# 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*

DRAFT

**[FINAL DRAFT VERSION]**

# 1 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. As a point of clarification, while European jurisdictions usually refer to "personal data"[1] and North American jurisdictions such as Canada to "personal information",[2] throughout this analysis, the term of reference will be "personal information" or PII and the words "information" or "data" (or "personal information" and "personal data") may be used interchangeably.

There is a general view that identifiability falls on a spectrum, from no risk of re-identification to fully identifiable[3], with many precedents in between[4]. This spectrum has been defined previously in terms of a probability[5]. Recently, many legal scholars have proposed different approaches to determine at what point information should be considered as "personal information", in many cases, using a risk based approach.[6] For instance, Schwartz and Solove define three specific states of data: identified, identifiable, and non-identifiable[7]. There is also the on-going distinctions between de-identified and pseudonymous data, whereby the latter is considered personal information by regulatory authorities.

In this article we examine these points or states on the spectrum of identifiability in more detail, and characterize the manner in which they differ. These identifiability "states of data" are colored by our experiences with health data[8], although they may nevertheless be useful much more broadly to other domains.

DRAFT

[1] European Data Protection Direcive, 1995, which uses the term 'personal data'.

[2] Although in the U.S., sectoral data protection laws (DPLs) usually refer to "personally identifiable information" ("PII") instead of *personal information* or *personal data*. As a matter of fact, U.S. laws protecting personal information often refer to "PII" which stands for "personally identifiable information*"*. See for example COPPA and *California Online Privacy Protection Act*, Bus & Prof. Code §§ 22575-22579 (2004).

[3] 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.

[4] Khaled El Emam, 'Heuristics for De-Identifying Health Data' [2008] IEEE Security and Privacy 72.

[5] CJ Skinner and MJ Elliot, 'A Measure of Disclosure Risk for Microdata' (2002) 64 Journal of the Royal Statistical Society: Series B (Statistical Methodology) 855.

[6] 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; see also 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.

[7] Paul Schwartz and Daniel Solove (n 1).

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

The objective of our analysis, in addition to being descriptive, is to formulate criteria for the sharing of pseudonymous data. These criteria are consistent with existing notions of re-identification risk and therefore do not require changes to the current framework for defining identifiability.

## *1.1 Terminology*

To present an analysis of identifiability, basic concepts need to be defined and a terminology established. It is useful to differentiate among the different types of variables in a data set. We make a distinction among three types of variables[9].

**Directly identifying variables.** One or more direct identifiers can be used to uniquely identify an individual, either by themselves or in combination with other readily available information. For example, there are more than 200 people named “John Smith” in Ontario (based on a search in the White Pages), therefore the name by itself would not be directly identifying, but in combination with the address it would be directly identifying information. A telephone number is not directly identifying by itself, but in combination with the readily available White Pages it becomes so. Other examples of directly identifying variables include email address, health insurance card number, credit card number, and social insurance number. These numbers are identifying because there exist public and/or private databases that an adversary can plausibly get access to where these numbers can lead directly, and uniquely, to an identity.

**Indirectly identifying variables (quasi-identifiers).** The quasi-identifiers are the background knowledge variables about individuals in the data set that an adversary can use, individually or in combination, to probabilistically re-identify a record. If an adversary does not have background knowledge of a variable then it cannot be a quasi-identifier. The manner in which an adversary can obtain such background knowledge will determine which attacks on a data set are plausible. For example, the background knowledge may be available because the adversary knows a particular target individual in the data set, an individual in the data set has a visible characteristic that is also described in the data set, or the background knowledge exists in a public or semi-public registry.

Examples of quasi-identifiers include sex, date of birth or age, locations (such as postal codes, census geography, information about proximity to known or unique landmarks), language spoken at home, ethnic origin, aboriginal identity, total years of schooling, marital status, criminal history, total income, visible

DRAFT

[9] Tore Dalenius, ‘Finding a Needle In a Haystack or Identifying Anonymous Census Records’ (1986) 2 Journal of Official Statistics 329.

minority status, activity difficulties/reductions, profession, event dates (such as admission, discharge, procedure, death, specimen collection, visit/encounter), codes (such as diagnosis codes, procedure codes, and adverse event codes), country of birth, birth weight, and birth plurality.

**Other variables.** These are the variables that are not really useful for determining an individual's identity. For example, someone's blood pressure value at 1pm on Tuesday would not be useful for re-identification.

Individuals can be re-identified because of the directly identifying variables and the quasi-identifiers[10].

## *1.2 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 it would be considered personal information or not.

To characterize the states of data, we consider five characteristics about 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. This can be achieved through signing a contract or by more automated identity authentication schemes. The main purpose of identity verification is to be able 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. For example, the names may be removed and unique identifiers are converted to pseudonyms[11]. Another term that is used for this kind of data is "pseudonymous" or "pseudonymized" data.
3. **Application of de-identification.** De-identification techniques perturb the quasi-identifiers, such as the dates, ZIP codes, and other demographic and socio-economic data[12]. This perturbation is intended to be minimal in that it would not change the conclusions drawn from the analytics run on the data. However,

DRAFT

[10] Lawrence Cox and Gordon Sande, 'Techniques for Preserving Statistical Confidentiality', *Proceedings of the 42nd Session of the International Statistical Institute* (1978).
[11] 'Health Informatics. Pseudonymization' (ISO 2008) International Standard ISO/TS 25237:2008.
[12] Khaled El Emam, *Guide to the De-Identification of Personal Health Information* (CRC Press (Auerbach) 2013).

it can also range from low perturbation (e.g. a date of birth converted to month and year of birth) to high perturbation (.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 is considered a strong control[13]. Other requirements in the contract may include: prohibition on attempting to contact data subjects, limits or prohibitions on linking the data without prior approval of the data custodian, and limits on further sharing the data without prior approval of the data custodian. The addition of audit requirements (e.g., to verify that the data recipient implements all of the controls in the point below, and these can be third party audits for example) would further strengthen the contractual controls.

5. **Security and Privacy Controls.** Such controls can vary in their strength[14], but they reduce the likelihood of a rogue employee at the data recipient attempting a re-identification and reduces the likelihood of a data breach occurring[15]. Standard security and privacy practices would make up this set.

These characteristics are not independent. For example, it is not possible to have contractual controls unless the data recipient can be identified, and there is no mechanism to enforce security and privacy controls unless there are also contractual controls.

Using these characteristics, we can 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.

DRAFT

[13] 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.

[14] HITRUST Alliance, 'HITRUST Common Security Framework'.

