Big Data and Innovation, Setting the Record Straight: De-identification *Does* Work

Information and Privacy Commissioner of Ontario

Senior Analyst, Information Technology and Innovation foundation

June 16, 2014

ACKNOWLEDGEMENTS

The co-authors would like to acknowledge the contributions of the following individuals to the development of this paper: Michelle Chibba, Director of Policy & Special Projects, IPC; David Weinkauf, Policy & Information Technology Officer, IPC; Robert Atkinson, President, ITIF; and Jordan Misra, Policy Intern, Center for Data Innovation.

Information and Privacy Commissioner
Ontario, Canada

2 Bloor Street East
Suite 1400
Toronto, Ontario
M4W 1A8
Canada

416-326-3333
1-800-387-0073
Fax: 416-325-9195
TTY (Teletypewriter): 416-325-7539
Website: www.ipc.on.ca
Privacy by Design: www.privacybydesign.ca

Big Data and Innovation, Setting the Record Straight: De-identification *Does* Work

TABLE OF CONTENTS

Introduction

In the coming years, analytics will offer an enormous opportunity to generate economic and social value from data. But much of the success of data analytics will depend on the ability to ensure that individuals' privacy is respected. One of the most effective ways in which to do this is through strong "de-identification" of the data — in essence, storing and sharing the data without revealing the identity of the individuals involved.

A number of researchers have been investigating techniques to re-identify de-identified datasets. Unfortunately, some commentators have misconstrued their findings to suggest that de-identification is ineffective. Contrary to what misleading headlines and pronouncements in the media almost regularly suggest, datasets containing personal information may be de-identified in a manner that minimizes the risk of re-identification, often while maintaining a high level of data quality.

Despite earlier publications illustrating the effectiveness of de-identification,[1] the myth that datasets cannot be reliably de-identified regardless of the methods employed continues to be promulgated. It is increasingly apparent that one of the reasons for the staying power of this myth is not factual inaccuracies or errors within the primary literature, but rather a tendency on the part of commentators on that literature to overstate the findings. While nothing is perfect, the risk of re-identification of individuals from properly de-identified data is significantly lower than indicated by commentators on the primary literature.

At the same time, advancements in data analytics are unlocking opportunities to use de-identified datasets in ways never before possible. Where appropriate safeguards exist, the evidence-based insights and innovations made possible through such analysis create substantial social and economic benefits. However, the continued lack of trust in de-identification and focus on re-identification risks may make data custodians less inclined to provide researchers with access to much needed information, even if it has been strongly de-identified; or worse, to believe that they should not waste their time even attempting to de-identify personal information before making it available for secondary research purposes. This could have a highly negative impact on the availability of de-identified information for potentially beneficial secondary uses.

For these reasons, it is imperative that we get the story right about de-identification. In this paper, we will discuss a select group of academic articles often referenced in support of the myth that de-identification is an ineffective tool to protect the privacy of individuals. While these articles raise important issues concerning the use of proper de-identification techniques, reported findings do not suggest that de-identification is impossible or that de-identified data should be classified as personally identifiable information. We then provide a concrete example of how data may be effectively de-identified — the case of the U.S. Heritage Health Prize. This example shows that in some cases, de-identification can maximize both privacy and data quality, thereby enabling a shift from zero-sum to positive-sum thinking — a key principle of *Privacy by Design*.

1 See, e.g., Ann Cavoukian and Khaled El Emam, "Dispelling the Myths Surrounding De-identification: Anonymization Remains a Strong Tool for Protecting Privacy," June 2011, http://www.ipc.on.ca/images/Resources/anonymization.pdf; Ann Cavoukian, "Looking Forward: De-identification Developments – New Tools, New Challenges," May 2013, http://www.ipc.on.ca/images/Resources/pbd-de-identifcation_developments.pdf.

