GUIDE TO BASIC ANONYMISATION

SG:DIGITAL

pdpc PERSONAL DATA PROTECTION COMMISSION SINGAPORE

CONTENTS

INTRODUCTION

INTRODUCTION

This guide is meant to provide an introduction and practical guidance to organisations that are new to anonymisation on how to appropriately perform basic anonymisation and de-identification of structured[1], textual[2], non-complex datasets[3]. It presents the anonymisation workflow in the context of four common use cases.

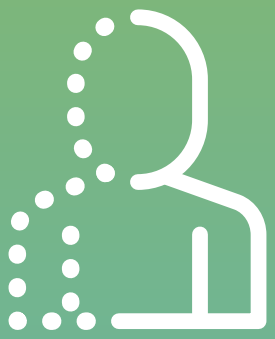
This guide is not exhaustive in dealing with all the issues relating to anonymisation, de-identification and re-identification of datasets. Organisations are advised to consider hiring anonymisation experts, statisticians or independent risk assessors to perform the appropriate anonymisation techniques or assessment of re-identification risks, where anonymisation issues are complex (e.g. large datasets containing a wide range of longitudinal or sensitive personal data).

Implementation of the recommendations in this guide does not imply compliance with the Personal Data Protection Act (PDPA).

Different jurisdictions view anonymisation differently and hence, the recommendations provided in this guide may not apply to data protection laws in other countries.

This guide should be read together with the Personal Data Protection Commission's (PDPC) *Advisory Guidelines on the Personal Data Protection Act for Selected Topics*.

1. "Structured" refers to data in a defined and tabular format, such as a spreadsheet or relational database (e.g. XLSX and CSV).
2. "Textual" refers to text, numbers, dates, etc., that is, alphanumeric data already in digital form. Anonymisation techniques for non-textual data like audio, video, images, biometric data, etc., create additional challenges and require different anonymisation techniques, which are outside the scope of this guide.
3. The approach recommended in this guide applies only to treatment of structured data (i.e. textual data in a tabular form in columns and rows, such as Microsoft Excel spreadsheets). Digital photographs, for example, do not fall under this category of data.



ANONYMISATION VERSUS DE-IDENTIFICATION

ANONYMISATION VERSUS DE-IDENTIFICATION

Anonymisation refers to the conversion of personal data into data that cannot be used to identify any individual. PDPC views anonymisation as a risk-based process, which includes applying both anonymisation techniques and safeguards to prevent re-identification.

De-identification[4] refers to the removal of identifiers (e.g. name, address, National Registration Identity Card (NRIC) number) that directly identify an individual. De-identification is sometimes mistakenly equated to anonymisation, however it is only the first step of anonymisation. A de-identified dataset may easily be re-identified when combined with data that is publicly or easily accessible.

Re-identification refers to the identification of individuals from a dataset that was previously de-identified or anonymised.

Anonymised data is not considered personal data and thus, is not governed by the PDPA. For more information, please refer to the topic on anonymisation in the PDPC's *Advisory Guidelines on the Personal Data Protection for Selected Topics*.

AN EXAMPLE OF DE-IDENTIFICATION

Albert uses food ordering apps frequently. His favourite food ordering app — SuperHungry — decides to publish some information about its users for a hackathon.

Albert's data record at SuperHungry:

Name	Favourite eatery	Favourite food
Albert Phua	Katong Fried Chicken	3-Piece Chicken Set, 33 past orders

Date of birth	Gender	Company
01/01/1990	Male	ABC Pte Ltd

4. The de-identification process may include assigning of pseudonyms.

SuperHungry de-identifies the dataset by removing the names before publishing, thinking that this equates to anonymising the dataset.

Albert's de-identified record published by SuperHungry:

Name	Favourite eatery	Favourite food	Date of birth	Gender	Company
Albert Phua	Katong Fried Chicken	3-Piece Chicken Set, 33 past orders	01/01/1990	Male	ABC Pte Ltd

However, Albert can be re-identified by combining his de-identified record with other records (e.g. personal information from his social media profile).

Albert's social media profile:

Name	Date of birth	Gender	Company
Albert Phua	01/01/1990	Male	ABC Pte Ltd

Any person with sufficient motivation can easily identify[5] the person as Albert from the de-identified data if there are other publicly or easily available information to enable such re-identification. If the dataset or combined dataset is sensitive, further anonymisation will be required.