[15] 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.

| | | 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 | HIGH |
| | "Vanilla" Pseudonymized Data | YES | YES | NONE | NO | NONE |
| | Personal Data | YES | NO | NONE | NONE | NONE |

**Table 1:** The current six states of data.

One of the factors that determines whether data is personally identifying information or not is whether a data set is considered anonymized or not. As shown in Table 1, data where the direct identifiers have been removed or pseudonymized (i.e., masked) and the indirect identifiers left intact is considered pseudonymous. All known successful re-identification attacks have been performed on pseudonymous data[16]. In the EU, the Article 29 Working Party had made clear that pseudonymous data is considered personal information[17]. Under the US HIPAA, the limited data set is effectively pseudonymous data and it is still considered protected health information[18]. Therefore, all variants of pseudonymous data are considered personal information.

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 de-identification. Because it is public data, there are no contractual controls, and consequently there are no security and privacy controls. When the masking and de-identification are performed properly, the probability of re-identifying individuals in such data is very small, and therefore it is not considered personally identifiable information (PII). For example, the European Medicines Agency has mandated that drug manufacturers who submit

DRAFT

[16] K 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>.

[17] 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.

[18] 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>.

their drugs for approval to the agency must also submit an anonymized version of the clinical study information which will be made public[19] Even though there are terms of use that data users need to agree to, the agency has said it will not enforce these, which makes this effectively public data.

**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. The terms of use would prohibit re-identification attempts, require reasonable efforts to protect the data, have a prohibition on sharing the data (so that everyone who has a copy of the data needs to register), and a prohibition on linking the data with other public or private data sets. This data set would still be masked and will still be subject to de-identification, especially to ensure that public information cannot be used to re-identify individuals in the data. A good example of this is the Heritage Health Prize data set, which was longitudinal claims data made available as part of a competition to build models to predict hospital re-admissions[20].

**Non-Public Data Release.** When data is released in a non-public manner it means that not just anyone can ask for the data—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 who have a legitimate purpose to use the data. Because strong contractual controls will be in place, specific security and privacy controls may be imposed by the data custodian. Because strong contractual, as well as security and privacy controls, are in place the amount of de-identification that would need to be applied is less than for quasi-public data and therefore the data quality that comes out of this process is higher. For example, clinical trial data is being made available by drug manufacturers through a hosted secure portal, and researchers need to sign an agreement before they get access to the data[21].

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 more. For non-public data the level of perturbation is small, and therefore the data quality can be quite high.

DRAFT

[19] European Medicines Agency. "Policy 0070: European Medicines Agency policy on publication of data for medicinal products for human use". 2014.

[20] El Emam K, Arbuckle L, Koru G, Eze B, Gaudette L, Neri E, et al. De-identification Methods for Open Health Data: The Case of the Heritage Health Prize Claims Dataset. Journal of Medical Internet Research 2012;14:e33. doi:10.2196/jmir.2001.

[21] http://www.clinicalstudydatarequest.com/

Up to this point the data can credibly be said to be not personal information, on the assumption that proper de-identification methods have been applied, and the contractual, security, and privacy controls are reasonable. The data states below are, at least under existing definitions, considered to be personal information. Data in these states would not be shared publicly, and therefore these states only pertain to non-public data:

> **Protected Pseudonymous Data.** Pseudonymous data is where only masking has been applied and no de-identification is used. *Protected* pseudonymous data has additional contractual, security, and privacy controls in place. These controls reduce the risk of re-identification considerably, but do not necessarily bring them below the threshold to be considered non-PII. Also, pseudonymous data is assumed to be non-public.
>
> **"Vanilla"Pseudonymous Data.** This is pseudonymous data without any of the additional contractual, security or privacy controls in place. This is a non-public data set and therefore the identify of the recipient is already assumed to be known. Any additional conditions in the form of a terms-of-use agreement can therefore be evaluated under the state of protected pseudonymous data.
>
> **Personal Data.** Data that has had not been modified in any way or that has been modified so little that the probability of re-identification is still very high, and still includes all direct and indirect identifiers moreorless intact, is considered personal data.

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

DRAFT

# 2 Pseudonymous Data

On-going debate in the privacy community is how to treat "protected pseudonymous data", and whether it should receive special treatment. The notion of protected pseudonymous data is already articulated in current regulations.

The EU Article 29 Working Party has advocated that for the equivalent of protected pseudonymous data as we have defined it[22], which is consistent with other interpretations by the Working Party and the disclosure control community[23]:

> *"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 it is not made clear what this additional flexibility would mean. Perhaps such flexibility would be determined at a national level.

In the proposed EU general data protection regulation (GDPR), pseudonymous data is defined as[24]:

> *"(2a) 'pseudonymous data' means personal data that cannot be attributed to a specific data subject without the use of additional information, as long as such additional information is kept separately and subject to technical and organisational measures to ensure non-attribution;"*

This definition is also 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, pseudonymous data is still considered personal information.

Under the HIPAA Privacy Rule there is the concept of a limited data set, which requires that: (a) the following 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[25]. 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.

DRAFT

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

[23] 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.

[24] European Parliament, 'Legislative Resolution of 12 March 2014 on the Proposal for a Directive of the European Parliament and of the Council on the Protection of Individuals with Regard to the Processing of Personal Data by Competent Authorities for the Purposes of Prevention, Investigation, Detection or Prosecution of Criminal Offences or the Execution of Criminal Penalties, and the Free Movement of Such Data (COM(2012)0010 – C7-0024/2012 – 2012/0010(COD))' <http://www.europarl.europa.eu/sides/getDoc.do?pubRef=-//EP//TEXT+TA+P7-TA-2014-0219+0+DOC+XML+V0//EN>.

[25] 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).

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.

Limited data sets are effectively pseudonymous data, with dates, location, as well as other quasi-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[26].

Therefore, pseudonymous data with specific conditions imposed on it, such as contractual as well as security and privacy controls, are still considered personal information. These conditions have not been specified precisely (e.g., “use appropriate safeguards”), however as personal information there is an expectation that it be treated with the same level of controls. Furthermore, pseudonymous data are also expected to receive special or “flexible” treatment[27], although this is not specified precisely either.

# 3 Additional Conditions on Protected Pseudonymous Data

If pseudonymous data with additional protections is still considered personal information, then that does not create strong incentives for data custodians to invest in these protections. If, however, protected pseudonymous data can be shared without express consent of the individual data subjects[28], then that is a significant simplification for data custodians.

DRAFT

[26] ibid.

[27] 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.

[28] Although there is an exception in that for LDS there is no requirement to obtain consent.