The Risk of Re-identification Has Been Greatly Exaggerated

Some commentators routinely state as a fact that no data can be reliably de-identified, relying on the results of a few studies that allegedly make this claim. Even the recent report from the White House on big data and privacy makes this claim.[2] Therefore, it is not surprising that policymakers are now starting to believe this to be true. For example, a recent petition submitted to the U.S. Federal Communications Commission (FCC) referenced a select group of academic articles, which will be discussed below, and quoted as a "given" the "increasing ease with which datasets purged of personally identifying information can be re-identified."[3] Likewise, the title of a TechDirt article — "There's No Such Thing as an Anonymized Dataset" — is emblematic of many media stories on the topic.[4]

One recent example of primary literature whose findings have been interpreted as a basis for questioning the effectiveness of de-identification is a 2013 study on mobility data entitled "Unique in the Crowd."[5] The study found that when an individual's location is specified hourly within the reception area of a mobile phone antenna, knowing as few as four random spatio-temporal points was enough to uniquely identify 95 per cent of the mobility traces in the dataset of one and a half million individuals.[6]

However, while the authors of this study successfully demonstrated that an individual's mobility data is highly unique, they did not *actually* re-identify any individuals from their mobility traces. While they suggest that re-identification could be done by linking the mobility data to other outside information (e.g., home address, workplace address, or geo-localized tweets or pictures), they neither performed this task, nor demonstrated how it could be done. In addition to having access to the comprehensive dataset of mobility traces, an adversary would have to know at least four spatio-temporal pieces of information (e.g., the person's home address, work address, etc.) about each individual in the sample in order to re-identify 95 per cent of that population. Needless to say, amassing such information from publicly available sources would not be a trivial undertaking.

Moreover, it is important to note that only minimal efforts were made to de-identify the original dataset of mobility traces. The authors used what they refer to as a "simply anonymized dataset [that] does not contain name, home address, phone

2 For example, the authors of the report write, "When data is initially linked to an individual or device, some privacy-protective technology seeks to remove this linkage, or 'de-identify' personally identifiable information—but equally effective techniques exist to pull the pieces back together through 're-identification.'" See "Big Data: Seizing Opportunities, Preserving Values," Executive Office of the President, May 2014, http://www.whitehouse.gov/sites/default/files/docs/big_data_privacy_report_may_1_2014.pdf.

3 See Public Knowledge et al., "Petition for Declaratory Ruling Stating that the Sale of Non-Aggregate Call Records by Telecommunications Providers without Customers' Consent Violates Section 222 of the Communications Act," December 11, 2013, http://apps.fcc.gov/ecfs/document/view?id=7520963695, p. 8.

4 See Timothy Lee, "There's No Such Thing as an Anonymized Dataset," *TechDirt*, November 30, 2007, http://www.techdirt.com/articles/20071130/114005.shtml.

5 Yves-Alexandre de Montjoye, César A. Hidalgo, Michel Verleysen & Vincent D. Blondel, "Unique in the Crowd: The Privacy Bounds of Human Mobility," *Scientific Reports* 3, http://dx.doi.org/10.1038/srep01376.

6 Ibid., p. 2.

number or other obvious identifiers."[7] However, it is a well-known fact that the removal of direct identifiers alone is generally insufficient to properly de-identify datasets.[8] For this reason, while the researchers raise important points concerning the uniqueness of mobility data, it is not possible to generalize the findings of the study to the myriad types of other data that may be properly de-identified and used for research purposes. Such media claims as "it is remarkably easy to identify a mobile phone user from just a few pieces of location information"[9] and "a new study has revealed just how little information is required to determine an individual's personal identity"[10] rarely reflect the limited scope of the research.

While there is no known standard for de-identifying mobility data, there are additional techniques, such as obfuscation, that can significantly help to preserve the anonymity of location data. Mobility data can also be aggregated spatially and temporally to decrease the risk of re-identification. Despite these techniques, it is admittedly very difficult to de-identify mobility traces, while maintaining a sufficient level of data quality necessary for most secondary purposes, due to their high degree of uniqueness. In this regard, mobility data and other high-dimensional data such as genetic data are quite different from other types of low-dimensional data (e.g., date of birth, gender, home address, postal/ZIP code, etc.). The latter can be readily de-identified through processes such as aggregation or generalization, while maintaining a sufficient level of data quality necessary for secondary purposes. In the case of high-dimensional data, additional arrangements may need to be pursued, such as making the data available to researchers only under tightly restricted legal agreements.[11]

Another example of primary literature often cited to cast doubt on the effectiveness of de-identification is Latanya Sweeney's 2000 study that used 1990 U.S. census data to show that 87 per cent of the U.S. population could be uniquely identified through the combination of gender, date of birth, and ZIP code.[12] This finding means that given two hypothetical databases containing data on virtually everyone in the United States, 87 per cent of the records could be matched using just these three data points. This demonstrates the need to ensure that one strongly de-identifies these three fields, among many others. However, commentators should not simply rely on this early research as definitive evidence that datasets containing personal information cannot be effectively de-identified.