5. This example re-identifies the record only if this information is unique to Albert in the population.

INTRODUCTION TO BASIC DATA ANONYMISATION CONCEPTS

INTRODUCTION TO BASIC DATA ANONYMISATION CONCEPTS

Data anonymisation requires a good understanding of the following elements, which should be taken into consideration when determining what constitutes suitable anonymisation techniques and appropriate anonymisation levels.

A Purpose of anonymisation and utility

The purpose of anonymisation must be clear, because anonymisation should be done specifically for the purpose at hand. The process of anonymisation, regardless of techniques used, reduces the original information in the dataset by some extent. Hence, as the degree of anonymisation increases, utility (e.g. clarity and/or precision) of the dataset is generally reduced. Therefore, the organisation needs to decide on the degree of the trade-off between acceptable (or expected) utility and the risk of re-identification.

It should be noted that utility should not be assessed at the level of the entire dataset as it is typically different for different attributes. One extreme is that the accuracy of a specific data attribute is crucial and no generalisation or anonymisation technique should be applied (e.g. medical conditions and drugs administered to individuals may be crucial data when analysing the hospital admission trends). The other extreme is that the data attribute is of no use for the intended purpose and may be dropped entirely without affecting the utility of the data to the recipient (e.g. date of birth of individuals may not be important when analysing the purchase transaction trends).

Another important consideration in determining the trade-off between utility and anonymisation is whether it poses an additional risk if the recipient knows which anonymisation techniques and what degree of granularity have been applied; on one hand, knowing this information may help the analyst understand the results and interpret them better, but on the other hand it may contain hints, which could lead to a higher risk of re-identification.

B Reversibility

Typically, the process of data anonymisation would be “irreversible” and the recipient of the anonymised dataset would not be able to recreate the original data. However, there may be cases where the organisation applying the anonymisation retains the ability to recreate the original dataset from the anonymised data; in such cases, the anonymisation process is “reversible”.

C Characteristics of anonymisation techniques

The different characteristics of the various anonymisation techniques mean that certain techniques may be more suitable for a particular situation or data type than others. For instance, certain techniques (e.g. character masking) may be more suitable for use on direct identifiers and others (e.g. aggregation) for indirect identifiers. Another characteristic to consider is whether the attribute value is a continuous value (e.g. height = 1.61m) or discrete value (e.g. "yes" or "no"), because techniques such as data perturbation work much better for continuous values.

The various anonymisation techniques also modify data in significantly different ways. Some modify only part of an attribute (e.g. character masking); some replace the value of an attribute across multiple records (e.g. aggregation); some replace the value of an attribute with an unrelated but unique value (e.g. pseudonymisation); and some remove the attribute entirely (e.g. attribute suppression).

Some anonymisation techniques can be used in combination (e.g. suppressing or removing (outlier) records after generalisation is performed).

D Inferred information

It may be possible for certain information to be inferred from anonymised data. For example, masking may hide personal data, but it does not hide the length of the original value in terms of the number of characters.

Organisations may also wish to consider the order in which the anonymised data is presented. For example, if the recipient knows that the data records were collected in serial order (e.g. registration of visitors as they come), it may be prudent (as long as it does not affect utility) to reshuffle the entire dataset to avoid inference based on order of the data records.

Inference is not limited to a single attribute, but may also apply across attributes even if anonymisation techniques had been applied to all. The anonymisation process must, therefore, take note of every possibility that inference may occur, both before deciding on the actual techniques and after applying the techniques.

E Expertise with the subject matter

Anonymisation techniques basically reduce the identifiability of one or more individuals from the original dataset to a level acceptable by the organisation's risk portfolio.

An identifiability and re-identifiability[6] assessment should be performed before and after anonymisation techniques are applied. This requires a good understanding of the subject matter which the data pertains to. For example, if the dataset is healthcare data, the

6. "Identifiability" refers to the degree to which an individual can be identified from one or more datasets containing direct and indirect identifiers while "re-identifiability" refers to the degree to which an individual can be re-identified from anonymised dataset(s).

organisation would likely require someone with sufficient healthcare knowledge to assess a record's uniqueness (i.e. to what degree it is identifiable or re-identifiable).

The assessment before the anonymisation process ensures that the structure and information within an attribute is clearly identified and understood, and the risk of explicit and implicit inference from such data is assessed. For example, an attribute containing the year of birth implicitly provides age, as does an NRIC number to some extent. The assessment after the anonymisation process will determine the residual risk of re-identification from the anonymised data.

Another instance is when data attributes are swapped between records and it takes a subject-matter expert to recognise if the anonymised records make sense.

