Office of the Victorian Information Commissioner

Protecting unit-record level personal information

The limitations of de-identification and the implications for the *Privacy and Data Protection Act 2014*

Foreword

This report is one of a number of publications on de-identification produced by, or for, the Victorian public sector. Notably, in early 2018 Victoria's Chief Data Officer issued a de-identification guideline to point to what 'reasonable steps' for de-identification looks like in the context of data analytics and information sharing under the *Victorian Data Sharing Act 2017* (**VDS Act**).

This paper is not aimed at the work conducted by the Victorian Centre for Data Insights (**VCDI**), where information sharing occurs within government with appropriate controls, and it is not intended to inhibit that work. Rather, it speaks to the use of de-identification more broadly, in circumstances where so-called 'de-identified' data is made freely available through *public* or other less inhibited release of data sets, which occurs in so-called "open data" programs. This report should be interpreted in that context.

De-identification is a subject that has received much attention in recent years from privacy regulators around the globe. Once touted as a silver bullet for protecting the privacy of personal information, the reality is that when it involves the release of data to the public, the process of de-identification is much more complex.

As improvements in technology increase the type and rate at which data is generated, the possibility of re-identification of publicly released data is greater than ever. Auxiliary information – or secondary information – can be used to connect an individual to seemingly de-identified data, enabling an individual's identity to be ascertained. Auxiliary information can come from anywhere, including other publicly available sources online.

In recent examples of successful re-identification that we have seen in Australia, it is clear that those releasing de-identified data did not appreciate the auxiliary information that would be available for re-identification – in that they did not expect re-identification would be possible. Individual data elements may be non-distinct and recognisable in many people, but a combination of them will often be unique, making them attributable to a specific individual. This is why de-identification poses a problem for unit-record level data.

This report has been produced to demonstrate the complexities of de-identification and serve as a reminder that even if direct identifiers have been removed from a data set, it may still constitute 'personal information'. The intention is not to dissuade the use of de-identification techniques to enhance privacy, but to ensure that those relying on and sharing de-identified information to drive policy design and service delivery, understand the challenges involved where the husbandry of that data is not managed.

The Office of the Victorian Information Commissioner (**OVIC**) supports the use of data to improve policy and decision making, service design, and enhance efficiencies across government. In some cases, these pursuits will be fueled by personal information. Depending on the nature of the information and the potential risks should re-identification occur, it is appropriate to rely on access controls and secure research environments to provide an addition layer of protection, in addition to de-identifying the information before it is used and shared. This becomes most acute outside research environments or data labs.

Public release of de-identified information may not always be a safe option, depending on the techniques used to treat the data and the auxiliary information that the public may have access to. Wherever unit level data – containing data related to individuals – is used for analysis, OVIC's view is that this is most appropriately performed in a controlled environment by data scientists. Releasing the data publicly in the hope that 'de-identification' provides protection from a privacy breach is, as this paper demonstrates, a risky enterprise.

As a general principle, we encourage agencies, where dealing with individual level data, to work with the VCDI or another expert in analytics, and to avoid the use of individual-level data sets, whether de-identified or not, in any data released for analysis outside these expert environments.

The report has been written to appeal to both technical and non-technical audiences. While some of the concepts explored throughout the paper can be complex, they are explained in a way that also makes them digestible to those without a background in the subject matter. Where possible, links to additional resources have been provided in footnotes, and suggested further readings are contained at the end of the document.

OVIC would like to thank Dr Vanessa Teague, Dr Chris Culnane and Dr Benjamin Rubinstein of the University of Melbourne for preparing this report on our behalf.

Rachel Dixon
Privacy and Data Protection Deputy Commissioner
May 2018

Protecting unit-level record personal information

The limitations of de-identification and the implications for the *Privacy and Data Protection Act 2014*

Protecting unit-record level personal information

The limitations of de-identification and the implications for the *Privacy and Data Protection Act 2014*

1. Summary

A detailed record about an individual that has been de-identified, but is released publicly, is likely to be re-identifiable, and there is unlikely to be any feasible treatment that retains most of the value of the record for research, and also securely de-identifies it. A person might take reasonable steps to attempt to de-identify such data and be unaware that individuals can still be reasonably identified.

The word 'de-identify' is, unfortunately, highly ambiguous. It might mean removing obvious identifiers (which is easy) or it might mean achieving the state in which individuals cannot be 'reasonably identified' by an adversary (which is hard). It is very important not to confuse these two definitions. Confusion causes an apparent controversy over whether de-identification "works", but much of this controversy can be resolved by thinking carefully about what it means to be secure. When many different data points about a particular individual are connected, we recommend focusing instead on restricting access and hence the opportunity for misuse of that data. Secure research environments and traditional access control mechanisms are appropriate.

Aggregated statistics, such as overall totals of certain items (even within certain groups of individuals) could possibly be safely released publicly. Differential privacy offers a rigorous and strong definition of privacy protection, but the strength of the privacy parameters must be traded off against the precision and quantity of the published data.

This paper discusses de-identification of a data set in the context of release to the public, for example via the internet, where it may be combined with other data. That context includes the concept of "open data", in which governments make data available for any researchers to analyse in the hope they can identify issues or patterns of public benefit.

Therefore, it's important to emphasise that this document should not be read as a general warning against data sharing within government, or in a controlled research environment where the combination of the data set with other data can be managed. It is not intended to have a chilling effect on sharing of data in those controlled environments.

2. Background and scope

Victoria's primary information privacy law, the *Privacy and Data Protection Act 2014* (**PDP Act**), places obligations on the Victorian public sector in relation to the collection, management and use of personal information.[1]

The definition of 'personal information' is:

> *information or an opinion (including information or an opinion forming part of a database), that is recorded in any form and whether true or not, about an individual whose identity is apparent, or can reasonably be ascertained, from the information or opinion, but does not include information of a kind to which the Health Records Act 2001 applies.*[2]

[1] Specifically, Part 3 of the PDP Act applies to 'personal information'.
[2] Section 3, PDP Act.

The PDP Act recognises the concept of de-identification. Section 3 of the PDP Act defines it as:

> *de-identified, in relation to personal information, means personal information that no longer relates to an identifiable individual or an individual that can be reasonably identified.*

The term 'de-identify' is also used under Information Privacy Principle 4, requiring organisations to "destroy or permanently de-identify personal information if it is no longer needed for any purpose".[3]

It is widely believed that:

a) personal information can be de-identified (i.e. that a process or technique can be applied to personal information to reliably convert it from personal information to non-identifying information); and

b) the de-identified information can be used for a range of purposes, including the development of public policy, for research in the public interest etc.