Therefore, we propose three precise conditions that would allow one to make a credible claim that the risk of re-identification is very small for protected pseudonymized data so that it can be used and disclosed for secondary purposes without consent.

## *3.1 Adding Flexibility to the Processing of Protected Pseudonymized Data*

In this section, we define very specific criteria that would reduce the risk of re-identification for protected pseudonymous data to a very small value.[29]

### *3.1.1 No Processing by Humans*

We can utilize an existing risk assessment framework[30]. For non-public data there are three types of attacks that can be launched by an adversary:

> **Deliberate attack.** This is when the data recipient deliberately attempts to re-identify data subjects. This can be a corporate decision or an act by a rogue employee. The contractual controls mitigate against the former, and demonstrably strong security and privacy controls mitigate against the latter. Therefore, protected pseudonymized data would have a low probability for this type of attack.

> **Breach.** This is when there is a data breach and the data ends up in the “wild”. Strong security and privacy controls would mitigate against that type of risk materializing. Therefore, protected pseudonymized data would have a low probability for this type of attack.

DRAFT

[29] 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.

[30] 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); ‘HITRUST De-Identification Framework’ (*Hitrust*) <https://hitrustalliance.net/de-identification/>.

> **Inadvertent re-identification.** This occurs when a data analyst spontaneously or 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 in a geography where they are very unlikely to know a data subject. With respect to the latter, if all the data analysts are based in India, say, and all of the data subjects are from Minneapolis, then the likelihood that an analyst will know someone in the data will be small. Although this logic falls apart for famous data subjects where the analyst in India may still know the data subject in the U.S. data.

Therefore, if protected pseudonymized data is processed only by machine and not by individual analysts then the risk from the three plausible attacks can be considered to be very small.

### *3.1.2 No PII Leakage from Results*

A key criterion for sharing processed data or the results of data analysis is that these results do not leak information. There are a number of ways that analysis results can leak information, and these have been documented thoroughly elsewhere[31]. For example, information leakage has been documented from the release of tabular data[32].

We assume that the analysis results are shared broadly with no restrictions. We will refer to the consumer of these analysis results as the end-users. If the analysis is simple cross-tabulations, and the end-user is able to get multiple cross-tabulations from a data set, then it is possible for the end-user to start inferring information that can be potentially identifying. This is a well-known problem of inference from statistical databases[33].

If the automated analysis produces a model, such as a linear regression model or a logistic regression model, then under certain conditions these models can reveal information that can potentially identify individuals in the data set[34].

DRAFT

[31] 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.

[32] Muralidhar K, Sarathy R. Privacy Violations in Accountability Data Released to the Public by State Educational Agencies. Federal Committee on Statistical Methodology Research Conference, 2009; Algranati D, Kadane J. Extracting Confidential Information from Public Documents: The 2000 Department of Justice Report on the Federal Use of the Death Penalty in the United States. Journal of Official Statistics 2004;20:97–113.

[33] Subcommittee on Disclosure Limitation Methodology (n 10).

[34] 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>.

Therefore, it is important to ensure that the analysis performed on the data in an automated fashion does not produce results that would leak identifying information. In practice, this is a challenging problem that has not been solved in general (i.e., for all types of statistical analyses). However, there are techniques that can be applied for specific types of analysis to avoid the leakage of such identifying information[35].

### *3.1.3 No Sensitive Data*

Another condition is that the data itself is not considered sensitive. We suggest that there are two types of sensitive information: first, the type of information which may trigger a subjective type of harm (usually upon personal data being disclosed), and a second type which may trigger an objective type of harm (usually upon personal data being used).

#### *3.1.3.1 Information triggering Subjective Harm*

A first type of harm, which is a more subjective type of harm in the sense that it is associated with feelings of humiliation or embarrassment, would most likely take place upon personal information being disclosed.

Warren and Brandeis in their famous article about privacy and the right to be let alone, referred to the disclosure of private facts in new press, contending that privacy involved "injury to the feelings"[36]. 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"[37]. 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 [38]. Observation may include the activity of collecting personal information but this also includes the disclosure of personal information [39]. Calo suggests that many of the harms we associate with a person seeing us, such as "embarrassment, chilling effects or a loss of solitude", flow from the mere belief that one is being observed [40]. Gavison refers to an observation with an "inhibitive effect on most individuals that makes them more formal and uneasy"[41]. Recently, in *Jones v. Tsige* [42], the Court of Appeal for Ontario hinted that there was a subjective component to an invasion of privacy, assimilated to "distress,

DRAFT

[35] Christine M O'Keefe and Donald B Rubin, 'Individual Privacy Versus Public Good: Protecting Confidentiality in Health Research' [2015] Statistics in Medicine n/a.
[36] Warren and Brandeis, 'The Right to Privacy' (1890) IV Harvard Law Review <http://groups.csail.mit.edu/mac/classes/6.805/articles/privacy/Privacy_brand_warr2.html>.
[37] William Prosser, 'Privacy' (1960) 48 California Law Review 383.
[38] R Calo, 'The Boundaries of Privacy Harm' (2011) 86 Indiana Law Journal.
[39] ibid.
[40] MR Calo, 'People Can Be So Fake: A New Dimension to Privacy and Technology Scholarship' (2010) 114 Penn State Law Review 809.
[41] Ruth E Gavison, 'Privacy and the Limits of Law' (1980) 89 The Yale Law Journal 421.
[42] *Jones v Tsige* (ONCA).

humiliation or anguish” (which is therefore subjective in nature) [43]. Information may be sensitive if, a disclosure of this information may result in some type of humiliation or embarrassment.

In order to be harmful to an individual, 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” and which may be essential to the identification of this kind of harm: These additional criteria are the “**intimate**” nature of the information, and the extent of its “**availability**” to third parties or the public upon being disclosed.[44]“Intimate Nature” of the Information

To trigger the feeling of humiliation or embarrassment upon being disclosed, the data usually needs to focus on something of an “intimate nature”[45]. As mentioned earlier, Warren and Brandeis were specifically concerned with protecting information about “the private life, habits, acts, and relations of an individual”[46] and Prosser discussed the presence of a tort of privacy invasion in cases where there had been a “[p]ublic disclosure of embarrassing private facts.”[47] In the late 1960s, the conclusions of the Nordic Conference on the Right of Privacy (1967) referred to the kind of harm resulting from an attack on the honour and reputation of an individual and the “disclosure of irrelevant embarrassing facts relating to his private life”[48]. In Europe, Resolution 428 (1970) *containing a declaration on mass communication media and human rights* suggested that the right to privacy was the protection of one’s “private, family and home life” which consisted, among other things, of the “non-revelation of irrelevant and embarrassing facts, unauthorized publication of private photographs, (…) [and the] protection from disclosure of information given or received by the individual confidentially.”[49]

DRAFT

[43] ibid.