The right choice of anonymisation techniques, therefore, depends on awareness of the explicit and implicit information contained in the dataset and the amount or type of information intended to be anonymised.

F Competency in anonymisation process and techniques

Organisations that wish to share anonymised datasets should ensure that the anonymisation process is undertaken by employees who have undergone training and are familiar with anonymisation techniques and principles. If the necessary expertise is not found within the organisation, external help should be engaged.

G The recipient

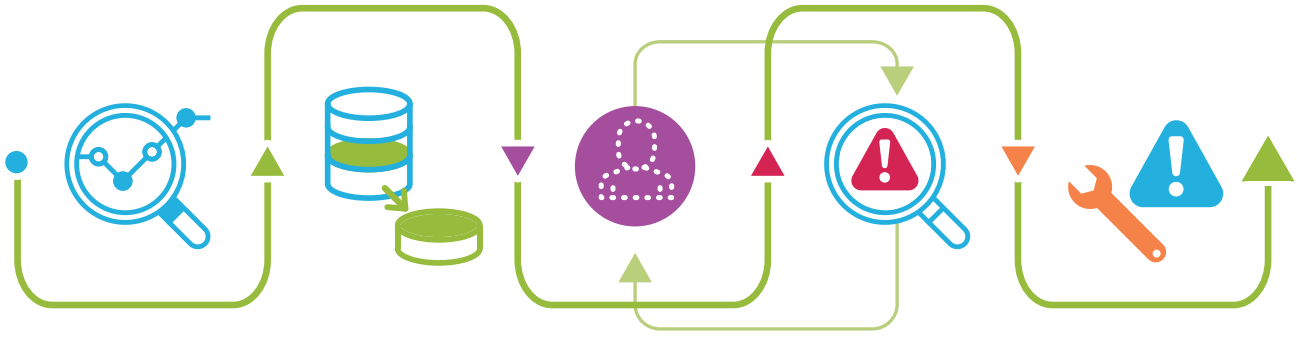
Factors such as the recipients' expertise in the subject matter and controls implemented to limit the quantity of recipients and to prevent the data from being shared with unauthorised parties play an important role in the choice of anonymisation techniques. In particular, the expected use of the anonymised data by the recipient may impose limitations on the applied techniques because the utility of the data may be lost beyond acceptable limits. Extra caution needs to be taken when making public releases of data and organisations will require a much stronger form of anonymisation compared to the data shared under a contractual arrangement.

H Tools

Software tools can be very useful to aid in executing anonymisation techniques. Refer to Annex E for some anonymisation tools that are available in the market.

THE ANONYMISATION PROCESS

THE ANONYMISATION PROCESS



Step 1	Step 2	Step 3	Step 4	Step 5
Know your data	**De-identify your data**	**Apply anonymisation techniques**	**Compute your risk**	**Manage your risks**

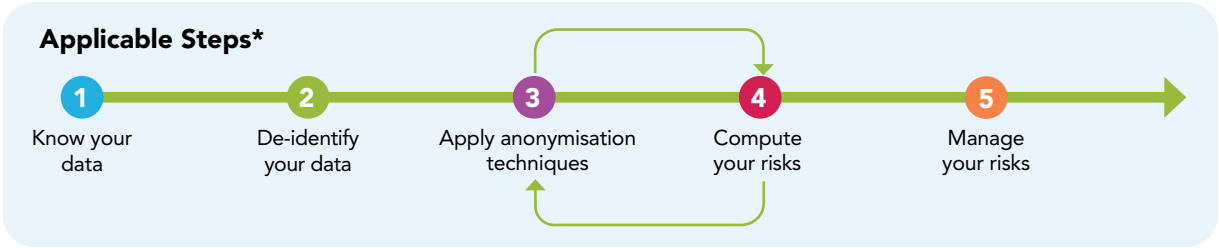
You can use these five steps to anonymise your datasets where appropriate, depending on your use case. In this guide, we explain these steps using five common data use cases by organisations.

In all data use cases, you should ensure:

- Data minimisation, such that only necessary data attributes and an extract (where possible) of your dataset is shared to third parties;
- Any identifying information of the dataset that you are anonymising should not be publicly available (e.g. if you are anonymising information on a membership database, the profiles of your membership base should not be publicly available); and
- The appropriate level of protection and safeguard is given to the anonymised dataset and identity mapping table for pseudonyms to prevent re-identification. Generally, the less you modify a dataset through anonymisation, the more you need to protect and safeguard the dataset as the re-identification risks are higher.