The VDS Act was enacted to promote data sharing across Victorian government departments and agencies to inform policy making, service planning and design. The VDS Act also establishes the role of the Chief Data Officer who leads the VCDI.

The VDS Act requires that:

> *The Chief Data Officer or a data analytics body must take reasonable steps to ensure that data received from data sharing bodies and designated bodies under this Act no longer relates to an identifiable individual or an individual who can reasonably be identified before using that data for the purpose of data analytics work.*[4]

In taking 'reasonable steps', a data custodian must have regard to not only the mathematical methods of de-identifying the information, but also "the technical and administrative safeguards and protections implemented in the data analytics environment to protect the privacy of individuals".[5]

Therefore, there is a possibility that in some circumstances, a dataset in which 'reasonable steps' have been taken for de-identification under the VDS Act may not be de-identified according to the PDP Act, because individuals may still be 'reasonably identified' if the records are released publicly outside the kinds of research environments described in the VDS Act.

In this report, we describe the main techniques that are used for de-identifying personal information. There are two main ways of protecting the privacy of data intended for sharing or release: removing information, and restricting access. We explain when de-identification does (or does not) work, using datasets from health and transport as examples. We also explain why these techniques might fail when the de-identified data is linked with other data, so as to produce information in which an individual is identifiable.

3. Framing the controversy: Does de-identification work?

Does de-identification work? In one sense, the answer is obviously yes: de-identification can protect privacy by deleting all the useful information in a data set. Conversely, it could produce a valuable data set by removing names but leaving in other personal information. The question is whether there is any middle ground; are there techniques for de-identification that "work" because they protect the privacy of unit-record level data while preserving most of its scientific or business value?

[3] IPP 4.2, Schedule 1, PDP Act.
[4] Section 18(1), VDS Act.
[5] Section 18(2), VDS Act.

Controversy also exists in arguments about the definitions of 'de-identification' and 'work'. De-identification might mean:

- *following a process* such as removing names, widening the ranges of ages or dates, and removing unusual records; or
- *achieving the state* in which individuals cannot be 'reasonably identified'.

These two meanings should not be confused, though they often are. A well-intentioned official might carefully follow a de-identification process, but some individuals might still be 'reasonably identifiable'. Compliance with de-identification protocols and guidelines does not necessarily imply proper mathematical protections of privacy. This misunderstanding has potential implications for privacy law, where information that is assumed to be de-identified is treated as non-identifiable information and subsequently shared or released publicly.

De-identification would work if an adversary who was trying to re-identify records could not do so successfully. Success depends on 'auxiliary information' – extra information about the person that can be used to identify their record in the dataset. Auxiliary information could include age, place of work, medical history etc. If an adversary trying to re-identify individuals does not know much about them, re-identification is unlikely to succeed. However, if they have a vast dataset (with names) that closely mirrors enough information in the de-identified records, re-identification of unique records will be possible.

4. Can the risk of re-identification be assessed?

For a particular collection of auxiliary information, we can ask a well-defined mathematical question: can someone be identified uniquely based on just that auxiliary information?

There are no probabilities or risks here – we are simply asking what can be inferred from a particular combination of data sets and auxiliary information. This is generally not controversial. The controversy arises from asking what auxiliary information somebody is likely to have.

For example, in the Australian Department of Health's public release of MBS/PBS billing data, those who prepared the dataset carefully removed all demographic data except the patient's gender and year of birth, therefore ensuring that demographic information was not enough on its own to identify individuals. However, we were able to demonstrate that with an individual's year of birth and some information about the date of a surgery or other medical event, the individual could be re-identified.[6] There was clearly a mismatch between the release authority's assumptions and the reality about what auxiliary information could be available for re-identification.

5. How re-identification works

Re-identification works by identifying a 'digital fingerprint' in the data, meaning a combination of features that uniquely identify a person. If two datasets have related records, one person's digital fingerprint should be the same in both. This allows linking of a person's data from the two datasets – if one dataset has names then the other dataset can be re-identified.

[6] Chris Culnane, Benjamin Rubinstein & Vanessa Teague, *Health data in an open world*, December 2017, available at https://arxiv.org/abs/1712.05627.

Computer scientists have used linkage to re-identify de-identified data from various sources including telephone metadata,[7] social network connections,[8] health data[9] and online ratings,[10] and found high rates of uniqueness in mobility data[11] and credit card transactions.[12] Simply linking with online information can work.[13]

Most published re-identifications are performed by journalists or academics. Is this because they are the only people who are doing re-identification, or because they are the kind of people who tend to publish what they learn? Although by definition we won't hear about the unpublished re-identifications, there are certainly many organisations with vast stores of auxiliary information. The database of a bank, health insurer or employer could contain significant auxiliary information that could be of great value in re-identifying a health data set, for example, and those organisations would have significant financial incentive to do so. The auxiliary information available to law-abiding researchers today is the absolute minimum that might be available to a determined attacker, now or in the future.

This potential for linkage of one data set with other data sets is why the federal Australian Government's draft bill to criminalise re-identification is likely to be ineffective, and even counterproductive.[14] If re-identification is not possible then it doesn't need to be prohibited; if re-identification is straightforward then governments (and the people whose data was published) need to find out.

The rest of this report examines what de-identification is, whether it works, and what alternative approaches may better protect personal information. After assessing whether de-identification is a myth, we outline constructive directions for where to go from here. Our technical suggestions focus on differential privacy and aggregation. We also discuss access control via secure research environments.

6. Three medical billing datasets and the Opal Transport Data

This section takes three different databases of medical billing records, plus one open transport dataset, as concrete examples of privacy risks and information gains associated with de-identification. Although health information is governed in Victoria under the *Health Records Act 2001,* and not the PDP Act, the mathematical techniques and questions are the same: Why are some datasets useful and others not? How does re-identification work? Why are some datasets harmless and others not?

6.1. MBS/PBS data

Most of Australia's public health system is billed through a centralised system called the Medical Billing System (**MBS**) and a scheme for government-subsidised medications called the Pharmaceutical Benefits Scheme (**PBS**). Although it is only billing data, without any lab results or doctors' notes, it is still highly sensitive.

---

[7] Mudhakar Srivatsa & Mike Hicks, 'Deanonymizing Mobility Traces: Using Social Networks as a Side-Channel', *Computer and Communications Security*, October 2012, available at http://www.cs.umd.edu/~mwh/papers/GraphInfoFlow.CCS2012.pdf.