[44] Gratton, *Understanding Personal Information: Managing Privacy Risks* (LexisNexis 2013) <http://store.lexisnexis.ca/store/ca/catalog/booktemplate/productdetail.jsp?pageName=relatedProducts&catId=cacat_70_en&prodId=prd-cad-6116>, 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.

[45] Pierre Trudel and Karim Benyekhlef, ‘Approches et stratégies pour améliorer la protection de la vie privée dans le contexte des inforoutes’ (CAI 1997) <https://papyrus.bib.umontreal.ca/xmlui/handle/1866/71>.

[46] Warren and Brandeis (n 27).

[47] Prosser (n 28).

[48] Kenneth Younger, ‘Report of the Committee on Privacy, Appendix K: “Definitions of Privacy”.’ : “2. (…) The right of the individual to lead his own life protected against (…) the disclosure of irrelevant embarrassing fact relating to his private life (…).”

[49] Council of Europe Parliamentary Assembly, ‘RESOLUTION 428 Containing a Declaration on Mass Communication Media and Human Rights’ <http://assembly.coe.int/main.asp?Link=/documents/adoptedtext/ta70/eres428.htm>.:“The right to privacy consists essentially in the right to live one's own life with a minimum of interference. It concerns (…) non-revelation of irrelevant and embarrassing facts (…).”

In Europe, Directive 95/46/EC has included at article 8 categories of “sensitive” data, and acknowledge that certain types of personal data are more privacy sensitive and more likely to harm the data subject in cases of unauthorized processing[50]. These categories 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”. The categories of so-called inherently “sensitive” information are usually of an “intimate” nature[51]. Interestingly, these categories are similar to the categories or elements determined by Canadian courts as relating to the intimate or the private life of individuals[52]. While the usual metric in Canadian DPLs (under the *Personal Information Protection and Electronic Documents Act* (PIPEDA) and similar provincial laws from B.C., Alberta and Quebec) to establish whether certain information is protected is the notion of an “identifiable individual”, courts (in Canada and even in the U.S.) have adopted a rather different threshold in the context of the “reasonable expectation of privacy”. In *R. v. Plant*,[53] Sopinka J. of the Supreme Court of Canada establishes the framework for evaluating informational privacy claims. According to Sopinka, a reasonable expectation of privacy depends on whether the information in question reveals “a biographical core of personal information (…) [that] (…) would include information which tends to reveal intimate details of the lifestyle and personal choices of the individual”[54].

In Canada, in *Stevens v. SNF Maritime Metal Inc*.[55], the Federal Court of Canada took the position that the individual had not put into evidence the fact that his personal information disclosed in breach of PIPEDA triggered a subjective harm, since the information at stake was not “deeply personal” or “intimate”. In the recent case of *Jones v. Tsige*[56], the Court of Appeal for Ontario illustrates that in the case of an invasion of privacy, the fact that the information disclosed is of “intimate” nature is crucial:

DRAFT

[50] 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.

[51] Gratton, *Understanding Personal Information: Managing Privacy Risks* (LexisNexis 2013) <http://store.lexisnexis.ca/store/ca/catalog/booktemplate/productdetail.jsp?pageName=relatedProducts&catId=cacat_70_en&prodId=prd-cad-6116>, section entitled “Information that Is Inherently Intimate” at 291 (and following), which elaborates on the fact that information of *intimate* nature would include medical and health information, information pertaining to one’s family and personal life, information pertaining to love, sex and sexual orientation, religion, political and philosophical opinions, race and ethnicity, personal affiliations, financial information, private communications and location information.

[52] Pierre Trudel, ‘Privacy Protection on the Internet: Risk Management and Networked Normativity’ in Prof Serge Gutwirth and others (eds), *Reinventing Data Protection?* (Springer 2009) <http://link.springer.com/chapter/10.1007/978-1-4020-9498-9_19>; Personal Information and Electronic Documents Act (PIPEDA) 2000 c.5. In Canada, PIPEDA suggests that the form of the consent sought by the organization may vary, depending upon the circumstances and the type of information, and that in determining the preferred form of consent, organizations shall take into account the sensitivity of information. In principle 4.3.4. it goes on to say that any information can be sensitive depending on the context, and provides the following example: “For example, the names and addresses of subscribers to a newsmagazine would generally not be considered sensitive information. However, the names and addresses of subscribers to some special-interest magazines might be considered sensitive.” It is interesting to note that when referring to the “special-interest” magazine, PIPEDA is probably implying reference to information of an “intimate” nature.

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

[54] At p. 16. Under section 8 of the *Canadian Charter of Rights and Freedoms*, information is therefore only worthy of constitutional protection if it forms part of a “biographical core” of intimate details or lifestyle choices.

[55] *Stevens v SNF Maritime Metal Inc* 1137 (FC).

[56] *Jones v. Tsige* (n 33).

> *"A claim for intrusion upon seclusion will arise only for deliberate and significant invasions of personal privacy. (…) it is only intrusions into matters such as one's financial or health records, sexual practices and orientations, employment, diary or private correspondence that, viewed objectively on the reasonable person standard, can be described as highly offensive."*[57]

In the U.S., there is no general DPL overseeing all commercial activities of organizations (such as there are in Canada and France) although so-called "sensitive" information is accorded special recognition through a series of sectoral privacy statutes. More specifically, the particular categories of information most likely to require protection against disclosure to third parties are often information which would be considered as being of "intimate" nature: government records [58], cable company records[59] (i.e. personal communications), video rental records [60], and personal health information [61]. Various U.S. states would also restrict the disclosure of particular forms of information, such as medical data and alcohol and drug abuse [62]. Ohm in his recent article suggests that the types of inherently sensitive information would include, on top of information pertaining to health, sex, financial and political opinions, information such as criminal records, education, geolocation and metadata[63] which is information which may disclose information of intimate nature triggering subjective harm.

#### *3.1.3.2 "Availability" of the information*

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)[64].

Some are claiming that changes with regards to how individuals view their privacy have recently taken place and contend that the social changes inherent to web 2.0 (with individuals voluntarily sharing their personal information) may perhaps reflect a changing mentality with regards to privacy [65]. As early as 1970, Resolution 428 *containing*

DRAFT

[57] ibid.