USE CASES: HOW YOU CAN USE ANONYMISED OR DE-IDENTIFIED DATA

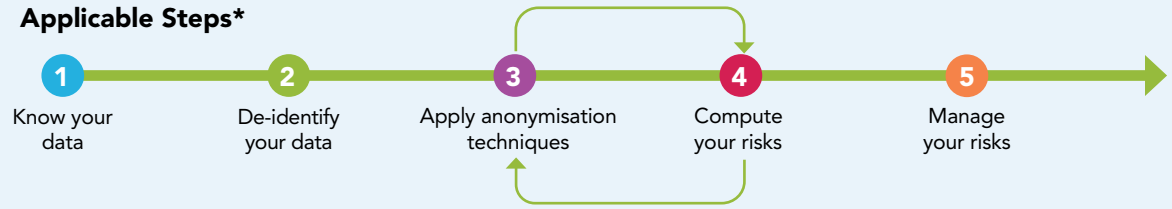
Here are some ways that anonymised or de-identified data can be used in your organisation.



Use case	**Internal data sharing (de-identified data)** (e.g. De-identified customer data shared between sales and marketing departments for analysis and in-house development of targeted marketing campaigns).
Description	Data is only de-identified to support record-level data sharing and use within the organisation, which may require most details in the data to be left untouched. The de-identified data is still personal data as it is likely to be easily re-identifiable. However, it is still good practice to de-identify the data as it provides an additional layer of protection.
Are additional controls needed to prevent re-identification?	Yes
Is the end result considered anonymised data?	No
***Applicable**	1 2 5

Use case	**Internal data sharing (anonymised data)** (e.g. Anonymised data on the demographics of high value consumers and their respective spending patterns shared with loyalty teams to develop differentiated customer value propositions).
Description	Organisations could consider anonymised data instead of de-identified data for internal sharing under the following cases where: • Internal data sharing does not require detailed de-identified personal data (e.g. for trend analysis); • Data involved is more sensitive in nature (e.g. financial information); or • Larger datasets shared with more than one department. In such cases, organisations may apply the anonymisation process suggested for external data sharing to their internal data sharing use case to reduce the risk of re-identification and disclosure.
Are additional controls needed to prevent re-identification?	Yes
Is the end result considered anonymised data?	Yes
***Applicable steps**	1 2 3 4 5

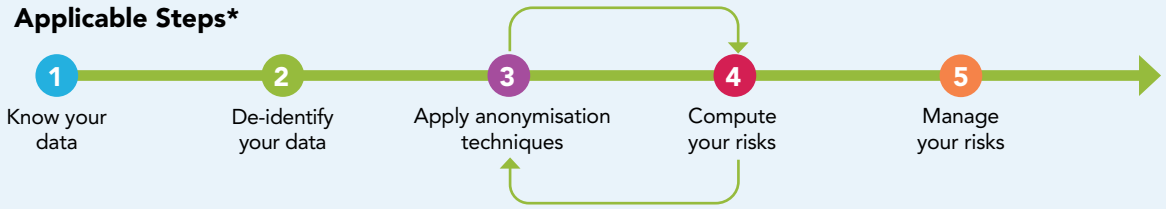
Applicable Steps*



Use case	**External data sharing** (e.g. Anonymised customer data shared between sales department and external business partner for analysis of customer profiles and development of co-branded products).
Description	Record-level data shared with an authorised external party for business collaboration purposes. Anonymisation techniques are used to convert personal data to non-identifying data.
Are additional controls needed to prevent re-identification?	Yes
Is the end result considered anonymised data?	Yes
***Applicable steps**	1 2 3 4 5

Use case	**Long-term data retention** for data analysis (e.g. Historical analysis of customer trends).
Description	Anonymisation techniques are used to convert personal data to non-identifying data, and allow the data to be kept at record-level beyond the retention period for long-term data analysis.
Are additional controls needed to prevent re-identification?	Yes
Is the end result considered anonymised data?	Yes
***Applicable steps**	1 2 3 4 5

Applicable Steps*



Use case	**Synthetic data**[7] for application development and testing purposes, where replication of statistical characteristics of the original data is not required (e.g. Used for testing by outsourced vendor engaged to develop and test payroll application).
Description	Record-level synthetic data can be created from the original data by heavily anonymising all data attributes using the anonymisation techniques in this guide, such that all data attributes are modified very significantly and all records created do not match any individual's record in the original data. In this case, the application of anonymisation techniques would not retain the statistical characteristics of the original data and thus is not suitable for sophisticated purposes such as AI model training or data analytics.
Are additional controls needed to prevent re-identification?	No[8]
Is the end result considered anonymised data?	Yes
***Applicable steps**	1 2 3

Note: In synthetic data, the "fake" direct identifiers used should not relate to an actual person, i.e. a randomly generated NRIC with a randomly generated name should not be the same as the NRIC and name combination of an actual person.