[8] Arvind Narayanan, Elaine Shi & Benjamin Rubinstein, 'Link Prediction by De-anonymization: How We Won the Kaggle Social Network Challenge', *The 2011 International Joint Conference on Neural Networks*, October 2011, available at: https://arxiv.org/pdf/1102.4374.pdf.

[9] Latanya Sweeney, '*k*-anonymity: A model for protecting privacy' *International Journal on Uncertainty, Fuzziness and Knowledge-based Systems*, Vol. 10, No. 5, 2002, available at: https://dataprivacylab.org/dataprivacy/projects/kanonymity/kanonymity.pdf.

[10] Arvind Narayanan & Vitaly Shmatikov, 'Robust De-anonymization of Large Sparse Datasets', *Security and Privacy*, May 2008, available at: https://www.cs.cornell.edu/~shmat/shmat_oak08netflix.pdf.

[11] Yves-Alexandre de Montjoye, César A. Hidalgo, Michel Verleysen & Vincent D. Blondel, 'Unique in the Crowd: The privacy bounds of human mobility', *Scientific reports*, March 2013, available at: https://www.nature.com/articles/srep01376?ial=1.

[12] Yves-Alexandre de Montjoye, Laura Radaelli, Vivek Kumar Singh & Alex "Sandy" Pentland, 'Unique in the shopping mall: On the reidentifiability of credit card metadata', *Science*, Vol. 347, No. 6221, 2015, available at: http://science.sciencemag.org/content/347/6221/536.full.

[13] Michael Barbaro & Tom Zeller Jr, 'A Face Is Exposed for AOL Searcher No. 4417749' *NY Times*, 9 August 2006, available at: https://www.nytimes.com/2006/08/09/technology/09aol.html; Charles Duhigg, 'How Companies Learn Your Secrets' *NY Times*, 16 February 2012, available at: https://www.nytimes.com/2012/02/19/magazine/shopping-habits.html.

[14] See Privacy Amendment (Re-identification Offence) Bill 2016, https://www.aph.gov.au/Parliamentary_Business/Bills_Legislation/Bills_Search_Results/Result?bId=s1047.

In its raw form it contains a list, for each individual, of medical tests taken, doctors' visits, prescriptions and surgeries, along with the price they paid, the government benefit and the exact date. Clearly such data is of great value for scientific research and public health policy development, yet it has the potential to expose personal information of a damaging nature. Some MBS/PBS item codes are generic – code 00023 occurs millions of times, indicating a visit to the GP – while others are highly specific and sensitive, such as prescriptions that are only for HIV patients.

The Australian Government openly released two different versions of this data:

- **Group statistics, complete**: a complete list of the frequencies of each billing code for the whole population. These are aggregated into patient age ranges of ten years, and reported for each state and each month, for the years 1984 – 2014;[15] and
- **Longitudinal unit records, 10% sample**: 'de-identified' longitudinal patient records, for the years 1984—2014, for 10% of the Australian population. These include the patient's gender, year of birth, and complete billing history. Some rare items were removed, the dates were randomly perturbed by up to two weeks, and the supplier (doctor) IDs were encrypted.[16]

The two datasets could be regarded as two different ways of de-identifying the same unit-record level data. One does a reasonable job of preserving privacy, while the other retains most of the scientific value of the dataset.

Longitudinal records and how they can be re-identified

A typical record in the longitudinal dataset includes a gender, age and encrypted patient ID, for example like this:

(Encrypted) patient ID	0345952108
Gender	F
Year of birth	1963

Then, for each encrypted patient ID, it contains every billing record. A typical billing record looks like this:

(Encrypted) patient ID	0345952108
State	Vic-Tas[17]
Date	7 Aug 1992
(Encrypted) supplier ID	2340981234
Item code	00023 (GP visit)
Price paid by patient	$85
Price reimbursed by Medicare	$60
Various other details	…

[15] See https://data.gov.au/dataset/medicare-benefits-schedule-mbs-group.
[16] This dataset was taken offline when we showed that the encryption was insecure.
[17] States are coded by number, but we use names here.

Many combinations of characteristics in the longitudinal records are unique, including billing totals, childbirth dates and some surgeries given year of birth and state. Patient records can be easily re-identified given a little auxiliary information, which has serious consequences for the privacy of the patient.[18]

Of course, a unique match may not be correct – it may be that the target person is not in the 10% sample, but shows a coincidental resemblance to someone who is. However, it is possible to use yet more auxiliary data to gain very high confidence in some cases.

Records in the group statistics dataset are much simpler. This is a real complete record:

Age range	45—54
State	Qld
Month	August
Year	2011
Gender	M
Item code	38556
Price reimbursed by Medicare	$2240

It says that a man aged between 45 and 54 had an aortic valve replacement (coded by 38556) in Queensland in August of 2011. In the group statistics database, some codes are very uncommon, and are either absent or charged only once in particular months for particular ranges of patient age. Overall, 27% are unique. Yet compared to the longitudinal records, there is much less risk to patient privacy – the group statistics are fairly harmless data.

Why? It isn't because re-identification is impossible in one data set but possible in the other. For example, former Australian Prime Minister Kevin Rudd had an aortic valve replacement in Brisbane in August 2011. This was described in an online news article that gave an exact date. According to the group statistics dataset, this surgery was unique in that month and age range. We can therefore identify the exact record that corresponds to his surgery – it is the one in the previous paragraph. However, from this record we learn only what we knew already from Wikipedia – his age, gender, and the state in which the surgery was performed.

Dimensionality and why it matters

Kevin Rudd's record is not included in the longitudinal 10% sample, because no such surgery for the right age, state and gender appears there. However, if Mr Rudd's record had been chosen (and there was a 10% chance it would have been), the re-identification of that record would have implied the retrieval of the entire 30 years' worth of his medical billing history – information that is not otherwise available online.

The difference between the two datasets is *dimensionality* – the sheer number of independent facts in each record. The group statistics dataset has a very low dimension, so no single record says much. The dataset simply omits a lot of the information in each longitudinal record – crucially, it doesn't indicate which records belong to the same patient. The longitudinal records have a much higher dimension, so uniqueness based on a few characteristics can be leveraged to retrieve much more information about the person.

[18] A non-technical write up of the main ideas is here: https://pursuit.unimelb.edu.au/articles/the-simple-process-of-re-identifying-patients-in-public-health-records.

Multiple independent data points allow even quite vague information to be combined together to re-identify a person. For example, there are about 300,000 childbirths in Australia each year, so the set of people with a baby born in any given year is large. However, the number of people with childbirths in any two particular years is much smaller, and even smaller again for three. If we combine this information with the patient's year of birth (which matches 2-300,000 people) and their state of residence (which they share with a few million people), the combination of these three or four vague facts might match a unique individual or a very small set.

