The Seven States of Data: When is Pseudonymous Data Not Personal Information ?

Khaled El Emam[1], Eloise Gratton[2], Jules Polonetsky[3], Luk Arbuckle[4]

[1] University of Ottawa & Privacy Analytics Inc.
[2] Borden Ladner Gervais LLP
[3] Future of Privacy Forum
[4] Children's Hospital of Eastern Ontario Research Institute

I. INTRODUCTION

There has been considerable discussion about the meaning of personal information and the meaning of identifiability. This is an important concept in privacy because it determines the applicability of legislative requirements: data protection laws ("DPL") around the world protect and govern personal information. A common definition of personal information as "information pertaining to an identifiable individual" can be found in all DPL around the world.[1] Consequently, the notion of "identifiability" becomes key in the interpretation and application of these laws.

There is a general view that identifiability falls on a spectrum, from no risk of re-identification to fully identifiable[2], with many precedents in between[3]. Recently, a number of legal scholars have proposed different approaches to determine at what point information should be considered as "personal", in many cases using a risk based approach.[4] For instance, Schwartz and Solove define three specific states of data: identified, identifiable, and non-identifiable[5]. Identified information is that which "singles out a specific individual from others".[6] Identifiable information under Schwartz and Solove's definition is information that does not currently "single out" an individual but could be used to identify an individual at some point in the future. Finally, non-identifiable information is that which cannot reasonably be linked to an individual. Rubenstein acknowledges that a range of identifiability, or conversely de-identification, exists and that these concepts need to be more clearly defined in terms of the risk posed in order to ensure that the risk can be effectively mitigated.[7] Polonetsky, Tene and Finch define various points on the spectrum of identifiability, from "explicitly personal" information which contains direct and indirect identifiers without any safeguards or controls to "aggregated anonymous" information from which direct and indirect identifiers have been removed or transformed and no controls are required due to the

[1] US Congress, 'The Health Insurance Portability and Accountabibid; Personal Information and Electronic Documents Act (PIPEDA) 2000 c. 5; 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). 2016.ility Act of 1996; 42 U.S. Code § 1320d - Definitions' <http://www.law.cornell.edu/uscode/text/42/1320d>.

[2] Paul Schwartz and Daniel Solove, 'The PII Problem: Privacy and a New Concept of Personally Identifiable Information' (2011) 86 New York University Law Review 1814.

[3] Khaled El Emam, 'Heuristics for De-Identifying Health Data' [2008] IEEE Security and Privacy 72; Jules Polonetsky, Omer Tene and Kelsey Finch, 'Shades of Gray: Seeing the Full Spectrum of Practical Data de-Identification' (2016) In press Santa Clara Law Review 593.

[4] Ira Rubinstein and Woodrow Hartzog, *Anonymization and Risk*, Washington Law Review, Vol. 91, No. 2, 2016 NYU School of Law, Public Law Research Paper No. 15-36; Voir également Paul Ohm, *Broken Promises of Privacy : Responding to the Surprising Failure of Anonymization* (August 13, 2009). UCLA Law Review, Vol. 57, p. 1707, 2010 ; U of Colorado Law Legal Studies Research Paper No. 9-12; Boštjan Bercic & Carlisle George, *"Identifying Personal Data Using Relational Database Design Principles"* (2009) 17:3 International Journal of Law and Information Technology 233; Lundevall-Unger and Tranvik, *"IP Addresses: Just a Number?"* (2011) 19:1 International Journal of Law and Information Technology 53; Paul Schwartz & Daniel Solove, "*The PII Problem: Privacy and a New Concept of Personally Identifiable Information"* (2011) 86 N.Y.U. Law Review 1814; 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"*, Albany Law Journal of Science & Technology, Vol. 24, No. 1, 2013.

[5] Paul Schwartz and Daniel Solove (n 2).

[6] ibid.

highly aggregated nature of the data.[8] Along this spectrum, they place pseudonymous data somewhere near the middle and argue that it can pose more or less risk depending on the methods used to transform direct identifiers and the safeguards and controls applied.