---

7 Ibid., p. 1.

8 See Article 29 Working Party, "Opinion 05/2014 on Anonymisation Techniques," April 10, 2014, http://ec.europa.eu/justice/data-protection/article-29/documentation/opinion-recommendation/files/2014/wp216_en.pdf, p. 9.

9 Jason Palmer, "Mobile location data 'present anonymity risk'," *BBC News*, March 25, 2013, http://www.bbc.com/news/science-environment-21923360.

10 Lisa Zyga, "Study shows how easy it is to determine someone's identity with cell phone data," *Phys.org*, March 25, 2013, http://phys.org/news/2013-03-easy-identity-cell.html.

11 For example, the Data for Development (D4D) Challenge, co-led by Alex "Sandy" Pentland at MIT, distributed aggregated anonymous mobility and call pattern data to researchers under a legal contract that further restricted its use and disclosure to the researchers involved and for the purposes proposed by them. See Elizabeth Bruce, "Big Data for the Rest of Us," *MIT Big Data Initiative*, May 16, 2013, http://bigdata.csail.mit.edu/node/100.

12 Latanya Sweeny, *Uniqueness of Simple Demographics in the U.S. Population*, Carnegie Mellon University, Laboratory for International Data Privacy, Pittsburgh, PA, 2000.

It is important to note that when researchers at the Palo Alto Research Center replicated Sweeney's study using more recent census data from 2000, they found that only 63 per cent of the U.S. population is uniquely identifiable given those data categories.[13] More importantly, these researchers found that the risk of unique identification drops off sharply when given slightly more abstract data. For instance, if an individual's date of birth is replaced with only the month and year of birth, the percentage of those uniquely identifiable drops to 4.2 per cent.[14] Similarly, if one further replaces the ZIP code with an individual's county, then the percentage of the population capable of being uniquely identified drops dramatically to 0.2 per cent.[15]

This simple example illustrates that different de-identification methods result in vastly different outcomes, especially when more recent efforts are compared to earlier efforts at de-identification. The more effectively the data is de-identified, the lower the percentage of individuals who are at risk of re-identification. The risk of re-identification for weakly de-identified data, such as datasets released with gender, ZIP code, and date of birth, is not the same as for strongly de-identified data. Sweeney's study contributed importantly to improving the quality of de-identification; however, articles in the media that reference it alone, while ignoring more recent de-identification standards and practices, are clearly misguided. For example, articles with headlines such as "'Anonymized' Data Really Isn't"[16] and "You're Not So Anonymous"[17] fail to appreciate the improvements in de-identification techniques that have taken place as a result of Sweeney's 2000 study.

De-identification techniques generally address three privacy risks.[18] First, they protect an individual's records from being uniquely identified in the dataset. Second, they prevent an individual's records from being linked to other datasets. If a set of attributes uniquely identifies an individual within a de-identified dataset and those same attributes are found in a personally identifying dataset, then that individual may be re-identified by linking the two datasets together. Third, they make it difficult to infer sensitive information about an individual from the de-identified dataset. For example, if groups of individuals are identified in a dataset and all the individuals in a certain group have a certain property, then if an individual is known to belong to that group, one could easily find out the value of his/her group property. Various protections reduce the risk that an individual will be re-identified and that sensitive information about the individual will be discovered. Effective de-identification methods use appropriate techniques to protect against these privacy risks, taking into consideration the specific threats and uses of the datasets in the particular context of the data processing.[19]

13 Phillippe Golle, *Revisiting the Uniqueness of Simple Demographics in the US Population*, Palo Alto Research Center, 2006, http://crypto.stanford.edu/~pgolle/papers/census.pdf.