This remains true even when the information isn't strictly independent in the statistical sense, for example because people of a certain age are more likely to become parents. This general pattern occurs in data privacy research across a variety of different domains, and was best expressed in the study "Unique in the Crowd" – each data point may match a large crowd of others, but a combination of a few data points about an individual is very often unique.[19] This is why de-identification does not work on complex unit-record level data.

Records of high dimension are both very useful for science and very problematic for privacy. This has been called 'the curse of dimensionality'.[20]

6.2. The Heritage Health Prize Dataset: Does de-identified data preserve utility?

Between 2011 and 2013, a US-based organisation, Heritage Provider Network, ran the Heritage Health Prize challenge, offering a $3 million prize to the team that did the best job of predicting hospital stay length among people in a de-identified dataset. Records included (encrypted) patient and supplier IDs, the (capped) length of time the patient had spent in hospital, low-precision data about the type of medical event they were treated for, and other details such as the delay between billing and payment.

Although the dataset is no longer openly available, its properties can be estimated by looking at the success of the analytics solutions submitted for the contest. Success was defined by a scoring formula[21] and published on an online leader board.[22]

One obvious benchmark for scoring was the Optimized Constant Value Benchmark, which represents the best score available from simply guessing that everyone went to the hospital for the same number of days. The optimum guess is the average (according to the scoring formula). This guess produces a score of 0.486, which is close to the standard deviation (on a log scale) in the dataset.

The $3 million dollar prize was offered for the best prediction, if there was one that scored better (smaller) than 0.4. Nobody came even close. Of the 1353 applicants, nearly 500 scored higher (i.e. worse) than the Optimized Constant Value Benchmark. In other words, their sophisticated analysis and prediction based on the de-identified data was a less effective prediction than simply taking the average. The best score was 0.443. This means that the best analytics could explain less than 10% of the standard deviation in hospital stays – nearly 20% would have been necessary to collect the prize.

Of course, none of this proves that the de-identification itself destroyed the predictive quality of the data – perhaps no accurate predictions would have been possible on the raw data anyway. Nevertheless, it should

[19] Yves-Alexandre de Montjoye, César A. Hidalgo, Michel Verleysen & Vincent D. Blondel, 'Unique in the Crowd: The privacy bounds of human mobility', *Scientific reports*, March 2013, available at: https://www.nature.com/articles/srep01376?ial=1.
[20] Paul Ohm, 'Broken promises of privacy: Responding to the surprising failure of anonymization', *UCLA Law Review*, Vol. 57, 2010, available here: http://paulohm.com/classes/techpriv13/reading/wednesday/OhmBrokenPromisesofPrivacy.pdf.
[21] See http://www.heritagehealthprize.com/c/hhp/rules.
[22] These are the not-quite-final results, the only results that are public. If the final results were better, they have never been publicly announced as such.

encourage caution in the use of this example to justify de-identification as the solution to the problem of making data available for analysis.

There is also good evidence that the data would have been re-identifiable given reasonable auxiliary information. In an independent analysis, Narayanan showed "a significant fraction of members are vulnerable if the adversary knows demographic information and a sufficient number of diagnosis codes (and year of claim for each one)".[23] The number of patients who are unique given a particular combination of auxiliary facts is easy to check. The real disagreement is over what auxiliary information the adversary is likely to have. Those who de-identified the Heritage Health dataset wrote "we assume that the adversary does not know the order of the quasi-identifier values. For example, the adversary would not know that the heart attack occurred before the broken arm...". Narayanan replies, "I cannot fathom the reason for this. If the auxiliary information comes from online review sites, for example, detailed timeline information is very much available."[24]

We could not find any example of securely de-identified detailed unit-record level data that demonstrably preserved the useful information for scientific study.

6.3. NSW Opal tap-on and tap-off tallies

The last example involves the application of differential privacy to an aggregated transport dataset, derived from the NSW Opal system.

Transport NSW recently released data from its Opal ticketing system as open data.[25] The underlying data would have included, for each individual Opal card, the complete sequence of tap-on and tap-off locations and times. This is highly sensitive, because it could reveal a person's entire travel history. Instead of releasing the raw data, Transport NSW chose to release a highly aggregated version; rather than publish the detailed record of any individuals, they broke the links between different events on the same card. They released only the total number of tap-ons and tap-offs at various locations and times.

As an added protection, the totals were randomly perturbed to produce a differentially private version of the data. This concept is described in the section on differential privacy later in this report.

A significant amount of information needed to be removed in order for the data to be safely publishable online; a scientist who wants to link the beginning and end of a trip cannot do so without applying directly to Transport NSW for the raw data.[26]

This is an example of both good technical treatment (beginning from low-dimensional data) and a good process. Making the algorithms public, and having them open to scrutiny, means that any errors or weaknesses are likely to be found and corrected before more data is released.

6.4. Is there a good solution?

The controversy probably stems from a different understanding of what it means for de-identification to "work". Scientists who regard de-identification as working well tend to come from statistical or medical disciplines; the charge that it doesn't work tends to come from computer scientists with training in more adversarial notions of information security, cryptography, and privacy. The notion of risk in these two

[23] Arvind Narayanan, An Adversarial Analysis of the Reidentifiability of the Heritage Health Prize Dataset, 2011, available at: http://randomwalker.info/publications/heritage-health-re-identifiability.pdf.
[24] Arvind Narayanan & Edward W Felton, *No silver bullet: De-identification still doesn't work*, 2014, available at: http://randomwalker.info/publications/no-silver-bullet-de-identification.pdf.
[25] The dataset is available here: https://opendata.transport.nsw.gov.au/dataset/opal-tap-on-and-tap-off.
[26] Data61 wrote a detailed report of Transport NSW's de-identification techniques, available at: https://arxiv.org/abs/1705.05957. An analysis of the technical methods used before the data was released is available at: https://arxiv.org/abs/1704.08547.

communities is very different – one is statistical and random, while the other envisages a determined and ingenious adversary with full access to a wide collection of auxiliary information.

For detailed unit-record level data, there is no de-identification method that "works" in the sense that it preserves the scientific value of the data while preventing re-identification by a motivated attacker with auxiliary information. Evaluating the re-identification risk of publicly releasing unit-record-level data is impossible, since it requires knowledge of all contexts in which the data could be read.

7. Survey of privacy-enhancing techniques and definitions

There are two main ways of protecting the privacy of data when it is to be made publicly available: removing information and restricting access.

This section gives an overview of three methods for removing information:

- the Health Insurance Portability and Accountability Act (**HIPAA**) standard
- k-anonymity, and
- differential privacy.