In this article, we extend the previous work in this area by:

a) mapping the spectrum of identifiability to a risk-based approach for evaluating identifiability which is consistent with practices in the disclosure control community

b) defining precise criteria for evaluating the different levels of identifiability

c) proposing a new point on this spectrum using the same risk-based framework that would allow broader uses of pseudonymous data under certain conditions.

We aim to strengthen the existing literature on the spectrum of identifiability by proposing a precise framework that is consistent with contemporary regulations and best practices. The identifiability "states of data" proposed here are colored by our experiences with health data[9], although they may nevertheless be useful much more broadly to other domains.

The Current States of Data

The context we assume is that of a *data custodian* who is sharing data with a *data recipient*. We need to understand the state of the data that is being exchanged, and whether or not it should be considered personal information.

To characterize the states of data, we consider five interdependent characteristics of the data or the data sharing transaction itself:

1. **Verifying the identity of the data recipient.** This determines whether the data custodian knows with high confidence the identity of the individual or organization that is receiving the data, enabling the custodian to hold the data recipient accountable for any breach of contract or terms-of-use

2. **Application of masking.** Masking techniques perturb the direct identifiers in a data set. Another common term that is used for this kind of data is "pseudonymous" or "pseudonymized" data.

[9] Khaled El Emam and Luk Arbuckle, Anonymizing Health Data: Case Studies and Methods to Get You Started (O'Reilly 2013).

3. **Application of de-identification.** De-identification techniques perturb the indirect identifiers, such as the dates, ZIP codes, and other demographic and socio-economic data[10]. This perturbation can range from low (e.g. a date of birth converted to month and year of birth) to high (e.g., a date of birth converted to a 5 year interval).

4. **Contractual controls.** Having the data recipient sign an enforceable contract that prohibits re-identification attempts, among other safeguards, is considered a strong control.[11]

5. **Establishment of Security and Privacy Controls.** Such controls can vary in their strength[12], but they reduce the likelihood of a rogue employee at the data recipient attempting a re-identification attack and the likelihood of a data breach occurring[13]. Standard security and privacy practices would make up this set of controls.

Using these characteristics, we describe the current six different states of data as shown in **Table 1**. These six states reflect the type of data sharing that is happening today based on our observations.

		Verify Identify of Data Recipient	**Masking (of Direct identifiers)**	**De-identification (of Quasi-identifiers)**	**Contractual Controls**	**Security & Privacy Controls**
Not-PII	Public Release of Anonymized Data	NO	YES	HIGH	NO	NONE
	Quasi-Public Release of Anonymized Data	YES	YES	MEDIUM-HIGH	YES	NONE
	Non-Public Release of Anonymized Data	YES	YES	LOW-MEDIUM	YES	LOW-HIGH
PII	Protected Pseudonymized Data	YES	YES	NONE	YES	MEDIUM
	"Vanilla" Pseudonymized Data	YES	YES	NONE	NO	NONE
	Raw Personal Data	YES	NO	NONE	NONE	NONE

Table 1: The current six states of data.

One of the factors that determine whether data is personally identifying information is whether the data is considered to be de-identified. As shown in Table 1, data in which the direct identifiers have been removed or pseudonymized (e.g., masked) and the indirect identifiers left intact is considered pseudonymous. Most known

[10] Khaled El Emam, *Guide to the De-Identification of Personal Health Information* (CRC Press (Auerbach) 2013).
[11] ibid; Health Information Trust Alliance, 'HITRUST De-Identification Framework' (HITRUST Alliance 2015) <https://hitrustalliance.net/de-identification/>; HITRUST Alliance, 'HITRUST Common Security Framework'.
[12] HITRUST Alliance (n 11).
[13] Juhee Kwon and M Eric Johnson, 'Security Practices and Regulatory Compliance in the Healthcare Industry' (2013) 20 Journal of the American Medical Informatics Association 44.