7. Another approach not addressed in this guide is to create synthetic data from scratch. This can be done by randomly generating a dataset that merely fulfils the data format requirements, or by generating a dataset that also retains the statistical characteristics of the original dataset using Artificial Intelligence or other methods.
8. Refer to assumptions in Step 5: Manage your re-identification and disclosure risks.

KNOW YOUR DATA

Applicable to:

- **Internal data sharing (de-identified data)**
- **Internal data sharing (anonymised data) or External data sharing**
- **Long-term data retention**
- **Synthetic data**

A personal data record is made up of data attributes that have varying degrees of identifiability and sensitivity to an individual.

Anonymisation typically involves removal of direct identifiers and modification of indirect identifiers. Target attributes are usually left unchanged, except where the purpose is to create synthetic data. The table and examples below illustrate how a data attribute is typically classified within a data record.

	Direct identifiers	Indirect identifiers	Target attributes
Classification of data attributes in a dataset	These are data attributes that are unique to an individual and can be used as key data attributes to re-identify an individual.	These are data attributes that are not unique to an individual but may re-identify an individual when combined with other information (e.g. a combination of age, gender and postal code).	These are data attributes that contain the main utility of the dataset. In the context of assessing adequacy of anonymisation, this data attribute may be sensitive in nature, and may result in a high potential for adverse effect to an individual when disclosed.
Accessibility of data	These data attributes are usually public or easily accessible.	These data attributes may be public or easily accessible.	These data attributes are usually not public or easily accessible. They cannot be used for re-identification as they are typically proprietary.

Common examples in a dataset	• Name • Email address • Mobile phone number • NRIC number • Passport number • Account number • Birth certificate number • Foreign Identification Number (FIN) • Work Permit number • Social media username	• Age • Gender • Race • Date of birth • Address • Postal code • Job title • Company name • Marital status • Height • Weight • Internet Protocol (IP) address • Vehicle license plate number • In-vehicle Unit (IU) number • Global Positioning System (GPS) location	• Transactions (e.g. purchases) • Salary • Credit rating • Insurance policy • Medical diagnosis • Vaccination status

EXAMPLE 1: CLASSIFICATION OF DATA ATTRIBUTES IN AN EMPLOYEE DATA RECORD

Staff ID	Name	Department	Gender	Date of birth	Start date of service	Employment type
39192	Sandy Thomas	Research & Development	F	08/01/1971	02/03/1997	Part-time
37030	Paula Swenson	Engineering	F	15/05/1976	08/03/2015	Full-time
22722	Bosco Wood	Engineering	M	31/12/1973	30/07/1991	Full-time
28760	Stef Stone	Engineering	F	24/12/1970	18/03/2010	Part-time
13902	Jake Norma	Human Resource	M	15/07/1973	28/05/2012	Part-time

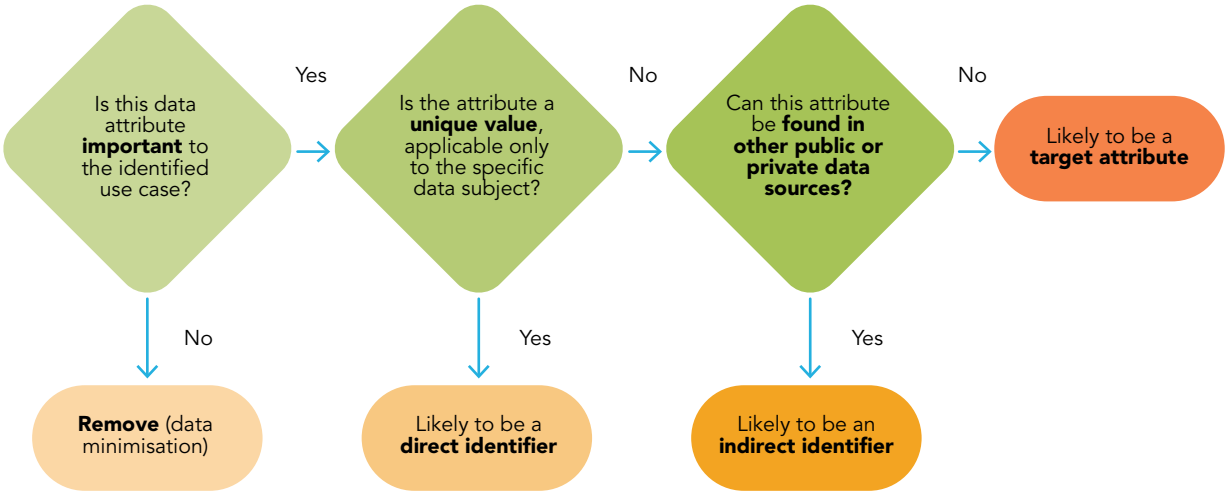
Direct identifiers | Indirect identifiers | Target variables