[58] US Congress, 'The Privacy Act of 1974; 5 U.S. Code § 552a(e)(10)' <http://www.law.cornell.edu/uscode/text/5/552a>.

[59] US Congress, 'The Cable Communications Policy Act of 1984; 47 U.S. Code § 551 (b)-(c) - Protection of Subscriber Privacy' <http://www.law.cornell.edu/uscode/text/47/551>.

[60] 'The Video Privacy Protection Act of 1988, 18 U.S.C. § 2710(b)(1) - Wrongful Disclosure of Video Tape Rental or Sale Records' <http://www.law.cornell.edu/uscode/text/18/2710>.

[61] US Congress, 'The Health Insurance Portability and Accountability Act of 1996; 42 U.S. Code § 1320d - Definitions' <http://www.law.cornell.edu/uscode/text/42/1320d>.

[62] State of California, 'The California Health and Safety Code § 199.21 (West 1990) (repealed 1995)'; 'New York Public Health - Title 2 - § 17 Release of Medical Records' <http://law.onecle.com/new-york/public-health/PBH017_17.html> accessed 23 February 2015; State Government, '71 PA. STAT. ANN. § 1690.108 (West 1990) - Confidentiality of Records' <https://govt.westlaw.com/pac/Document/N6E0D5A40343911DA8A989F4EECDB8638?viewType=FullText&originationContext=documenttoc&transitionType=CategoryPageItem&contextData=%28sc.Default%29> accessed 23 February 2015.

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

[64] 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.

[65] N Robinson and others, 'Review of the European Data Protection Directive' (RAND Corporation 2009) at 15: "(..) for example individuals willing to give up personal information for small gains such as by telling personal stories to become part of a trusted community of shared interests, and sharing content increasingly via userfriendly and accessible platforms such as YouTube and SNS". L Gordon Crovitz, 'Privacy

*a declaration on mass communication media and human rights* suggested that individuals who "by their own actions, have encouraged indiscreet revelations about which they complain later on, cannot avail themselves of the right to privacy."[66] Certain European jurisdictions (such as France) provide for certain exclusions (no consent is required) for personal data rendered public by the individual concerned[67]. This provision implicitly acknowledges the fact that the disclosure of personal information, that was already made available or rendered public by the individual, may not be as harmful as the disclosure of information that has remained confidential.

In the Information Age, with new technologies and the web, most information that is disclosed may have been previously available to a certain extent. Instead of data being "disclosed", we can therefore speak of data being "increasingly available". Solove suggests that in such a situation "One must focus on the extent to which the information is made more accessible"[68].

In a 2009 finding, the Office of the Privacy Commissioner of Canada allowed enrichment of phone book information with demographic information from Statistics Canada and refused to impute a consent requirement[69]. The PIAC has shared its concern with this decision involving "publicly available" information enriched by data aggregators and data miners[70]. In France, a different outcome took place in a similar situation, when publicly available directory data was to be merged with other available information. France's Data Protection Authority, the CNIL, announced on September 23, 2011 that it had found the French provider of universal telephone directory services, *Pages Jaunes,* guilty of violating several provisions of the French DPL[71]. Pages Jaunes' web crawler function captured information contained on Facebook, Twitter and LinkedIn profiles of individuals having the same name as the individual being looked up in the directory service and "more complete profiles" were made

DRAFT

Isn't Everything on the Web' *Wall Street Journal* (24 May 2010) <http://www.wsj.com/articles/SB10001424052748704546304575260470054326304> accessed 23 February 2015; Stan Karas, 'Privacy, Identity, Databases: Toward a New Conception of the Consumer Privacy Discourse' (2002) 52 American University Law Review. Stan Karas articulates the view that "if people tend to treat different kinds of private information differently, perhaps should the law."

[66] Council of Europe Parliamentary Assembly (n 39) 428.

[67] 'Loi Informatique et Liberté, C. II, S. 2, Art. 8 (II) (4)' <http://www.cnil.fr/documentation/textes-fondateurs/loi78-17/#CHAPITRE2>.

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

[69] Office of the Privacy Commissioner of Canada, 'Commissioner's Findings - PIPEDA Case Summary #2009-004: No Consent Required for Using Publicly Available Personal Information Matched with Geographically Specific Demographic Statistics (January 9, 2009)' (8 June 2009) <https://www.priv.gc.ca/cf-dc/2009/2009_004_0109_e.asp>.

[70] Public Interest Advocacy Centre, '2010 Consumer Privacy Consultations: Comments of PIAC on Behavioural Targeting' (15 March 2010) <http://www.piac.ca/privacy/piac_comments_to_privacy_commissioner_of_canada_on_behavioural_targeting>."However, the Office of the Privacy Commissioner of Canada, in bestowing the title of "publicly available" upon this type of personal information (directory information) and then refusing to require consent for the new use the information after its "enrichment" with yet more personal information simply guts PIPEDA Principle 4.5. It ignores the general safeguards that the CRTC sought to uphold over the years in many decisions on directories. It allows an entire industry to be constructed with the express purpose of doing indirectly what PIPEDA forbids directly." [footnotes omitted].

[71] CNIL, 'Carton Rouge Pour Les Pages Jaunes' (23 September 2011) <http://www.cnil.fr/linstitution/actualite/article/article/carton-rouge-pour-les-pages-jaunes/>. The CNIL did not fine *Pages Jaunes*, but published a detailed warning, listing each privacy violation that the CNIL had identified during its investigation of *Pages Jaunes's* activities.

available online without the requisite consent.[72] The CNIL's decision illustrates the concerns which can take place with the "availability" criteria. More specifically, an organization, prior to disclosing information, must assess if the data to be disclosed has been mined, analyzed and whether the disclosure of the information will release additional information or increase the "knowledge" with regards to the individual concerned or disclose something new about the individual.

While many privacy regulators may be hesitant to apply and take into account this "availability" criteria given that DPLs regulate all "personal information", in many cases, even if this information is publicly available,[73] this criteria may still be useful. For instance, it may be used by courts to determine the extent of a privacy harm to an individual following a use or disclosure of his or her personal information. Moreover, it may be use by an organization to determine the risk behind a data handling activity, in the sense that an individual would most likely be less concerned with the disclosure of his or her personal information if such information was already widely available to third parties.[74] Certain case law rendered also confirm that information disclosed will create a *risk of harm* only if it is not already available or known to the individuals to which it is disclosed. In the U.S., if an intimate fact about a person is known to others, many courts conclude that it is no longer private (and concurrently that there is no harm in disclosing it or making it available). This was the U.S. case in *Sipple v. Chronicle Publishing Co.*[75] where 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.[76]

DRAFT

#### *3.1.3.3 Information triggering 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

[72] For example, if someone were to look up the telephone number of Eloïse Gratton, Pages Jaunes would show Gratton's phone number, and would also show information on social media sites relating to individuals named Eloïse Gratton. The information displayed included photos, the name of employer, schools attended, geographic location, profession, etc.