successful re-identification attacks have been performed on pseudonymous data[14]. In the EU, the Article 29 Working Party has made clear that pseudonymous data is considered personal information[15], and this interpretation has been incorporated into the recent General Data Protection Regulation (GDPR) [16]. Under the US HIPAA, the limited data set is effectively pseudonymous data and is still considered to be protected health information.[17] Therefore, all variants of pseudonymous data are considered personal information. Conversely, when the data has been de-identified and both direct and indirect identifiers have been perturbed to ensure that the risk of re-identification is very small, the data is no longer considered to be personal information and therefore is no longer subject to the regulations.

We can now examine the six states in more detail:

Public Release of Anonymized Data. This is essentially open data, where files containing individual-level data (known as "microdata") are made available for download by anyone. There are no restrictions on who gets the data and there is no request for the identity of the entity or individual who is downloading the data. This type of data must be masked and is subject to extensive perturbation in order to de-identify it.

Quasi-Public Release of Anonymized Data. Quasi-public data is still public in that anyone can request access to the files with individual-level information. However, the identity of these data recipients needs to be verified and they must sign on to a terms-of-use agreement that prohibits re-identification.

Non-Public Data Release. When data is released in a non-public manner, those requesting the data must be qualified first. For example, they may have to be research investigators at a recognized academic institution. Or they may have to be recognized businesses that have a legitimate purpose to use the data. Such data releases require strong contractual controls and the data custodian would impose specific security and privacy controls on the data recipient.

[14] Khaled El Emam and others, 'A Systematic Review of Re-Identification Attacks on Health Data' (2011) 6 PLoS ONE <http://www.plosone.org/article/info%3Adoi%2F10.1371%2Fjournal.pone.0028071>.

[15] Article 29 Data Protection Working Party, 'Opinion 05/2014 on Anonymization Techniques' (2014) WP216; K El Emam and C Alvarez, 'A Critical Appraisal of the Article 29 Working Party Opinion 05/2014 on Data Anonymization Techniques' [2014] International Data Privacy Law.

[16] 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). (n 1).

[17] US Congress, 'The Health Insurance Portability and Accountability Act of 1996; 45 Code of Federal Regulations 164.154(b)5(e) Limited Data Set' <https://www.law.cornell.edu/cfr/text/45/164.514>.

The level of data perturbation is different among these three data states. For public data the level of perturbation is high, which means that data quality is degraded. For non-public data the level of perturbation is small, and therefore the data quality can remain quite high.

Up to this point the data can credibly be determined not to be personal information, on the assumption that proper de-identification methods have been applied, and the contractual, security, and privacy controls are reasonable.

The non-public data states below are, at least under existing definitions, considered to be personal information.

> **Protected Pseudonymous Data.** Data is pseudonymous when only masking of direct identifiers has been applied and no de-identification methods are used. *Protected* pseudonymous data has additional contractual, security, and privacy controls in place. These controls reduce the risk of re-identification considerably, but may not necessarily bring it below the threshold to be considered non-PII.
>
> **"Vanilla" Pseudonymous Data.** This is pseudonymous data without any of the additional contractual, security or privacy controls in place. This means that only masking of direct identifiers has been applied and any indirect identifiers included in the data set remain intact.
>
> **Raw Personal Data.** This is data that has not been modified in any way or that has been modified so little that the probability of re-identification is still very high.

Based on our categorization above, the question is whether there is a form or state of pseudonymous data that could be considered not-PII? Or, if it is considered PII, can that information nonetheless be used and disclosed for secondary purposes without consent or authorization under certain conditions?

II. Pseudonymous Data

On-going debate in the privacy community is how to treat "protected pseudonymous data", and whether it should receive special treatment.

The EU Article 29 Working Party has advocated that for the equivalent of protected pseudonymous data as we have defined it[18], which is consistent with other interpretations by the Working Party and the disclosure control community[19]:

[18] Article 29 Data Protection Working Party, 'Opinion 4/2007 on the Concept of Personal Data' (2007) WP136.