EXAMPLE 2: CLASSIFICATION OF DATA ATTRIBUTES IN A CUSTOMER DATA RECORD

Customer ID	Name	Gender	Date of birth	Postal code	Occupation	Income	Education	Marital status
56833	Jenny Jefferson	F	05/08/1975	570150	Data scientist	$13,000	Masters	Widowed
50271	Peter G	M	14/12/1973	787589	University lecturer	$12,000	Doctorate	Married
53041	Tim Lake	F	02/03/1985	408600	Researcher	$7,000	Doctorate	Divorced
17290	Remy Bay	M	27/03/1968	570150	Database administrator	$8,000	Bachelor	Married
52388	Walter Paul	M	25/06/1967	199588	Architect	$10,000	Masters	Single

Direct identifiers | Indirect identifiers | Target variables | Indirect identifiers

Any data attribute that is not required in the resultant dataset should be removed as part of data minimisation. A simple flowchart is provided below to assist you in classifying your data attribute appropriately.



STEP 2 DE-IDENTIFY YOUR DATA

Applicable to:

- **Internal data sharing (de-identified data)**
- **Internal data sharing (anonymised data) or External data sharing**
- **Long-term data retention**
- **Synthetic data**

This step is always performed as part of the anonymisation process.

First, remove all direct identifiers. In the following example, all names are removed. Where the dataset includes other direct identifiers such as NRIC number and email address, these should also be removed.

Name	Age	Favourite show
~~Alex~~	25	*The Big Bang Theory*
~~Bosco~~	54	*Friends*
~~Charlene~~	42	*Grey's Anatomy*

Optionally, assign a pseudonym to each record if there is a need to link the record back to a unique individual or to the original record for use cases such as:

a. Data merger;

b. Analysis of multiple records relating to unique individuals; or

c. Generation of synthetic datasets where direct identifier values are required for the development and testing of applications. For this use case, replace all necessary direct identifiers with pseudonyms.

The pseudonyms should be unique for each unique direct identifier (as illustrated below). Assignment of pseudonyms should also be robust (i.e. not be reversible by unauthorised parties through guessing or computing the original direct identifier values from the pseudonyms).

~~Name~~	Token	Age	Favourite show
~~Alex~~	1234	25	*The Big Bang Theory*
~~Bosco~~	5678	54	*Friends*
~~Charlene~~	5432	42	*Grey's Anatomy*

If you want to retain the ability to link the de-identified data record back to the original record at a subsequent point in time, you will need to keep the mapping between the direct identifiers and the pseudonyms. The identity mapping table (illustrated below) should be kept securely as it permits re-identification.

Name	Token
Alex	1234
Bosco	5678
Charlene	5432

STEP 3

APPLY ANONYMISATION TECHNIQUES

Applicable to:

- ○ Internal data sharing (de-identified data)
- ✓ Internal data sharing (anonymised data) or External data sharing
- ✓ Long-term data retention
- ✓ Synthetic data

In this step, you will apply anonymisation techniques to the indirect identifiers so that they cannot be easily combined with other datasets that may contain additional information to re-identify individuals. For the synthetic data use case, anonymisation techniques should also be applied to the target attributes.

Do note that application of these techniques will modify the data values and may affect utility of the anonymised data for some use cases (e.g. data analytics). The anonymisation techniques recommended below take into consideration potential utility required for record-level data in each use case. Organisations may use other anonymisation techniques beyond what is recommended, if relevant to their use case.