[73] Gratton, *Understanding Personal Information: Managing Privacy Risks* (LexisNexis 2013) <http://store.lexisnexis.ca/store/ca/catalog/booktemplate/productdetail.jsp?pageName=relatedProducts&catId=cacat_70_en&prodId=prd-cad-6116>, section entitled "publicly available information" at 302 (and following).

[74] Gratton, *Understanding Personal Information: Managing Privacy Risks* (LexisNexis 2013) <http://store.lexisnexis.ca/store/ca/catalog/booktemplate/productdetail.jsp?pageName=relatedProducts&catId=cacat_70_en&prodId=prd-cad-6116>, section entitled "availability of information" at 300 (and following).

[75] 201 Cal. Rptr. 665 (Ct. App. 1984) at 666 [*Sipple*].

[76] *Ibid*. at 669: "[P]rior to the publication of the newspaper articles in question [Sipple]'s homosexual orientation and participation in gay community activities had been known by hundereds of people in a variety of cities (…)."

for the individual behind the information.[77] This second type of harm would most likely take place upon personal information being used. Reidenberg aptly observes:

> *"(…) the creation of special protection is also understood as requiring attention not only to whether information identifies particular aspects of a person's life that are sensitive, but how data will actually be used."[78]*

Calo explains that while at the collection or disclosure levels, the corresponding harm may be subjective in nature[79], the consequence of a third party using data would be much more concrete and in many cases, would have financial implications[80]. According to Calo, the objective category of privacy harm would be the unanticipated or forced use of personal information against a given person:

> *"The second category is "objective" in the sense of being external to the person harmed. This set of harms involves the forced or unanticipated use of information about a person against that person. Objective privacy harms can occur when personal information is used to justify an adverse action against a person, as when the government leverages data mining of sensitive personal information to block a citizen from air travel, or a neighbor forms a negative judgment from gossip. They can also occur when such information is used to commit a crime, such as identity theft or murder."* [81]

It is often the use of information that leads to a more tangible kind of harm. For example, if the criminal record of a bank employee is disclosed to his co-workers, this employee may feel embarrassed and humiliated (subjective harm resulting from the disclosure). Once the information is then used by the bank to dismiss the employee, the resulting harm will be objective in nature (in this case, a financial or economical harm).

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. In 1973, the *Report of the Secretary's Advisory Committee on Automated Personal Data Systems* (U.S.) mentioned that

DRAFT

[77] This particular activity (or the term "using") is not defined in the Canadian or French DPLs analyzed. In Europe, the activity of "processing" the information includes the "use" of personal information. As a matter of fact, EC, Directive 95/46/EC defines "processing of personal data" as "any operation or set of operations which is performed upon personal data, whether or not by automatic means, such as collection, recording, organization, storage, adaptation or alteration, retrieval, consultation, use, disclosure by transmission, dissemination or otherwise making available, alignment or combination, blocking, erasure or destruction."

[78] Joel R. Reidenberg and Paul M. Schwartz, 'Data Protection Law and Online Services: Regulatory Responses' <http://ec.europa.eu/justice/data-protection/document/studies/files/19981201_dp_law_online_regulatory_en.pdf>.

[79] Calo (n 15) at 20. "Subjective privacy harms are injuries individuals experience from being observed. But why does the belief that one is being observed cause discomfort or apprehension? In some instances, the response seems to be reflexive or physical. The presence of another person, real or imagined, creates a state of 'psychological arousal' that can be harmful if excessive and unwanted."

[80] *TJX Companies Retail SEC Breach Litigation, 564 F3d 489* Court of Appeals for the First Circuit 07-2828; Calo (n 29). For example, when TJX was hit with a security breach, its customers were worried about a potentially costly identity theft. In January 2007, TJX Companies, Inc. ('TJX'), a major operator of discount stores, revealed that its computer systems had been hacked and that credit or debit card data for millions of its customers had been stolen.

[81] Calo (n 29).

privacy was directly affected by the kind of “uses” made of personal information[82]. 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”[83], and that “outdated data” should be discarded especially when “used for making decisions which affect the data subject”[84].

Solove argues that the use of personal information in databases presents a different set of problems than does government surveillance[85] and, therefore, the Big Brother metaphor fails to capture the most important dimension of the database problem[86]. He uses the metaphor of Franz Kafka’s *The Trial*, to illustrate the problem (or the harm) resulting from databases and the activity of “using” personal information[87]. In *The Trial*, an unscrupulous bureaucracy uses personal information to take important decisions, while denying the relevant people the ability to participate in how their information is being used. Solove states that this problem is derived from information processing (which he defines as the storage, use and analysis of data) rather than simply information collection[88]. According to him, this sort of information processing (or use of information) would affect power relationships between people and the institutions of the modern state. The individual would be frustrated by a “sense of helplessness” and “powerlessness”. Social structure would also be affected by altering the kinds of relationships people have with the institutions that make important decisions about their lives[89].

A broad range of harms can be inflicted on data subjects emerging out of the use of their personal information. Van den Hoeven believes that the first type of moral reason for thinking about constraining the flow of personal information is concerned with the prevention of information-based harm, which includes financial harm such as theft or identity fraud[90]. When discussing the type of harm that may result from the use of personal information,