14 Ibid., p. 2. Note Sweeney's earlier finding of 3.7 per cent in *Uniqueness of Simple Demographics*, p. 30.

15 Ibid. Note Sweeney's earlier finding of 0.04 per cent in *Uniqueness of Simple Demographics*, p. 31.

16 Nate Anderson, "'Anonymized Data Really Isn't—and Here's Why Not," *Ars Technica*, September 8, 2009, http://arstechnica.com/tech-policy/2009/09/your-secrets-live-online-in-databases-of-ruin/.

17 Caroline Perry, "You're Not So Anonymous," *Harvard Gazette*, October 18, 2011, http://news.harvard.edu/gazette/story/2011/10/you%E2%80%99re-not-so-anonymous/.

18 See Article 29 Working Party, "Opinion 05/2014 on Anonymisation Techniques," p. 11–12.

19 See also Article 29 Working Party, "Opinion 4/2007 on the concept of personal data," June 20, 2007, http://ec.europa.eu/justice/policies/privacy/docs/wpdocs/2007/wp136_en.pdf, especially example no. 13.

In addition, it is important to note that existing de-identification standards err on the side of privacy when it comes to balancing the utility of the data and the risk of re-identification. For example, in the United States, the Safe Harbor Standard under the Health Insurance Portability and Accountability Act (HIPAA) requires the modification or removal of 17 specific data elements, including replacing an individual's date of birth with year only (i.e., no month or day), prohibiting the non-aggregated disclosure of ages over 89, and restricting the disclosure of ZIP codes to the initial three digits if the resulting population is greater than 20,000 or changing it to "000" if not.[20] According to the estimates of one de-identification expert, only 0.04 per cent (4 in 10,000) of the individuals within datasets de-identified using the Safe Harbor Standard are uniquely identifiable.[21]

Even if an individual in a dataset is uniquely identifiable, that does not necessarily mean that the individual can be re-identified. A uniquely identifiable individual refers to someone who is the only person in a dataset with a given set of characteristics. To actually re-identify someone requires matching those uniquely identifiable characteristics to other data. Some commentators mistakenly assume that there are databases of the entire population available for this type of analysis. While it is certainly true that marketers, data brokers, and others maintain datasets about individuals, these datasets are often incomplete, making it difficult to positively identify someone with a high degree of confidence. Therefore it is incorrect to assume that merely because a small percentage of individuals in a dataset are uniquely identifiable, that an attacker could actually re-identify these records in practice. Re-identification is only possible if there is an alternative data source available for an attacker to use.

An example of this situation can be seen in the case of former Massachusetts Governor William Weld, whose hospital medical records were re-identified in 1996 from publicly available data about voters in Cambridge. As discussed below, this case is often cited as evidence for the claim that individuals can easily be re-identified from de-identified data, even though the technique used to re-identify Governor Weld would not work for the majority of the population. In addition, it is important to note that this case pre-dates the establishment of the HIPAA Privacy Rule in the United States, which makes this attack infeasible today. Moreover, it was instrumental in bringing about more rigorous rules and guidelines for de-identifying protected health information in order to prevent similar cases of re-identification from occurring. Thus, while it speaks to the inadequacy of certain de-identification methods employed in 1996, to cite it as evidence against current de-identification standards is highly misleading. If anything, it should be cited as evidence for the *improvement* of de-identification techniques and methods insofar as such attacks are no longer feasible under today's standards precisely because of this case.

---

20 See "Guidelines Regarding Methods for De-identification of Protected Health Information in Accordance with the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule," U.S. Department of Health & Human Services, http://www.hhs.gov/ocr/privacy/hipaa/understanding/coveredentities/De-identification/guidance.html.

21 See National Committee on Vital and Health Statistics Report to the Secretary of the U.S. Department of Health and Human Services, *Enhanced Protections for Uses of Health Data: A Stewardship Framework for 'Secondary Uses' of Electronically Collected and Transmitted Health Data*, December 19, 2007, http://www.ncvhs.hhs.gov/071221lt.pdf, p. 36.