[19] Khaled El Emam and Cecilia Álvarez, 'A Critical Appraisal of the Article 29 Working Party Opinion 05/2014 on Data Anonymization Techniques' (2015) 5 International Data Privacy Law 73.

> *"even though data protection rules apply, the risks at stake for the individuals with regard to the processing of such indirectly identifiable information will **most often be low**, so that the application of these rules will **justifiably be more flexible** than if information on directly identifiable individuals were processed." [emphasis added]*

Although the Working Party does not make clear what this additional flexibility would mean.

In the EU general data protection regulation (GDPR), pseudonymous data is defined in the following manner[20]:

> *"'pseudonymisation' means the processing of personal data in such a manner that the personal data can no longer be attributed to a specific data subject without the use of additional information, provided that such additional information is kept separately and is subject to technical and organisational measures to ensure that the personal data are not attributed to an identified or identifiable natural person".*

This definition is consistent with our state of protected pseudonymous data, in that it considers these additional controls (additional information is kept separately and subject to technical and organizational measures) as part of the definition. However, the interpretation of how to treat pseudonymous data under this regulation can be somewhat challenging as the treatment of pseudonymous data under several Articles of the GDPR (e.g. Articles 23, 32 and 34), is more akin to that of de-identified data rather than personal information. Although pseudonymous data remains personal information under the GDPR, the additional controls that are incorporated into its definition distinguish it from raw personal data and allow for some limited flexibility in its use.

Under the HIPAA Privacy Rule there is the concept of a limited data set, which requires that: (a) common direct identifiers be removed from the data set, (b) the purpose of the disclosure may only be for research, public health or health care operations, and (c) the person receiving the information must sign a data use agreement with the data custodian[21]. The data use agreement must require the recipient to use appropriate safeguards, not re-identify the data or contact individuals in the data, and ensure that the restrictions set forth in the agreement are passed on to any agents that they provide the data to, among other restrictions. See Table 2 for a comparison of the treatment of pseudonymous data under GDPR and HIPAA.

[20] 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). (n 1).

[21] US Congress, 'The Health Insurance Portability and Accountability Act of 1996; 45 Code of Federal Regulations 164.154(b)5(e) Limited Data Set' (n 15).

Limited data sets are effectively pseudonymous data, which include dates, location, as well as other indirect identifiers. HIPAA explicitly states that "a limited data set is **protected health information**" [emphasis added], despite the additional conditions imposed on the data sharing transaction[22].

	Conditions	**Flexibilities**
General Data Protection Regulation	"can no longer be attributed to a specific data subject without the use of additional information, provided that such additional information is kept separately and is subject to technical and organizational measures to ensure that the personal data are not attributed to an identified or identifiable natural person"	• Allows for "processing for a purpose other than that for which the personal data have been collected" (by the same controller, without authorization of the data subjects) • Allows for "the further processing of personal data for archiving purposes in the public interest, scientific or historical research purposes or statistical purposes"[23]
HIPAA Limited Data Set	Remove 16 data elements: 1) Names; 2) Postal address information, other than town or city, State, and zip code; 3) Telephone numbers; 4) Fax numbers; 5) Electronic mail addresses; 6) Social security numbers; 7) Medical record numbers; 8) Health plan beneficiary numbers; 9) Account numbers; 10) Certificate/license numbers; 11) Vehicle identifiers and serial numbers, including license plate numbers; 12) Device identifiers and serial numbers; 13) Web Universal Resource Locators (URLs); 14) Internet Protocol (IP) address numbers; 15) Biometric identifiers, including finger and voice prints; and 16) Full face photographic images and any comparable images. A data use agreement must be signed. Use limited to research, public health or health care operations.	No patient authorization required to use the data**.**

Table 2: Regulatory treatment of pseudonymous data

[22] ibid.

[23] 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). (n 1).