Use case	Suggested anonymisation techniques for record-level data
Internal data sharing (anonymised data) **or** **External data sharing**	• **Record suppression:** The removal of a record (i.e. row of data, especially where such data may contain unique data values that cannot be anonymised further). • **Attribute suppression:** The removal of a data attribute (i.e. column of data, especially where such data is not needed in the dataset and may contain unique data values that cannot be anonymised further). • **Character masking:** The replacement of some characters of the data value with a consistent symbol (e.g. * or x). For example, masking a postal code would involve changing it from "235546" to "23xxxx". • **Generalisation:** The reduction in granularity of the data (e.g. by converting a person's age into an age range). For example, generalising the age of a person from "26 years old" to "25-29 years old".

Long-term data retention	• **Record or attribute suppression** • **Character masking** • **Generalisation** • **Data perturbation:** The modification of the values in the data by adding "noise" to the original data (e.g. +/- random values to the data). The degree of perturbation should be proportionate to the range of values of the attribute. For example, data perturbation would involve modifying salary data of an individual from "$256,654" to "$260,000" by rounding the data up to the nearest $10,000. Alternatively, the individual's salary can be modified to "$250,554" by subtracting a random number within $10,000 from its original value. *Note: Data aggregation may also be performed for this use case when record-level data is not required (refer to Annex A for an example).*
Synthetic data	Apply heavy anonymisation to the original data to create synthetic data such that all data attributes (including target attributes) are modified significantly. The resulting dataset and individual records created using this methodology will not have any resemblance to any individual's record and does not retain the characteristics of the original dataset. Because of the resulting dataset's non-resemblance to the original, it is suitable for application development/testing but not AI model training. • **Data perturbation** • **Swapping:** The rearrangement of data in the dataset randomly such that the individual attribute values are still represented in the dataset, but generally do not correspond to the original records.

Refer to Annex A for more information on the various anonymisation techniques and how to apply them. Refer to Annex B for suggested anonymisation techniques to apply on a list of common data attributes.

Next Step: After applying the appropriate anonymisation techniques, proceed to step 4 to assess the risk level. Repeat steps 3 and 4 until you achieve a *k*-anonymity value of 3, 5 or more.

Note: You may also consider removing outlier records or attributes (using record or attribute suppression) that are "resistant" to other anonymisation techniques that have been applied, especially if there is a relatively low count of such outliers and the removal would not significantly impact the quality of the data for your use case.

STEP 4

COMPUTE YOUR RISK

Applicable to:

- ○ Internal data sharing (de-identified data)
- ✓ Internal data sharing (anonymised data) or External data sharing
- ✓ Long-term data retention
- ○ Synthetic data

k-anonymity[9] is an easy method[10,11] to compute the re-identification risk level of a dataset. It basically refers to the smallest number of identical records that can be grouped together in a dataset. The smallest group is usually taken to represent the worst-case scenario in assessing the overall re-identification risk of the dataset. A *k*-anonymity value of 1 means that the record is unique. Generally, only indirect identifiers are considered for *k*-anonymity computation.[12]

A higher *k*-anonymity value means there is a lower risk of re-identification while a lower *k*-anonymity value implies a higher risk. **Generally the industry threshold for *k*-anonymity value is at 3 or 5**.[13] Where possible, a higher *k*-anonymity threshold value should be set to minimise any re-identification risks.

Refer to Chapter 3 (Anonymisation) of PDPC's *Advisory Guidelines on the Personal Data Protection Act for Selected Topics* on the criteria for determining whether the data may be considered sufficiently anonymised.

Postal code	Age	Favourite show	
22xxxx	21 to 25	*Emily in Paris*	*k*=2
22xxxx	21 to 25	*Emily in Paris*	
10xxxx	41 to 45	*Brooklyn Nine-Nine*	*k*=4
10xxxx	41 to 45	*Brooklyn Nine-Nine*	
10xxxx	41 to 45	*Brooklyn Nine-Nine*	
10xxxx	41 to 45	*Brooklyn Nine-Nine*	
58xxxx	56 to 60	*Attenborough's Life in Colour*	*k*=3
58xxxx	56 to 60	*Attenborough's Life in Colour*	
58xxxx	56 to 60	*Attenborough's Life in Colour*	