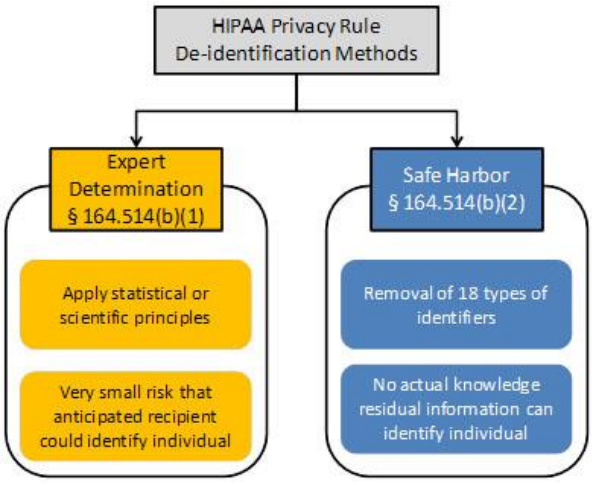
Of these, only differential privacy produces a rigorous and well-defined privacy guarantee.

We then briefly survey methods for restricting access, which may be a viable solution for data that remains re-identifiable.

7.1. Methods of removing personal information

Health Insurance Portability and Accountability Act Privacy Rule

One of the most notable, and often cited, de-identification standards is contained within the US HIPAA of 1996. HIPAA defines two options for performing de-identification: Expert Determination and Safe Harbor. Once either option has been performed the data is no longer covered by the HIPAA Privacy Rule and may be released. The US Department of Health and Human Services provides the following diagram to illustrate the options:[27]



[27] See https://www.hhs.gov/hipaa/for-professionals/privacy/special-topics/de-identification/index.html.

Expert Determination

The expert determination approach requires an expert with experience in the relevant methods for making information no longer individually identifiable. This approach acknowledges that there will remain a low risk of re-identification, the threshold of which is to be decided by the expert and the organisation releasing the data.

Crucially, the risk should be context-dependent and evaluated based on the recipient's ability to re-identify the data. Online guidance states that "…the risk of identification that has been determined for one particular data set in the context of a specific environment may not be appropriate for the same data set in a different environment or a different data set in the same environment. As a result, an expert will define an acceptable 'very small' risk based on the ability of an anticipated recipient to identify an individual."[28]

Safe Harbor

The Safe Harbor pathway is far more formulaic. It defines 18 types of identifiers that must be suppressed or modified for the information to be considered 'de-identified'. The 18 types of identifiers are as follows:

- Names
- Location – including geographic subdivisions smaller than a state. The first 3 digits of the zip code may be included if the Bureau of Census indicates the geographic unit formed by areas with the same first 3 digits has a population greater than 20,000, otherwise it must be set to 000.
- Date elements, except year, for dates directly related to an individual. Remove all date elements that are indicative of someone over the age of 89, or replace with a category of 90 or older.
- Telephone Numbers
- Vehicle Identifiers, including serial numbers and license plates
- Fax Number
- Device Identifiers and Serial Numbers
- Email Address
- URLs
- Social Security Number
- IP (Internet Protocol) Address
- Medical Record Numbers
- Biometric identifiers
- Health Plan beneficiary numbers
- Full-face photographs and comparable images

[28] US Department of Health and Human Services, *Guidance Regarding Methods for De-identification of Protected Health Information in Accordance with the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule*, 2012, available at: https://www.hhs.gov/sites/default/files/ocr/privacy/hipaa/understanding/coveredentities/De-identification/hhs_deid_guidance.pdf.

- Account numbers
- Any other unique identifying numbers, characteristics or code
- Certificate or license numbers

If the entity removes or modifies the above identifiers and has no actual knowledge of any ability to re-identify the data, then it is considered to be de-identified and no longer covered by the Privacy Rule, i.e. it is free for release.

Summary of privacy protection (HIPAA)

It should be clear by now that this formulaic notion of de-identification does not guarantee good privacy protection. If an adversary can find information about the data itself, including which medical events a person experienced on what dates, there is a significant risk that this could be combined with date and location information in the de-identified record to re-identify the person.

k-anonymity

k-anonymity is an early and influential sequence of frameworks for privacy protection.[29] It assumes that an adversary will have access to only a certain set of attributes, which can be predicted in advance. These are called 'quasi-identifiers'. They typically include information such as age, gender, and residential postcode.

Privacy is meant to be achieved through membership of a group of k 'similar' records, supposedly guaranteeing that each individual is hidden in a crowd of 'k' indistinguishable people – the larger the value of k, the less likely it seems that an individual's information can be inferred. A dataset achieves k-anonymity if, for each set of possible values of the quasi-identifiers, there are either no records or at least k records with the same values. For example, if age, gender and residential postcode are the quasi-identifiers, then a dataset achieves 5-anonymity if there are either 5 or zero 43-year-old men in each postcode (and likewise for other ages and genders).

If the raw data isn't sufficiently private, a number of approaches can be taken to achieve k-anonymity such as 'generalisation', whereby attribute values are made coarse (for example, an age of 15 being replaced by a range 12—18) and 'suppression', i.e. removal of records that aren't hidden in a large enough group.

For example, if we consider only name, gender and state as quasi-identifiers, most records in the MBS/PBS longitudinal 10% sample enjoy k-anonymity for k greater than 1,000. The exceptions are the (very few) records for people with years of birth in the 1880s or 1890s. If we were concerned about re-identification of those records, we could use generalisation – instead of listing the individual years of birth we could replace them with a value meaning "some time before 1900". If this still did not produce acceptable k-anonymity (for example in a small state), we could suppress (i.e. remove) those records altogether.

An advantage of k-anonymity is that it is straightforward to check and establish. The definition appears to protect privacy and is easily explained. However, this is not a secure guarantee of privacy, because an adversary may have access to much more information than the chosen quasi-identifiers.

[29] Latanya Sweeney, '*k*-anonymity: A model for protecting privacy' *International Journal on Uncertainty, Fuzziness and Knowledge-based Systems*, Vol. 10, No. 5, 2002, available at: https://dataprivacylab.org/dataprivacy/projects/kanonymity/kanonymity.pdf.

Criticisms and Limitations

The obvious shortcoming in principle, which has been applied to successfully re-identify k-anonymous data in practice, is that the adversary might know auxiliary information other than the chosen quasi-identifiers. Numerous linkage attacks have been demonstrated on k-anonymity and related concepts such as l-diversity, though they serve as the basis of HIPPA and derivative privacy legislation world-wide.