Therefore, pseudonymous data with conditions such as contractual, security and privacy controls imposed on it is still considered personal information. The required conditions have not been specified (e.g., “use appropriate safeguards”), however, as it remains personal information, there is an expectation that it be treated with the same level of controls as other personal information. Pseudonymous data with these additional controls may be afforded special or “flexible” treatment[24], although what this flexibility entails is not specified precisely either.

III. Additional Conditions on Protected Pseudonymous Data

If the use of pseudonymous data with additional protections imposed remains restricted to the same level as personal information, then there is little incentive for data custodians to invest in these protections. As we have seen, under current regulations there is some limited flexibility in regards to the use of protected pseudonymous data, as in the case of the LDS. But with the LDS, use and disclosure of this data is not unlimited; for example, sale of the data for secondary use is highly restricted. Under the GDPR, pseudonymized data can be used for secondary purposes by the same controller but cannot be shared with another party without consent (limited exceptions apply). If, however, protected pseudonymous data was allowed to be shared under certain conditions without express consent of the individual data subjects, the benefit to data custodians in terms of simplifying the data release process could be a strong motivating factor in the broad implementation of such protections.

Adding Flexibility to the Processing of Protected Pseudonymized Data

In this section, we define three specific criteria that would further reduce the risk of re-identification for protected pseudonymous data: (1) No processing by humans; (2) No PII leakage from analytics results; and (3) No sensitive data.[25]

[24] Soumitra Sengupta, Neil S Calman and George Hripcsak, ‘A Model for Expanded Public Health Reporting in the Context of HIPAA’ (2008) 15 Journal of the American Medical Informatics Association: JAMIA 569.

[25] See Daniel J. Solove, “Privacy and Power: Computer Databases and Metaphors for Information Privacy” (2001) 53 Stan. L. Rev. 1393 at 1418: “Being observed by an insect on the wall is not invasive for privacy; rather, privacy is threatened by being subject to *human* observation, which involves judgments that can affect one’s life and reputation. Since marketers generally are interested in aggregate data, they do not care about snooping into particular people’s private lives. Much personal information is amassed and processed by computers; we are being watched not by other humans, but by machines, which gather information, compute profiles, and generate lists for mailing, emailing, or calling. This impersonality makes the surveillance less invasive.” Ryan Calo also raises that perhaps the harm resulting from online surveillance by marketers is less important if the data is only viewed by a machine instead of an individual or a human making a judgment. See Ryan Calo, “The Boundaries of Privacy Harm” (2011) 86:3 Indiana Law Journal 1131 at 25. See also Éloïse Gratton, *Personalization, Analytics, and Sponsored Services: The Challenges of Applying PIPEDA to Online Tracking and Profiling Activities*, Comment, Canadian Journal of Law and Technology, Thompson-Carswell, November 2010, raising that the provisions of data protection laws should be interpreted in a flexible manner in order to ensure that online sponsored services and web analytics activities can be undertaken, especially if they are not harmful to individuals.

1. No processing by humans

We can utilize an existing risk assessment framework[26] to evaluate some of the processing risk. Non-public data may be vulnerable to three types of attacks:

> **Deliberate attack.** This is when the data recipient deliberately attempts to re-identify data subjects. Protected pseudonymized data would have a low probability for this type of attack due to the incorporation of contractual, security and privacy controls.
>
> **Breach.** This is when there is a data breach and the data ends up in the "wild". Protected pseudonymized data would have a low probability for this type of attack with strong security and privacy controls to mitigate the risk.
>
> **Inadvertent re-identification.** This occurs when a data analyst inadvertently recognizes someone they know in the data set. In principle, there are two ways to mitigate against this: (a) ensure that no humans are working with the released data (i.e., all processing is algorithmic), or (b) ensure that the data analysts are located in a geography in which they are very unlikely to know a data subject. With respect to the latter, geographic separation may leave some groups of people vulnerable as there may be special cases, such as when the data subject is an internationally known public figure, in which the geographically separated analyst may still be able to recognize the data subject (e.g.