Finally, critics of de-identification commonly cite a 2008 report by a pair of researchers who were able to re-identify Netflix users in an "anonymous" dataset by comparing the dates and ratings of movies watched by those users with similar, personally identifiable information available on the Internet Movie Database (IMDb).[22] While this study importantly exposed the risks of disclosing improperly de-identified data,[23] what articles citing it[24] often neglect to point out is that the researchers re-identified only two out of 480,189 Netflix users, or 0.0004 per cent of users, with confidence.[25] Here again, it is the statistical outliers that are most at risk of re-identification: the likelihood of re-identification goes up significantly for users who had rated a large number of unpopular movies.[26] Moreover, Netflix users who had not publicly rated movies in IMDb had no risk of re-identification. While those with an unusual taste in movies certainly deserve as much privacy as anyone, the point is that we should not assume that all non-aggregate data are automatically linkable to personally identifiable information available on the web, or that such information even exists.

The de-identification techniques used by Netflix removed "all personal information identifying individual customers" and replaced all customer IDs with "randomly-assigned IDs."[27] In addition, review dates were included in the dataset. This de-identification method does not align with any current standards. For example, under the HIPAA Safe Harbor Standard, dates are not allowed in the dataset.[28] Thus, while the researchers of the 2008 report discovered important re-identification risks associated with the released dataset, the conclusion to draw from their findings is not that de-identification as such is ineffective, but rather that a more rigorous standard should be applied to such a dataset. One must assess the quality and appropriateness of the de-identification techniques used on the dataset before weighing in on the effectiveness of de-identification.

All of the above examples of primary literature are research-based articles within the highly specialized field of computer science. An implicit assumption at work within the articles is that re-identification requires the knowledge of a highly trained, highly skilled "expert" in the field. This means that the "real-world" risk of re-identification may be far lower than expected. In the words of one judge's ruling on a case involving the release of de-identified data:

22 Arvind Narayanan and Vitaly Shmatikov, "Robust De-anonymization of Large Sparse Datasets," P*roceedings of the 2008 IEEE Symposium on Security and Privacy, 2008*, http://www.cs.utexas.edu/~shmat/shmat_oak08netflix.pdf.

23 See ibid., p. 9: "Netflix's claim that the data were perturbed does not appear to be borne out. One of the subscribers had 1 of 306 ratings altered, and the other had 5 of 229 altered."

24 See, e.g., "Differential Privacy: Motivation," Wikipedia, accessed March 27, 2014, https://en.wikipedia.org/wiki/Differential_privacy#Motivation.

25 Narayanan and Shmatikov, "Robust De-anonymization of Large Sparse Datasets," p. 13. The researchers used auxiliary data of only 80 IMDb users, so it may be difficult to draw useful conclusions from this fact alone.

26 While we refrain from commenting on broader issues like forum appropriateness and harm, it is worth mentioning that Netflix settled a class-action lawsuit regarding this data.

27 "The Netflix Prize Rules," Netflix, http://www.netflixprize.com/rules.

28 See Khaled El Emam, Elizabeth Jonker, Luk Arbuckle, Bradley Malin, "A Systematic Review of Re-Identification Attacks on Health Data," *PLoS ONE* 6 no. 12 (2011), doi:10.1371/journal.pone.0028071, p. 7.

> [T]he fact that one expert in data anonymity can manipulate the data to determine identity does not necessarily mean, without more, that a threat exists that other individuals will be able to do so as well, nor does it in any way define the magnitude of such a threat or whether that threat, if it in fact even exists, renders the release of the data an act that reasonably tends to lead to the identity of specific persons.[29]

Ultimately, this demonstrates that while research into re-identification is essential to ensuring the continued success of de-identification methods, interpreters of this literature have a tendency to jump to conclusions about the risks involved, inflating them dramatically. When all the factors are considered, de-identification remains a strong tool for protecting privacy, provided that it is employed effectively, with up-to-date tools and techniques. Indeed, as explained by Jane Yakowitz, a legal scholar critical of the types of arguments presented against de-identification, the risk of privacy harm from re-identification is actually significantly lower than many of the everyday risks we take for granted, such as those attendant on throwing out our trash.[30] Felix Wu, a professor at Benjamin N. Cardozo School of Law, claims that there is not much support for the "strongly pessimistic view" that no useful data can be de-identified.[31] He explains that "[a] closer look at the computer science … reveals that several aspects of that literature have been either misinterpreted, or at least over read, by legal scholars."[32] Identifiable data can indeed be de-identified through the use of strong de-identification techniques.[33]