The re-identifications in the MBS/PBS 10% sample is a clear example: although the quasi-identifiers (year of birth, gender) left large sets of matching records, we could easily find auxiliary information about other parts of the data (surgery dates, childbirths, state of residence). The assumptions about what data would be available to an adversary were wrong.

Another important limitation is that inferences may be possible even without access to auxiliary information: if the patient is located (given the quasi-identifiers) among a crowd of five people, but all five people have the same illness, then the fact that that person has that illness is exposed.

Summary of privacy protection (k-anonymity)

k-anonymity (and l-diversity, a related notion) protect privacy only if their assumptions about the adversary's limited knowledge are correct. This may be adequate protection against a well-meaning researcher in a controlled environment, but it is not a secure defence against a motivated adversary. It has been shown to fail in practice.

Differential Privacy

In contrast with k-anonymity and its variations, differential privacy[30] gives a proven limit on information leakage, even against an adversary with access to all relevant auxiliary information.

Importance of randomisation

Differential privacy leverages randomisation: a differentially-private mechanism takes a data set and, for example, adds random noise, then outputs the result. The average results can be very accurate, but details about an individual are obscured by randomness.

The release itself could be a number[31] (for example, the median body mass index of a cohort), a table[32] (for example, a marginal table of population statistics), a function[33] (for example, a classifier for diagnosing future cases of Ebola), or simple individual records.[34] The adversary sees the randomised output and tries to guess what is truly in the database.

[30] Cynthia Dwork, *Differential Privacy: A Survey of Results*, 2008, available at: https://www.microsoft.com/en-us/research/wp-content/uploads/2008/04/dwork_tamc.pdf.
[31] Cynthia Dwork, Frank McSherry, Kobbi Nissim & Adam Smith, Calibrating Noise to Sensitivity in Private Data Analysis, 2006, available at: https://iacr.org/archive/tcc2006/38760266/38760266.pdf.
[32] Boaz Barak, Kamalika Chaudhuri, Cynthia Dwork, Satyen Kale, Frank McSherry & Kunal Talwar, *Privacy, Accuracy, and Consistency Too: A Holistic Solution to Contingency Table Release*, 2007, available at: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.163.5503&rep=rep1&type=pdf.
[33] Francesco Aldà & Benjamin I. P. Rubinstein, *The Bernstein Mechanism: Function Release under Differential Privacy*, 2017, available at: http://www.aaai.org/ocs/index.php/AAAI/AAAI17/paper/download/14735/14325.
[34] Min Lyu, Dong Su & Ninghui Li, *Understanding the Sparse Vector Technique for Differential Privacy*, 2017, available at: http://www.vldb.org/pvldb/vol10/p637-lyu.pdf.

Intuitive Definition

Differential privacy is achieved if changes to a single record do not significantly affect the probability of any particular output. This means that observing the randomised value doesn't give the adversary much information.

Suppose that two databases differ in only one record (Kevin Rudd's, for example) and the adversary knows everything about the data, except whether that record is included or not – they do not know which of the two databases is the true one. The adversary is allowed to query the (randomised) differentially-private mechanism to try to discern whether Mr Rudd's record is present.

For example, the adversary could query the number of aortic valve replacements in August 2011 for men of Mr Rudd's age and state of residence. If there were no randomness, this would immediately reveal the truth.

The application of randomness to the answer introduces uncertainty. The idea is depicted by Figure 1 below. Possible database D_1 does not include Kevin Rudd's record – the true answer is zero. Possible database D_2 is the same as D_1, plus Kevin Rudd's record – the true answer is one. In both cases the differentially-private mechanism outputs a randomised value which averages to the true value, but might be a little more or less. If the mechanism outputs a 1, the adversary doesn't know whether this is D_2 (with zero randomness) or D_1 (with +1 added randomly). Any particular output value could have arisen from the (randomly perturbed) output of either database.

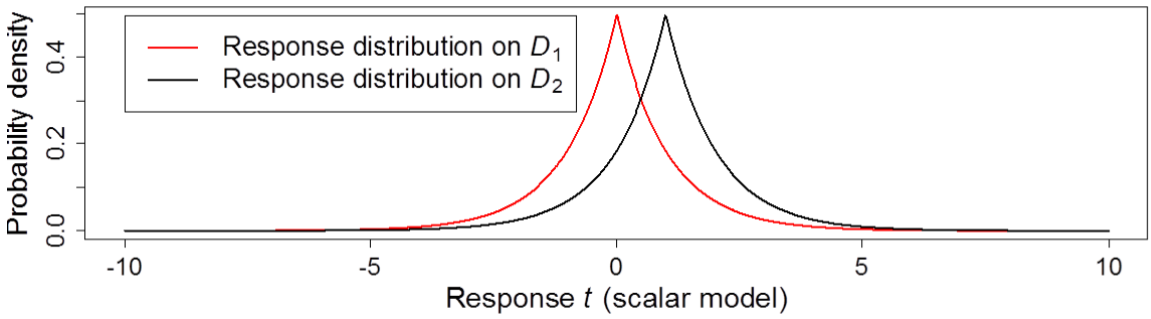


Figure 1: Differential privacy requires that on two datasets D_1 and D_2 that are identical except for one record difference, the distribution of output responses are similar. The privacy parameter ε is the worst-case ratio between these distributions.

It is important to understand that differential privacy is a restriction on information leakage, not a guarantee of perfect privacy. In Figure 1, for example, a 0 is twice as likely to indicate D_1 as D_2. Negative values are a little more likely to come from D_1 (the database without Mr Rudd's record); positive values are a little more likely if his record is present. The adversary knows this. Privacy is very well preserved when the distributions are very similar – in that case, any given answer is nearly equally likely to have come from either possible dataset. The privacy parameter ε (epsilon) central to differential privacy measures the worst-case information leakage – it is large when there are some outputs that are much more likely to derive from one dataset over the other, hence revealing a strong indication of which database is truly there.

General Threat Model

A powerful hypothetical adversary is permitted to have:

- **unbounded computational resources**: privacy guarantees hold up against adversaries with access to more than all computing facilities of the world combined;
- **almost complete knowledge of the dataset:** all of the dataset except one record – this covers cases where participants of a data set might collude to pool information.

Using these abilities, an adversary witnesses repeated responses from the release mechanism and tries to guess which data set is being used. This situation is represented in Figure 2 below. The adversary is trying to guess which of the possible data sets (on the left) is the true one. Each different data set has a slightly different response distribution (middle), but the adversary cannot reliably distinguish the observations from the true distribution (the histogram; right) with the right input data set rather than the others.