Therefore, if protected pseudonymized data is processed only by machine and not by humans then the risk from these attacks can be considered to be diminished.

2. No PII leakage from analytics results

We assume that the objective of sharing information for secondary purposes is most often to extend the analytic potential of the information by making it available for new and different types of analyses. A key criterion for the secure sharing of processed data or the results of data analyses is that these results do not leak information. There are a number of ways that analytics results can leak information, and these have been documented thoroughly elsewhere.[27]

[26] Committee on Strategies for Responsible Sharing of Clinical Trial Data; Board on Health Sciences Policy; Institute of Medicine., *Sharing Clinical Trial Data: Maximizing Benefits, Minimizing Risk* (Washington (DC): National Academies Press (US); 2015) <http://www.ncbi.nlm.nih.gov/books/NBK269030/>; The Expert Panel on Timely Access to Health and Social Data for Health Research and Health System Innovation, 'Accessing Health And Health-Related Data in Canada' (Council of Canadian Academies 2015); Health Information Trust Alliance (n 11).

[27] L Willenborg and T de Waal, *Elements of Statistical Disclosure Control* (Springer-Verlag 2001); Christine O'Keefe and James Chipperfield, 'A Summary of Attack Methods and Confidentiality Protection Measures for Fully Automated Remote Analysis Systems' (2013) 81 426; K Muralidhar and R Sarathy, 'Privacy Violations in Accountability Data Released to the Public by State Educational Agencies',

Therefore, it is important to ensure that any automated analysis performed on the data does not produce results that would leak identifying information to the end-user. There are techniques that can be applied for specific types of analyses to avoid the leakage of such identifying information[28]. This is a technical consideration and would require the confirmation of an expert to ensure that the likelihood of re-identification from analytic results is acceptably reduced.

3. No sensitive data

Another condition of data sharing is that the data itself is not considered sensitive. We suggest to build on Gratton's previous work on the issue of when should personal information be considered as sensitive, as well as on the type of criteria which should be considered in order to determine whether the information may trigger a risk of harm for the individual.[29] She maintains that there are two types of sensitive information: first, the type of information which may trigger a subjective type of harm (usually upon personal information being collected or disclosed), and a second type which may trigger an objective type of harm (usually upon personal information being used).

A. Information triggering a subjective harm

A first type of harm, which is a more subjective type of harm in the sense that it is associated with "injury to the feelings"[30], feelings of humiliation or embarrassment, would most likely take place upon personal information being collected or disclosed[31]. In order to be harmful to an individual (and therefore, be considered as sensitive data), a disclosure of personal information would therefore have to create some type of humiliation or embarrassment. Given that the *risk of harm* upon a disclosure is highly contextual and can be difficult to isolate, an option is to interpret the notion of "identifiable" in light of the overall sensivity of information in question, by using additional criteria relating to the information which may be used when interpreting the notion of "identifiable"

Federal Committee on Statistical Methodology Research Conference (2009); Subcommittee on Disclosure Limitation Methodology, 'Statistical Policy Working Paper 22 - Report on Statistical Disclosure Limitation Methodology' <http://www.fcsm.gov/working-papers/SPWP22_rev.pdf> accessed 28 October 2011; D Algranati and J Kadane, 'Extracting Confidential Information from Public Documents: The 2000 Department of Justice Report on the Federal Use of the Death Penalty in the United States' (2004) 20 Journal of Official Statistics 97; Khaled El Emam and others, 'A Secure Distributed Logistic Regression Protocol for the Detection of Rare Adverse Drug Events' [2012] Journal of the American Medical Informatics Association <http://jamia.bmj.com/content/early/2012/08/06/amiajnl-2011-000735>.

[28] Christine M O'Keefe and Donald B Rubin, 'Individual Privacy Versus Public Good: Protecting Confidentiality in Health Research' (2015) 34 Statistics in Medicine 3081.

[29] See generally, Gratton, *Understanding Personal Information: Managing Privacy Risks* (LexisNexis 2013).