DRAFT

[82] ‘Report of the Secretary’s Advisory Committee on Automated Personal Data Systems’ <https://epic.org/privacy/hew1973report/>.
[83] 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).
[84] ibid.
[85] D Solove, ‘“I’ve Got Nothing to Hide” and Other Misunderstandings of Privacy’ (2007) 44 San Diego Law Review 745.
[86] Daniel J Solove, ‘Privacy and Power: Computer Databases and Metaphors for Information Privacy’ [2001] Stanford Law Review at 1399: “Databases alter the way the bureaucratic process makes decisions and judgments affecting our lives; and they exacerbate and transform existing imbalances in power within our relationships with bureaucratic institutions. This is the central dimension of the database privacy problem, and it is best understood with the Kafka metaphor.”
[87] ibid. at 1429: “In sum, the privacy problem created by the use of databases stems from an often careless and unconcerned bureaucratic process—one that has little judgment or accountability—and is driven by ends other than the protection of people’s dignity. We are not heading toward a world of Big Brother or one composed of Little Brothers, but toward a more mindless process—of bureaucratic indifference, arbitrary errors, and dehumanization—a world that is beginning to resemble Kafka’s vision in *The Trial*.”
[88] Solove, ‘A Taxonomy of Privacy’ (n 55).
[89] Solove, ‘Privacy and Power’ (n 69).
[90] Jeroen Van Den Hoven, ‘Information Technology, Privacy, and the Protection of Personal Data’, *Information Technology and Moral Philosophy* (Cambridge University Press 2008) <http://dx.doi.org/10.1017/CBO9780511498725.016>. at 311: “In an information society, there is a new vulnerability to harm done on the basis of personal data – theft, identity fraud, or straightforward harm on the basis of identifying information. Constraining the freedom to access information of persons who could cause, threaten to cause, or are likely to cause information-based harm can be justified on the basis of Mill’s Harm Principle. Protecting identifying information, instead of leaving it in

RAND Corporation (U.K., 2009) also refers to an economic harm such as “financial damages suffered as a consequence of identity theft, loss of earnings.”[91] The Canadian breach notification guidelines and provisions discuss the fact that individuals should be notified in case of a security breach triggering a loss of employment, business or professional opportunities, financial loss, identity theft, negative effects on the credit record and damage to or loss of property[92]. Theft is another type of economic harm which may take place upon the use of personal information by thieves (e.g. home address, whereabouts of the home owner)[93]. Ohm in his recent article suggests that the types of inherently sensitive information would include information pertaining to public records[94].

The second type of harm is one that van den Hoeven refers to as “Information Inequality” (or discrimination)[95]. According to him, this type of moral reason to justify constraints on our actions with identity-relevant information is concerned with *equality and fairness*. As early as the 1970s, misuse of data and the resulting discrimination was of paramount importance; evidence of this can be found in the documents leading to the adoption of Convention 108[96].

Information may be used to discriminate, remove a benefit or tarnish a reputation and an individual may be subject to some type of discrimination, which could lead him to being refused for a job, refused for credit, mortgage or a loan, etc. Many have voiced their concerns about consumer profiling, as it may be a tool used to facilitate the practice of discrimination[97]. Ohm in his recent article suggests that sensitive information would include information which may be use to discriminate, such as information pertaining to criminal records[98].

DRAFT

the open, diminishes epistemic freedom of all to know, but also diminishes the likelihood that some will come to harm, analogous to the way in which restricting access to firearms diminishes both freedom and the likelihood that people will get shot in the street. In information societies, identity-relevant information resembles guns and ammunition. Preventing information-based harm clearly provides us with a strong moral reason to limit the access to personal data.”

[91] Robinson and others (n 52).

[92] Service Alberta, ‘PIPA Information Sheet 11 Notification of a Security Breach - infosheet11.pdf’ (2010) <http://servicealberta.ca/pipa/documents/infosheet11.pdf>; Industry Canada Government of Canada, ‘The Safeguarding Canadians’ Personal Information Act’ <https://www.ic.gc.ca/eic/site/ecic-ceac.nsf/eng/gv00571.html>. See also S-4, the recent *Digital Privacy Act* amending PIPEDA.

[93] Office of the Privacy Commissioner of Canada, ‘Key Steps for Organizations in Responding to Privacy Breaches’ <http://www.privcom.gc.ca/information/guide/2007/gl_070801_02_e.pdf>. “What is the context of the personal information involved? For example, a list of customers on a newspaper carrier’s route may not be sensitive. However, the same information about customers who have requested service interruption while on vacation may be more sensitive.”

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

[95] MJ Van den Hoven and J Weckert, *Information Technology and Moral Philosophy* (Cambridge University Press 2008).

[96] Council of Europe, ‘Explanatory Report: Resolution (73) 22’; Council of Europe, ‘Res (74) 29 on the Protection of the Privacy of Individuals Vis-À-Vis Electronic Data Banks in the Public Sector’ (1974) <https://wcd.coe.int/ViewDoc.jsp?id=660013&Site=CM&BackColorInternet=C3C3C3&BackColorIntranet=EDB021&BackColorLogged=F5D383> at Principle 3 of Annex refer to electronic data processing that “may lead to unfair discrimination”.

[97] Public Interest Advocacy Centre (n 57); Conseil de l’Europe, Comité consultatif de la convention pour la protection des personnes à l’égard du and traitement automatisé des données à caractère personnel, ‘Rapport Sur L’application Des Principes de Protection Des Données Aux Réseaux Mondiaux de Télécommunications. L’autodétermination Informationnelle À L’ère de l’Internet : Éléments Sur La Réflexion Sur La Convention No 108 Destinés Au Travail Futur Du Comité Consultatif’ (2004).

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

With the onslaught of new Internet technologies, online profiling activities are taking on a range of different forms. One discriminatory practice taking place online is known as “adaptive pricing” or “dynamic pricing”[99]. Amazon was suspected of using such practices, using cookies to identify the visiting consumers[100]. In the U.K., the now defunct OFT had also expressed its concern over price discrimination, especially if consumers are left in the dark[101]. Hoofnagle and Smith warn that information flows can be used to eliminate certain customers[102]. They claim that financial institutions may analyze and use information that they collect about their customers in order to target them for the purchase of products and services and that the data may potentially be used to deny consumers choice or to steer them towards choices not in their best interest[103]. Classifying people in such a way that their chances of obtaining certain goods, services or employment are diminished may also illustrate this type of harm[104].

A third type of objective harm is a physical one.[105] For example, individuals may become a victim of a crime against their person, in the event that their information (home or work address) are used by criminals such as stalkers and rapists[106]. The harm in question can be severe, a perfect example is the murder of actress Rebecca Schaeffer in 1989[107]. In the U.S. case *Remsburg v. Docusearch*[108], a stalker killed a woman after obtaining her work address from a data broker. Canadian breach notification provisions include, in the definition of “significant