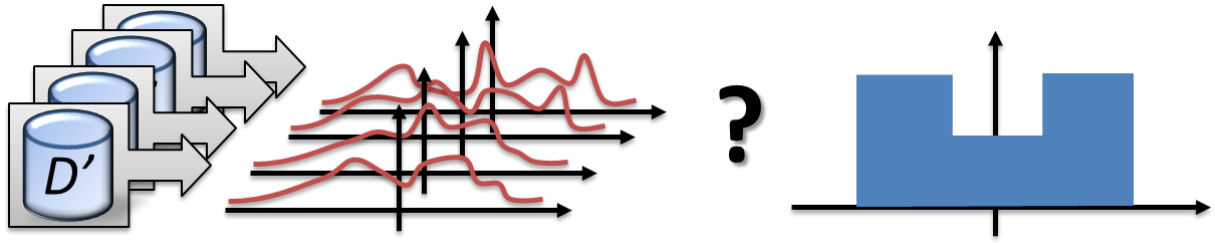


Figure 2: An adversary is trying to guess which dataset the observations came from, knowing their full probability distributions but observing only a few outputs.

The privacy parameter ε is chosen by the data curator. The smaller ε, the closer the response distributions and the more privacy preserved; larger values correspond to less privacy. Of course, smaller values usually require more randomness to be added, hence reducing the accuracy of the outputs.

For further technical reading on the topic of differential privacy, see Cynthia Dwork and Aaron Roth's *The Algorithmic Foundations of Differential Privacy*.[35]

Criticisms and Limitations

Differential privacy presents the following challenges:

- **Record independence**: Differential privacy does implicitly assume that all information about an individual participant is contained within a record and that records are independent. It is not (necessarily) appropriate to input a data set containing multiple records per person, into a differentially-private mechanism and expect it to work. If records are correlated, the privacy protections may be much weaker than they seem.
- **Cost on utility**: Since before the introduction of differential privacy, it has been known that there are fundamental limitations on using randomness to protect privacy. Dinur and Nisim[36] proved that there is always a tradeoff between utility (i.e. obtaining accurate answers to queries), privacy (i.e. limiting the possible inferences about individuals), and database size (small crowds make privacy

[35] Cynthia Dwork & Aaron Roth, 'The Algorithmic Foundations of Differential Privacy', *Foundations and Trends in Theoretical Computer Science*, Vol. 9, Nos. 3-4, 2014, available at: https://www.cis.upenn.edu/~aaroth/Papers/privacybook.pdf.
[36] Irit Dinur & Kobbi Nissim, *Revealing Information while Preserving Privacy*, 2003: http://www.cse.psu.edu/~ads22/privacy598/papers/dn03.pdf.

hard). These fundamental limitations apply to all randomised privacy treatments including differential privacy.

A number of empirical studies of differentially-private mechanisms have found that privacy came at a high cost to utility. For a given mechanism, this tradeoff is determined by the ε privacy parameter – smaller ε means better privacy, but generally requires larger random perturbations. However, better (high utility) mechanisms are often possible with sufficient care – this is particularly the case for aggregate statistics derived from large databases. Practical examples include movie recommendations,[37] products at Apple[38] and Google,[39] and the U.S. Census Bureau.[40]

Example: NSW Opal transport data

Randomness was added (or subtracted) to the tap-on and tap-off tallies of Transport NSW's open release of Opal data (described on page 12). This meant that even an attacker who knew everything about the totals except the presence or absence of a single person was unable to infer that that person had been present. For example, Figure 3 shows the difference between the tap-on and tap-off tallies of the Manly Ferry over a series of time intervals. We know that the underlying numbers must have been the same, assuming nobody jumped overboard. The graph shows what random perturbations were added by the differential privacy treatment.

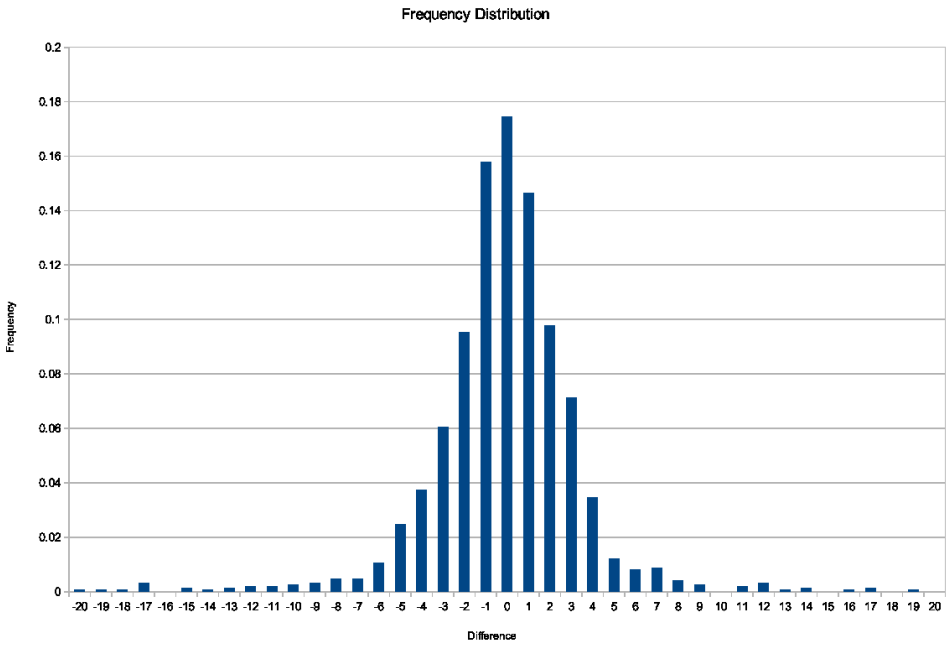


Figure 3: Difference between tap-on and tap-off totals on the Manly Ferry, after differential privacy treatment.[41] The underlying true difference is zero – the graph shows the differences introduced to hide the absence or presence of particular individuals.

[37] Frank McSherry & Ilya Mironov, *Differentially Private Recommender Systems: Building Privacy into the Netflix Prize Contenders*, 2009, available at: https://www.microsoft.com/en-us/research/wp-content/uploads/2009/06/NetflixPrivacy.pdf.
[38] See https://patents.google.com/patent/US9594741B1/en.
[39] Úlfar Erlingsson, Vasyl Pihur & Aleksandra Korolova, *Privacy: Theory meets Practice on the Map*, 2014, available at: https://static.googleusercontent.com/media/research.google.com/en//pubs/archive/42852.pdf.
[40] Ashwin Machanavajjhala, Daniel Kifer, John Abowd, Johannes Gehrke & Lars Vilhuber, *Privacy: Theory meets Practice on the Map*, 2008, available at: http://www.cse.psu.edu/~duk17/papers/PrivacyOnTheMap.pdf.
[41] Graph from Chris Culnane, Benjamin I. P. Rubinstein & Vanessa Teague, *Privacy Assessment of De-identified Opal Data: A report for Transport for NSW*, 2017, available at: https://arxiv.org/abs/1704.08547.