[30] Warren and Brandeis, 'The Right to Privacy' (1890) IV Harvard Law Review.

[31] William L. Prosser ("Prosser") discusses how the common law recognizes a tort of privacy invasion in cases where there has been a "[p]ublic disclosure of embarrassing private facts about the plaintiff" William Prosser, 'Privacy' (1960) 48 California Law Review 383. According to Calo, the subjective category of privacy harm (which is included in the activity of collecting and disclosing personal information) is the unwanted perception of observation, broadly defined and including the collection and disclosure if personal information [31]. R Calo, 'The Boundaries of Privacy Harm' (2011) 86 Indiana Law Journal.

and which may be essential to the identification of this kind of harm. Gratton articulates the view that these additional criteria are the “**intimate**” nature of the information, and the extent of its “**availability**” to third parties or the public upon this information being disclosed.[32]

Information of an **intimate nature** may include data “revealing racial or ethnic origin, political opinions, religious or philosophical beliefs, trade-union membership, and the processing of data concerning health or sex life” [33]. The type of information which tends to reveal intimate details of the lifestyle and personal choices of the individual[34], and criminal records, geolocation and metadata[35] may also be considered as information of intimate nature.

If a given set of information is already in circulation or already **available** to the party receiving the information, then the sensitivity of the information decreases since the risk of subjective harm that may be triggered by the disclosure of information is less substantial (in the sense that individuals will rarely be embarrassed nor humiliated following the disclosure of information already available or already known)[36]. In the Information Age, with new technologies and the web, most information that is disclosed may have been previously available to a certain extent. One must therefore focus on the extent to which the information is made more accessible[37] in order to assess its sensitivity.

B. Information triggering an objective harm

Another type of information which may be considered as sensitive, is information which may trigger a more objective harm upon being used against an individual or which, upon being used, may lead to a negative impact for the individual behind the information. The objective category of privacy harm would therefore be the unanticipated or forced use of personal information against a given person[38]. The type of harm that may result

[32] Gratton, *Understanding Personal Information: Managing Privacy Risks* (LexisNexis 2013) , at section entitled “ Risk of subjective Harm: Revisiting the Sensitivity Criteria” at p. 261 and following. See also 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”*, Albany Law Journal of Science & Technology, Vol. 24, No. 1, 2013. We note that the proposed criteria are very close to what Nissenbaum prescribes when she discusses how the principle of restricting access to personal information usually focuses on data that is “intimate”, “sensitive”, or “confidential”. Helen F. Nissenbaum, “Privacy as Contextual Integrity” (2004) 79:1 Washington Law Review 119 at 128.

[33] Directive 95/46/EC on the protection of individuals with regard to the processing of personal data and on the free movement of such data 1995 46.

[34] *R. v. Plant*, [1993] 3 SCR 281, 1993 CanLII 70 (SCC).

[35] Paul Ohm, *Sensitive Information*, Southern California Law Review, Vol. 88, 2015

[36] Sipple v Chronicle Publishing Co, 154 Cal App 3d 1040, 201 Cal Rptr 665 California Court of Appeal AO11998, newspapers disclosed the fact that Oliver Sipple, who heroically saved President Ford from an assassination attempt, was homosexual. The court concluded that his sexuality was not private because it was already known in the gay community.

[37] Daniel J Solove, ‘A Taxonomy of Privacy’ (2006) 154 University of Pennsylvania Law Review 477.

[38] Documents from the early 1970s produced in the context of the adoption of DPLs already raised the concern of having organizations use the information of individuals in a way which would be detrimental to them ‘Report of the Secretary’s Advisory Committee on Automated Personal Data Systems’ ; In the late 1970s, in the U.K., while discussing the adoption of a DPL or some type of regulation incorporating the FIPs, the Lindop Committee was already suggesting that individuals should be able to know if their data was to be used as the basis of “an adverse decision against them”, and that “outdated data” should be discarded especially when “used for making decisions which affect the data subject”. See Norman Lindop, ‘Report of the Committee on Data Protection: Presented to Parliament by the Secretary of State for the Home Department by Command of Her Majesty’ (HMSO 1978).

from the use of personal information, include financial damages suffered as a consequence of loss of employment, business or professional opportunities, identity theft, negative effects on the credit record and damage to or loss of property, information inequality[39] (discrimination) or physical harm (personal safety and geolocation)[40].