DRAFT

[99] Anthony Danna and Oscar H Gandy Jr, ‘All That Glitters Is Not Gold: Digging Beneath the Surface of Data Mining’ (2002) 40 Journal of Business Ethics 373. Some refer to this growing problem as first-degree price discrimination, a practice where organizations attempt to perfectly exploit the differences in price sensitivity between consumers.
[100] Conseil de l’Europe, Comité consultatif de la convention pour la protection des personnes à l’égard du and traitement automatisé des données à caractère personnel (n 79).
[101] Julia Kollewe and Richard Wray, ‘Office of Fair Trading to Probe Use of Personal Data by Online Retailers’ (*The Guardian*, 15 October 2009) <http://www.theguardian.com/business/2009/oct/15/retail-pricing-tactics-oft-investigation>. OFT is conducting two market studies into websites using behavioural data to set customized pricing, where prices are individually tailored using information collected about the user’s behaviour.
[102] Chris Jay Hoofnagle and Kerry E Smith, ‘Debunking the Commercial Profilers’ Claims: A Skeptical Analysis of the Benefits of Personal Information Flows’ (Social Science Research Network 2003) SSRN Scholarly Paper ID 504622 <http://papers.ssrn.com/abstract=504622>.
[103] ibid.
[104] Van Den Hoven (n 73); Tal Zarsky, ‘“MINE YOUR OWN BUSINESS!”: Making the Case for the Implications of the Data Mining of Personal Information in Hte Forum of Public Opinion’ (2003) 5 Yale Journal of Law and Technology <http://digitalcommons.law.yale.edu/yjolt/vol5/iss1/1>.
[105] These types of uses (physical harms), together with fraud and identity theft, are of a criminal nature, and they are governed by criminal laws. When certain objective harms resulting from the use of personal information are found to be very significant for individuals, they are often governed by laws, other than DPLs, which address these harms specifically. Still, acknowledging that certain disclosures may be harmful because criminals may use the information is relevant when assessing the risk of objective harm (or in assessing if there is a risk upon disclosing this information).
[106] Robinson and others (n 52); Van Den Hoven (n 73). at 311: “Stalkers and rapists have used the Internet and online databases to track down their victims. They could not have done what they did without access to electronic resources and without accessing some of the details of their victim’s lives.”
[107] *Margan v Niles, 250 F Supp 2d 63* (United States District Court for the Northern District of New York). It was discovered that her assailant located her home address through the records of the Department of Motor Vehicles.
[108] *Remsburg v Docusearch, Inc 816 A (2d) 1001* (NH).

harm", "physical harm"[109]. Ohm in his recent article suggests that the types of inherently sensitive information would include information pertaining to personal safety, information about children and geolocation[110].

If the information used may have a "negative impact" (objective harm) on the individual, the information may be considered as being sensitive. As a matter of fact, 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 not be governed by DPLs that there is less risk is using such information) or at least, that this information should be able to circulate without having to obtain the individual's prior consent.[111]

According to older as well as more recent documents (including DPLs), the central concern behind regulating the use of information has to do with the awareness of potentially negative impacts on the data subjects (objective harm). A number of provisions or principles lead us to this conclusion.

DPLs were to apply to information "used" in such a way which would have an impact on the individuals. For instance, recent DPLs provide that the information should be accurate, especially if it will be used in such a way which will create a negative impact on the individual. For example, PIPEDA provides that organizations should avoid that "inappropriate information (…) be used to make a decision about the individual."[112] Under the Civil code of Quebec, any person may examine and cause the rectification of a file kept on him by another person "with a view to making a decision in his regard or to informing a third person."[113] Under the B.C. DPL, an organization must make a reasonable effort to ensure that personal information collected by or on behalf of an organization is accurate and complete, "if the personal information is likely to be used by the organization to make a decision that affects the individual to whom the personal information relates."[114] The Directive 95/46/EC on the subject matter, has a similar provision: any individual is entitled to interrogate the data controller of his personal data in order to obtain information allowing him to know and to object to the reasoning involved in the automatic

DRAFT

[109] Service Alberta (n 76) at 2-3: "For example, if a women's shelter loses its client list, the possible harm might be much more significant than the possible harm if a fitness club loses its membership list."

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

[111] See generally, Gratton, *Understanding Personal Information: Managing Privacy Risks* (LexisNexis 2013) <http://store.lexisnexis.ca/store/ca/catalog/booktemplate/productdetail.jsp?pageName=relatedProducts&catId=cacat_70_en&prodId=prd-cad-6116>,

[112] Personal Information and Electronic Documents Act (PIPEDA) (n 42).

[113] 'Art. 38', *Civil Code of Quebec* (1991).

[114] 'Personal Information Protection Act (British Columbia), S.B.C. 2003' <http://www.bclaws.ca/Recon/document/ID/freeside/00_03063_01>.

processing, “in the case of a decision taken based on automatic processing and producing legal effects in relation to the data subject”[115].

These provisions were clearly meant to ensure that when personal information is used in assessments or decisions that may have a negative impact on an individual (what we refer to as sensitive information in the sense that its use may trigger an objective harm), the data in question should at least be accurate. It is interesting to note that under the Directive 95/46/EC, the decision has to either produce “legal effects” for, or “significantly affects”, an individual[116]. This means that there is an argument to be made that perhaps an organization using personal information which triggers a non significant impact for an individual should not be regulated in all instances by DPLs and a positive impact, even less.

The purpose of DPLs regulating the activity of using personal information was not to address situations or uses having a positive impact for the individual, as illustrated by van den Hoven:

> *“They do not mind if their library search data are used to provide them with better library services, but they do mind if these data are used to criticize their taste and character. They would also object to these informational cross-contamination when they would benefit from them, as when the librarian would advise them a book on low-fat meals on the basis of knowledge of their medical records and cholesterol values, or when a doctor asks questions on the basis of the information that one has borrowed a book from the public library about AIDS.”*[117]

## 3.2 Summary of Conditions

Therefore, one can define another type of protected pseudonymous data that has conditions on it. 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. The PII is pseudonymized by the data custodian and transferred securely to a data recipient who does automated processing on transient data, and then the data is destroyed.

DRAFT

[115] 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 (n 40).
[116] ibid 46.
[117] Van Den Hoven (n 73).

2. The analysis results do not reveal information which can lead to inferences of individual identity. There is a large body of work that describes methods of statistical disclosure control to protect against such inferences.

3. The data is non-sensitive according to the previous discussions. That is, there are no subjective or objective types of harm.

This type of data could arguably be treated more flexibly. Flexibility in this case means that it can be used for a secondary purpose without having to obtain consent. This does not mean that the data is not personal information per se – but there is added flexibility in its processing.

However, if there is a data breach then all of the protections (security and privacy controls, as well as contractual controls) are no longer applicable, and all of the quasi-identifiers would be intact in the data. As a result, breached pseudonymous data would still 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 2.

DRAFT

| | | **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 | HIGH |
| | "Vanilla" Pseudonymized Data | YES | YES | NONE | NO | NONE |
| | Personal Data | YES | NO | NONE | NONE | NONE |

**Table 2:** The Seven states of data defined.