Summary and comparison of methods for removing information from data

Differential privacy offers a provably secure solution against a very powerful and well-defined adversary. In many settings, it works well in practice, especially for simple aggregate statistics derived from large databases. However, it is not a method of securely publishing sensitive and complex unit-record level data.

Weaker treatments such as HIPAA and k-anonymity are highly dependent on assumptions about what information is available to the adversary – if those assumptions are wrong, the data may be easily re-identifiable.

7.2. Methods for restricting access

In the previous part of this report, we noted that there are two main ways of protecting personal information. The first is removing information, which we have discussed above, and the second is restricting access.

When there is a legitimate need for data access (for example by medical researchers who need longitudinal health data), but the data cannot be securely de-identified against a motivated adversary, restricted access to the data may be the only option.

We will not give a complete list of all possible techniques for access control here, but a good system would include:

- electronic access controls (user authentication, encryption of data)
- physical access controls (locked buildings, proper supervision); and
- processes for ethics approval, vetting of researchers, and checking of output data.

Many institutions offer secure research environments, including the UK Administrative Data Research Network,[42] the US Census Bureau's Secure Research Environment,[43] and the Victorian Centre for Data Insights here in Victoria.[44]

It is important to note that a de-identification method good enough for a controlled lab may not defend against a motivated attacker, and may not be secure enough for wider data sharing.

Trusted users

The Productivity Commission's 2017 report, *Data availability and use*, recommends distributing sensitive datasets to 'trusted users'.[45] However, recent news includes two stories that suggest caution about disclosing data sets to trusted users: 50 million users entrusted their data to Facebook, who allowed it to be collected by a researcher at Cambridge University. Somehow it was transferred to a private company

[42] See https://www.adrn.ac.uk/.
[43] See https://www.census.gov/about/adrm/fsrdc/about/secure_rdc.html.
[44] See https://www.vic.gov.au/datainsights.html.
[45] See https://www.pc.gov.au/inquiries/completed/data-access.

and allegedly used for political manipulation.[46] In other news, Telstra's Argus Software included hardcoded passwords which exposed health information.[47]

Trusted users can be a good model for some data sharing. However, these users should be chosen very carefully – they are being trusted for protection of the public interest and for the use of thorough data protection standards.

7.3. Misuse of data without explicit re-identification

Re-identification represents an end-point in privacy invasion. However, it is not necessary to reach that end-point before causing harm to an individual. This issue is more prevalent in the commercial realm than the public sector, but in principle it applies to both. The issue is that harm can be caused by just knowing attributes of an individual, without knowing their identity. Companies are increasingly showing an appetite to undertake sophisticated profiling of their customers, with a view to target marketing or apply price discrimination. There may be noble intentions in doing this – for example, there are use cases in which government might better target assistance to needy demographics – but this should be balanced with potential negative impacts or perceptions of bias through profiling.

In the commercial sphere, for example, if an individual is classified into a group that is considered affluent, a business can potentially charge more for the same service in the knowledge that prospective customer is likely to be able to pay. The classification need not be 100% accurate, it only needs to be right a sufficient number of times to result in increased profits. Coarse location information can be derived from the user's IP address. Even though not personally identifiable on its own, it is sufficient to allow classification of the customer, and for them to be treated differently, potentially to their detriment.

8. Summary and conclusion

Public release of de-identified information is often perceived as a solution to complex questions about balancing individual privacy with efficient data sharing in order to 'unlock value' through analysis outside governments. Unfortunately, there may not be a secure solution that works for all types of data and all applications. Numerous studies have demonstrated the re-identifiability of data that was released in the mistaken belief that it had been de-identified.

The same fallacy might also underpin the sharing of data in a commercial context – data that does not seem to be identifying may nevertheless be re-identifiable, and hence constitute personal information.

Aggregate statistics are amenable to rigorous privacy protection techniques such as differential privacy. Ongoing research is extending differential privacy techniques to other data types too. However, detailed unit-record level data about individuals may need to be protected by more traditional restricted access.

Data custodians must consider very carefully whether de-identification methods truly ensure that a person is not 'reasonably identifiable', or are merely a reasonable attempt at doing so. In some cases, seemingly de-identified information may still be capable of reasonably identifying an individual, and will attract obligations under privacy law.

[46] Nicholas Confessore, 'Cambridge Analytica and Facebook: The Scandal and the Fallout So Far', *NY Times*, 4 April 2018, available at: https://www.nytimes.com/2018/04/04/us/politics/cambridge-analytica-scandal-fallout.html.
[47] See https://www.smh.com.au/technology/medical-records-exposed-by-flaw-in-telstra-health-argus-software-20180322-p4z5ko.html.

9. Further reading

- The Chief Data Officer's de-identification guideline[48] explores 'reasonable steps' for de-identification in the context of the VDS Act.
- The European Union recently released guidelines for the release of public datasets.[49] Step 1 is "Understand the data. Consider potential use cases, the value of the data and potential risks."
- A US presidential commission on cybersecurity received a number of submissions on privacy and data sharing. An MIT submission[50] emphasised 'Open Algorithms' and 'Permissible Use'. Open algorithms means that details about the methods and processes should be available for public scrutiny; permissible use emphasises the consent or expectations of the people whose data is shared.
- The United Nations Special Rapporteur on the right to privacy recently released a report[51] and supporting technical detail[52] on big data and open data.

[48] Department of Premier and Cabinet, *De-identification guideline*, 2018, available at: https://www.vic.gov.au/system/user_files/Documents/vcdi/VCDI de-identification Guidelines_UPDATE as at 21 Feb.pdf.
[49] See https://www.europeandataportal.eu/en/content/how-address-privacy-concerns-when-opening-data.
[50] Alex Pentland, David Shrier, Thomas Hardjono & Irving Wladawsky-Berger, 'Towards an Internet of Trusted Data: A New Framework for Identity and Data Sharing', *MIT Connection Science*, 2016, available at: https://www.nist.gov/document/mitrfiresponsepdf.
[51] See http://www.ohchr.org/Documents/Issues/Privacy/A-72-43103_EN.docx.
[52] See http://www.ohchr.org/Documents/Issues/Privacy/A-72-Slot-43103.docx.