent to a phone carrier, operating system or location service provider, any unique identifiers are hashed. The hashed traffic data are then placed into a data vault and enriched with additional location data, before being returned to the user's mobile devices so that the user can, for example, see traffic conditions ahead or recalculate his commute to avoid congestion.

In order to provide a user with relevant, real-time traffic information, location services must track the user's geolocation at particular points in time, and must be able to link traffic reports back to that specific user. At the same time, combinations of technical, administrative, and legal controls can offer protections to users' sensitive location data. These include hashing individual identifiers, aggregating data after a set period of time, and applying contractual use restrictions. For example, before an app shares hashed traffic data with third parties for research or marketing purposes, the data recipients' data practices are reviewed and data are typically further aggregated into traffic reports.

C. Payment Processing

Payment processing companies aggregate billions of payment card transactions to provide value to both merchants and consumers. For example, the information can be used to detect theft and prevent fraud or to improve operational efficiencies. In some cases, payment data must remain quickly linkable to

104. *Id.*

an individual so that the services can readily confirm the purchaser's identity, for example, at a gas pump. Companies currently rely on practical de-identification and aggregation techniques to protect individual privacy as merchants, banks, and payment processing companies handle transaction information.

In order to maintain unique payment records without identifying users, payment processing companies remove all personally identifiable information from the transaction data and subject the account numbers to a one-way hash. Any non-transaction data, such as information from a loyalty rewards program, is aggregated and reviewed before being combined with de-identified transaction data. Information from these sources is placed into separate data warehouses, and can be aggregated into larger and larger datasets depending on the sensitivity of the transaction data. A data analytics team can then access outputs from the data warehouse, for example, to detect fraud. Aggregated reports can also be provided to merchants and banks, and can be combined with other macroeconomic data to gain further insights.

D. Medical Devices

Medical device manufacturers require potentially identifiable data for a range of critical purposes, including: monitoring device performance, conducting safety-related analysis, addressing customer escalations and concerns, and allowing advanced equipment troubleshooting. For example, when a patient complains about warming from a magnetic resonance (MRI) scan, data obtained from the relevant device is needed to understand and remediate this potentially hazardous situation. This, in turn, may require analyzing a wide range of data collected over time (including patient weight, the time and date of each relevant scan, the cumulative effect of multiple scans on the device, and the parameters used within each scan), so manufacturers must be able to link certain scan information to individual patients.

Engineers from medical device manufacturers have determined that more privacy-protective approaches—such as aggregating scans at the device level (as opposed to gathering scan-specific data), removing linkable patient information such as patient weight from the dataset, or replacing actual values with a range of values—are not feasible. Any of these

alternatives would increase machine downtime, adversely impact hospitals' efficiency in providing patient care, and impede device manufacturers' patient-safety and troubleshooting investigations. For example, the date and time of an exam scan are needed to correlate a discrete error event with a specific exam. The precise weight of a patient is needed to approximate various aspects of the physics that occur during an MRI scan, including those that help ensure that the scanner is operating within safe margins (for example, the amount of radio frequency power that may be needed to image a large adult would not be safe for a small child). In addition, a hashed patient ID is needed to group scans by patient to enable medical device manufactures understand whether a series of machine anomalies involved a single patient or different patients. A hashed patient ID is also needed to differentiate between actual exams and test exams, and to perform usage analyses. Using a range of values or aggregating patient IDs would disrupt a medical device manufacturer's ability to perform these needed tasks.

E. Genetic Research

In genetic research, maintaining a capacity to identify individual-level data is critically important. It enables participants in studies to withdraw their consent and allows researchers or clinicians to alert patients to incidental findings, as may be required by law or ethical codes of conduct. Given the sensitivity of individual medical and genetic data and the unsuitability of technically irreversible anonymization, robust de-identification techniques and strong administrative and legal privacy and security controls must be deployed to protect individual privacy.

The privacy certificate Schering AG received from the Schleswig-Holstein DPA in 2003 demonstrates how robust de-identification measures can pass muster with a data protection regulator in the context of pharmacogenetic research. The research-centered German pharmaceutical company collected blood and tissue samples "to determine the relationship between certain genetic information and the effect of pharmaceutical products upon it" and then "subsequently compared with the medical data from the participants which has been obtained in the context of clinical studies."[105] Because German

105. NORBERT LUTTENBERGER, BRIEF REPORT ON THE DATA PROTECTION

law required that participants in clinical trials have the right to withdraw consent and have their data verifiably deleted at any time, data could not be fully anonymized. Instead, recognizing "the fact that with scientific evaluation, an identification isn't required and a matching of genetic and medical data is enough in each case," the company developed a double-pseudonym process to protect the data from both external and internal attacks through role-based access permissions, firewall systems, encryption, and other controls. The certified program provided protection such that "Even in the case that genetic or clinical data fall into the hands of an unauthorized party, the activation of the matching procedure requires two separate and independent sites."

F. Mobile Devices

Mobile devices provide powerful connectivity, computing power and utility. These devices use a complex interplay between hardware, firmware, operating system, applications, and wireless networks (including cellular, WiFi, Bluetooth, GPS, and others). Each element of the wireless ecosystem is often provided by a different company competing to provide the best possible consumer experience—better connectivity, enhanced reliability and security, faster speeds, more computing power, more capabilities and functionalities through sensors and other technologies, and longer battery life. By analyzing data about how mobile devices are used in real-world settings, these companies reduce dropped calls or poor connections, enhance security, mitigate conflicts and crashes in mobile operating systems and apps, and extend battery life. Other identifiers are used for advertising or analyzing app usage.

In many cases, companies do not need to collect customer names, email or street addresses, phone numbers, or web browsing information—basic identifiers linked to the device and its usage provides a plethora of benefits. For example, a device identifier, such as a UDID or IMEI, software and firmware versions, make, model, and operating system could help identify issues across classes of devices as well as issues with

---

AUDIT: DATA PROCESSING INFRASTRUCTURE CONCEPT OF THE SCHERING CORPORATION FOR THE SECURE PSEUDONYM STORAGE AND KEEPING OF BLOOD AND TISSUE SAMPLES INTENDED FOR GENETIC ANALYSES, https://www.datenschutzzentrum.de/audit/kurzgutachten/a0303/a0303_engl.htm.

particular device configurations. Information about the network quality, cellular signal quality, and device location could be harnessed to improve network performance, while information about the CPU/GPU/DSP and other hardware usage and heat the device is generating used to improve hardware and software configurations. Knowing information about nearby WiFi signals and other wireless networks also could help offload data from cellular networks, as well as improve the performance of location services. Similarly, information on a device's configurations, hardware and software is key to reducing conflicts and crashes. Simply knowing the number of devices with a particular software version could help understand the risk for a known software flaw. Without this data, companies risk staying blind to the scope of a security problem.

While some mobile or advertising identifiers may be widely shared and linked to explicit personal information, others are subject to tight controls, easy for users to clear and reliably limited from being linked to other identifiers. Some identifiers are regularly rotated by the provider, limiting the ability of third parties to use them over time.