In addition to the issues raised above, literature is emerging questioning the accuracy and legitimacy of well publicized re-identification attacks. For example, a recent report identifies a "fatal flaw" in re-identification attacks that use population registers, such as the one involving Massachusetts Governor William Weld, since they assume that the population register used to re-identify individuals is a complete and accurate representation of the true population.[34] In the case of Governor Weld, the voter rolls purportedly used to confirm the Governor's identity contained only approximately half of the total population of Cambridge, making it impossible to confirm with certainty that the records in question actually belonged to the Governor. Instead of using a statistically robust method to identify his records, the researchers relied on publicly-reported information about the Governor's health stemming from his status as a public figure. Not only was the actual risk

29 *Southern Illinoisan v. Department of Public Health*, No. 5-02-0836, Appellate Court of Illinois, Fifth District, June 9, 2004, http://www.state.il.us/court/Opinions/AppellateCourt/2004/5thDistrict/June/Html/5020836.htm.

30 See Jane Yakowitz, "Tragedy of the Data Commons," *Harvard Journal of Law and Technology* 25 (2011): 1–40

31 Felix T. Wu, "Defining Privacy and Utility in Data Sets," *University of Colorado Law Review* 84 (2013): 1117–1175, p. 1124.

32 Ibid.

33 For example, in the health sector, Dr. Khaled El Emam, Canada Research Chair in Electronic Health Information at the University of Ottawa, has developed a framework for de-identifying personal health information in a manner that simultaneously minimizes both the risk of re-identification and the degree of distortion to the original database. See Khaled El Emam, *Guide to the De-Identification of Personal Health Information* (Boca Raton, FL: CRC Press, 2013).

34 See Daniel C. Barth-Jones, "The 'Re-identification' of Governor William Weld's Medical Information: A Critical Re-Examination of Health Data Identification Risks and Privacy Protections, Then and Now." June 4, 2012, http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2076397.

of re-identification from voter rolls for a randomly chosen individual rather low, there was zero risk for individuals not appearing in the voter database. As the study notes, without a complete and accurate population register, such attacks are "no better than the flip of a coin" if there is only one other person with the same indirect identifiers.

Moreover, a literature review by health data researchers identified 14 published accounts of re-identification attacks on de-identified data.[35] These attacks revealed that one quarter of all records and roughly one third of health records were re-identified. However, upon further study, the researchers found that only 2 out of the 14 attacks were made on records that had been properly de-identified, using existing strong standards. Further, only 1 of the 2 attacks had been made on health data, resulting in a very low re-identification rate of 0.013 per cent.[36] So the issue is not whether de-identification works (it does), but whether it is always employed effectively (it is not).

35 El Emam et al., "A Systematic Review of Re-Identification Attacks on Health Data."

36 Ibid., p. 7.

De-identification Plays a Role in Big Data and Innovation

Advances in computing technology are unlocking opportunities to collect and analyze data in ways never before imagined.[37] The analysis of data may lead to important insights and innovations that will benefit not only individual consumers, but society at large. While data is typically collected for a single purpose, increasingly it is the many different secondary uses of the data wherein tremendous economic and social value lies. For example, recent studies have shown that large-scale mobile phone data can help city planners and engineers better understand traffic patterns and thus design road networks that will minimize congestion.[38] In many cases, the use of big data, such as for optimizing industrial systems, improving public safety, or understanding the environment, does not involve the use of personally identifiable information. However, when datasets do include personally identifiable information, organizations will need the tools to both protect privacy and enable data analytics.

De-identifying the data is one way to enable its reuse by third parties. This may at times be done poorly. For example, removing only direct identifiers — i.e., variables that provide an explicit link to a data subject and that can directly identify an individual — is often insufficient to ensure that the information is truly de-identified. The problem of de-identification involves quasi-identifiers, those variables that may not directly identify individuals, but that are highly correlated with unique identities and still may be used for indirect re-identification. These quasi-identifiers can be used either by themselves or in combination with available information, to uniquely identify individuals.