Information may in certain cases be "used" by organizations for various purposes which may have no impact whatsoever on an individual, a very indirect and limited impact, or even a positive one. Gratton argues that in such cases, the information should less likely be considered as sensitive information. If the personal information used may have a "negative impact" (objective harm) on the individual (financial harm, discrimination or physical harm), the information may be considered as being sensitive.

Summary of Conditions

We therefore define another type of protected pseudonymous data that has conditions placed on its use and disclosure. The conditions would be threefold:

1. No humans actively work on the analysis of the data and all processing throughout the lifetime of the data is automated. The automation requirement also implies in practice that the data is transient and would be destroyed once the processing has been completed.
2. The analytic results do not reveal information which can lead to inferences of individual identity.
3. The data is non-sensitive according to the previous discussion. That is, there are no subjective or objective types of harm associated with the data.

Under the existing risk management framework applied extensively to health information, data that meets the conditions above would be considered to have a significantly lower risk of re-identification than other types of pseudonymous data. This type of data could arguably be treated with greater flexibly. Flexibility in this case means that it could be shared and/or sold for secondary purposes without having to obtain consent. This does not mean that the data is not personal information – it would still be considered personal information because the risk of re-identification would not generally fall below a common threshold – but because the risk is lowered and is close to the "de-identified" threshold, some added flexibility could be afforded in its processing.

[39] MJ Van den Hoven and J Weckert, *Information Technology and Moral Philosophy* (Cambridge University Press 2008). See also Paul Ohm, *Sensitive Information*, Southern California Law Review, Vol. 88, 2015.
[40] Paul Ohm, *Sensitive Information*, Southern California Law Review, Vol. 88, 2015

However, in the event of a data breach of flexible pseudonymous data, security and privacy controls, as well as contractual controls would be insufficient to protect the data, and personal information in the form of indirect identifiers would be disclosed. As a result, breached pseudonymous data may require notification to all affected individuals, unless it could be shown that the risk of re-identification was reasonably small (e.g., a data set with no demographic information could conceivably be of very low risk).

Based on this distinction, we can then propose the seven states of data as in Table 3.

·		**Verify Identify of Data Recipient**	**Masking (of Direct identifiers)**	**De-identification (of Quasi-identifiers)**	**Contractual Controls**	**Security & Privacy Controls**
Not-PII	Public Release of Anonymized Data	NO	YES	HIGH	NO	NONE
	Quasi-Public Release of Anonymized Data	YES	YES	MEDIUM-HIGH	YES	NONE
	Non-Public Release of Anonymized Data	YES	YES	LOW-MEDIUM	YES	MEDIUM-HIGH
No Consent Required	**Flexible Pseudonymized Data***	YES	YES	NONE	YES	HIGH
PII	Protected Pseudonymized Data	YES	YES	NONE	YES	MEDIUM
	“Vanilla” Pseudonymized Data	YES	YES	NONE	NO	NONE
	Raw Personal Data	YES	NO	NONE	NONE	NONE

***with conditions**

Table 3: The Seven states of data defined.

IV. Conclusions

In this article, we have defined the spectrum of identifiability and specific criteria for the placement of different types of data along this spectrum. Their criteria follow a well established framework for evaluating the risk of re-identification of data sets and are consistent with contemporary definitions of personal information. Under current regulations, “protected” pseudonymous data is granted some flexibility but this flexibility is limited and not clearly defined. By adding more conditions and safeguards to the existing state of protected pseudonymous data, we propose that more flexibility can be granted for the use and disclosure of the data while still being consistent with contemporary risk management frameworks.