However, de-identification can and should be done effectively. One important lesson from the primary literature is that creating anonymized datasets requires statistical rigor and should not be done in a perfunctory manner. Organizations should perform an initial risk assessment, taking into account the current state of the art in both de-identification techniques and re-identification attacks. Since de-identification is neither simple nor straightforward, policy makers should support the development of strong tools, training, and best practices so that these techniques may be more widely adopted. In particular, a governance structure should be in place that enables organizations to continually assess the overall quality of their de-identified datasets to ensure that their utility remains high, and the risk of re-identification sufficiently low.

It is possible that de-identifying data may reduce the utility of the dataset. There is a balance to be struck in de-identification between the utility of the data and the risk of re-identification. The more specific and less general the de-identified data is, the more useful it may be to researchers, but with that comes a greater risk of re-identification. For some specific applications, it may not be feasible to sufficiently de-identify a dataset while maintaining the degree of utility needed for

37 See Daniel Castro and Travis Korte, "Data Innovation 101," Center for Data Innovation (2013), http://www.datainnovation.org/2013/11/data-innovation-101/.

38 See Pu Wang, Timothy Hunter, Alexandre M. Bayen, Katja Schechtner, Marta C. González, "Understanding Road Usage Patterns in Urban Areas," *Scientific Reports* 2 (Dec. 20, 2012), doi:10.1038/srep01001.

a specific use. As in the case of high-dimensional data, the use of "data enclaves" in which datasets are protected by legal agreements and physical security controls may be a more appropriate means of protecting privacy. Organizations like the U.S. Census Bureau use this approach to balance privacy with the benefits of allowing qualified researchers access to highly-sensitive datasets.[39]

In many cases, however, it is possible to strongly de-identify the data (and thus achieve a high degree of privacy), while at the same time preserve the required level of data quality necessary for data analysis. Maximizing both privacy and data quality enables a shift from a zero-sum paradigm to a positive-sum paradigm, a key principle of *Privacy by Design*.[40] This doubly-enabling "win-win" strategy avoids unnecessary trade-offs and illustrates that it is often possible to de-identify personal information in a manner that maintains both privacy and data quality.

An excellent example of this approach to de-identification involves the case of the U.S. Heritage Health Prize (HHP) claims dataset.[41] The HHP was a global data mining competition to predict the number of days patients would be hospitalized in the subsequent year by using current and previous years' claims data. The core dataset consisted of 3 years' worth of de-identified demographic and claims data on 113,000 patients.

The task of the competition to predict future hospitalizations required a high quality dataset. At the same time, due to the sensitivity of the information in the dataset, as well as the global nature of the competition, protecting the privacy of individuals was of paramount concern. In order to achieve both privacy and data quality, the researchers responsible for de-identifying the dataset took a three-step approach.

Led by de-identification expert, Dr. Khaled El Emam, they first preprocessed the claims data by applying some basic de-identification techniques:

- replacing direct identifiers with irreversible pseudonyms;
- removing uncommonly high values in the dataset (top-coding);
- truncating the number of claims per patient;
- removing high risk patients and claims; and
- suppressing provider, vendor, and primary-care provider identifiers, where patterns of treatment were discoverable.

Second, they examined the risk of re-identification by plausible attacks on the preprocessed dataset. The attacks they considered were: the nosey neighbor

39 See, for example, "The Role of RDCs," U.S. Census Bureau, n.d., http://www.census.gov/ces/rdcresearch/roleofrdcs.html.

40 See Ann Cavoukian, "Privacy by Design. The 7 Foundational Principles," 2011, http://www.ipc.on.ca/images/Resources/7foundationalprinciples.pdf.

41 See Khaled El Emam et al., "De-identification Methods for Open Health Data: The Case of the Heritage Health Prize Claims Dataset," *Journal of Medical Internet Research* 14, no. 1 (2012), doi:10.2196/jmir.2001.

adversary, matching with the voter registration list, and matching with the state inpatient database.

Finally, based on an analysis of various possible attacks, they used an automated algorithm to de-identify the dataset through generalization.

After de-identifying the dataset, the researchers then simulated the attacks considered above. Based on this empirical evaluation, it was estimated that the probability of re-identifying an individual was .0084. In other words, at most, an attacker could only hope to re-identify less than 1 per cent of the individuals in the dataset. Given that the probability threshold specified for the competition was .05, the level of privacy protection afforded by the suite of de-identification techniques used by the researchers far exceeded expectations. This study demonstrated that use of proper de-identification tools that involve re-identification risk measurement techniques makes it is extremely unlikely that an individual in a de-identified dataset will ever be re-identified.

Conclusion

In some circles, it is treated as a given that de-identified data can always be re-identified. What is most disturbing about this assertion and its attempt to grab headlines with sensationalist assumptions is that policy makers who require accurate information to determine appropriate rules and regulations may be unduly swayed. While it is not possible to guarantee that de-identification will work 100 per cent of the time, it remains an essential tool that will drastically reduce the risk of personal information being used or disclosed for unauthorized or malicious purposes. In the same way that locking the doors and windows to one's home reduces the risk of unwanted entry but is not a 100 per cent guarantee of safety, so too does de-identification, properly applied, protect the privacy of individuals without guaranteeing anonymity 100 per cent of the time. For this reason, when commenting on research into the re-identification risks associated with de-identified datasets — research that is essential to the continued success of de-identification standards — we must be careful not to overstate the findings of specific, oftentimes problematic cases of re-identification.

Moreover, de-identification continues to improve with additional research. When researchers discover a new technique for re-identifying data, this should be seen as an opportunity to improve de-identification techniques, not as an indictment of the utility of creating anonymized datasets. Indeed, this type of back and forth between discovering new risks, followed by developing and deploying countermeasures to mitigate those risks, is the bedrock of the scientific process underlying much of computer security.

Despite the misleading headlines and assertions made by some of those reporting on this topic, de-identification continues to be a valuable and effective mechanism for protecting personal information. We must remain vigilant in denying the perception, on the part of some commentators, that de-identification is an ineffective tool to protect privacy — we cannot allow it to become a self-fulfilling prophecy. We must get the story right — the stakes are high. Strong de-identification remains an essential tool to protecting privacy and allowing research to be conducted on de-identified datasets.

Overview of Organizations

Office of the Information and Privacy Commissioner of Ontario (IPC)

The role of the Information and Privacy Commissioner of Ontario, Canada, is set out in three statutes: the *Freedom of Information and Protection of Privacy Act*, the *Municipal Freedom of Information and Protection of Privacy Act* and the *Personal Health Information Protection Act*. The IPC acts independently of government to uphold and promote open government and the protection of personal privacy. Under the three *Acts*, the Information and Privacy Commissioner: resolves access to information appeals and complaints when government or health-care practitioners and organizations refuse to grant requests for access or correction; investigates complaints with respect to personal information held by government or health-care practitioners and organizations; conducts research into access and privacy issues; comments on proposed government legislation and programs; and educates the public about Ontario's access and privacy laws. More at: www.ipc.on.ca and www.privacybydesign.ca.

Information Technology & Innovation Foundation (ITIF)

The Information Technology and Innovation Foundation (ITIF) is a Washington, D.C.-based think tank at the cutting edge of designing innovation strategies and technology policies to create economic opportunities and improve quality of life in the United States and around the world. Founded in 2006, ITIF is a 501(c) 3 non-profit, non-partisan organization that documents the beneficial role technology plays in our lives and provides pragmatic ideas for improving technology-driven productivity, boosting competitiveness, and meeting today's global challenges through innovation. More at: www.itif.org.

Office of the Information and Privacy Commissioner of Ontario
2 Bloor Street East
Suite 1400
Toronto, Ontario
Canada M4W 1A8
Telephone: 416-326-3333
Fax: 416-325-9195
E-mail: info@ipc.on.ca
Website: www.ipc.on.ca

Information Technology and Innovation Foundation
1101 K Street N.W. Suite 610
Washington, DC 20005
Telephone: 202-449-1351
E-mail: mail@itif.org
Website: www.itif.org

Privacy by Design: www.privacybydesign.ca

June 2014