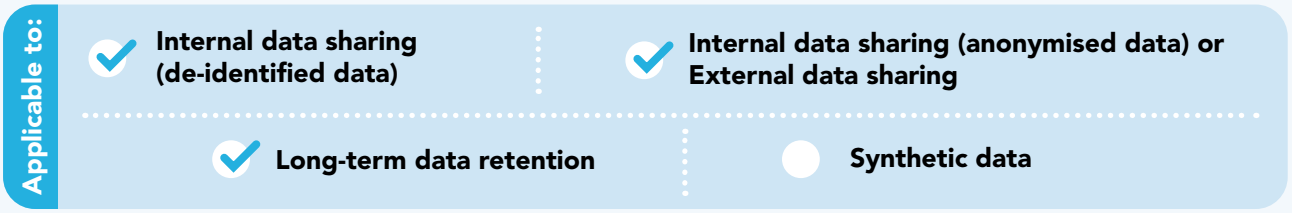
Overall *k*=2

The above diagram illustrates a dataset with three groups of identical records. The *k* value of each group ranges from 2 to 4. Overall, the dataset's *k*-anonymity value is 2, reflecting the lowest value (highest risk) within the entire dataset.[14]

Next Step: If the k -anonymity value threshold is achieved, proceed to step 5. If the k-anonymity value is lower than set threshold, return to step 3 and repeat.

Note: Where possible, you should set a higher k-anonymity value (e.g. 5 or more) for external data sharing, while a lower value (e.g. 3) may be set for internal data sharing or long term data retention. However, if you are not able to anonymise your data further to achieve that, you should put in place more stringent safeguards to ensure that the anonymised data will not be disclosed to unauthorised parties and re-identification risks are mitigated. Alternatively, you may engage experts to provide alternative assessment methods to achieve equivalent re-identification risks.

MANAGE YOUR RE-IDENTIFICATION AND DISCLOSURE RISKS



It is generally prudent to put in appropriate measures to safeguard your data against the risks of re-identification and disclosure. This is in view of future technological advances, as well as unknown datasets that could be used to match against your anonymised dataset and allow re-identification to be performed more easily than expected at the time of anonymisation.

9. More information on *k*-anonymity and how to use *k*-anonymity to assess the risk of re-identification can be found in Annex C and Annex D.
10. *k*-anonymity may not be suitable for all types of datasets or other complex use cases (e.g. longitudinal or transactional data where the same indirect identifiers may appear in multiple records). Special Uniques Detection Algorithms (SUDA) and μ-Argus are other approaches/tools to assess the risk of shared datasets.
11. A known limitation of using *k*-anonymity is attribute disclosure via homogeneity attacks which can be addressed using the *k*-anonymity extensions *l*-diversity and *t*-closeness. These topics are outside the scope of this guide.
12. Direct identifiers should have been removed in Step 2 and pseudonyms should not be included in the computation; otherwise, every record would be unique.
13. Reference from *The De-identification Decision-Making Framework* by Office of the Australian Information Commissioner, CSIRO and Data 61.
14. The guide adopts the more conservative approach of looking at the maximum risk. There are also other approaches (e.g. average risk and strict average risk).

As good practice, the details of the anonymisation process, parameters used and controls should also be clearly recorded for future reference. Such documentation facilitates review, maintenance, fine-tuning and audits. Note that such documentation should be kept securely as the release of the parameters may facilitate re-identification and disclosure of the anonymised data.

There are various types of re-identification and disclosure risks. The following explains some fundamental ones that you should assess when reviewing the sufficiency of protection measures that have been put in place.

1 Re-identification (Identity disclosure)

Determining, with a high level of confidence, the identity of an individual described by a specific record. This could arise from scenarios such as insufficient anonymisation, re-identification by linking or pseudonym reversal. For example, an anonymisation process which creates pseudonyms based on an easily guessable and reversible algorithm, such as replacing "1" with "a", "2" with "b" and so on.

2 Attribute disclosure

Determining, with a high level of confidence, that an attribute described in the dataset belongs to a specific individual even if the individual's record cannot be distinguished. Take, for example, a dataset containing anonymised client records of a particular aesthetic surgeon that reveals all his clients below the age of 30 have undergone a particular procedure. If it is known that a particular individual is 28 years old and is a client of this surgeon, we then know that this individual has undergone the particular procedure, even if the individual's record cannot be distinguished from others in the anonymised dataset.

3 Inference disclosure

Making an inference, with a high level of confidence, about an individual even if he or she is not in the dataset by statistical properties of the dataset. For example, if a dataset released by a medical researcher reveals that 70% of individuals above the age of 75 have a certain medical condition, this information could be inferred about an individual who is not in the dataset.

In general, most traditional anonymisation techniques aim to protect against re-identification and not necessarily other types of disclosure risks.

The following table explains when measures against re-identification and disclosure risks are recommended. A set of basic protection measures (technical, process and legal controls) for the use cases are outlined in the following paragraphs.

Use case	Do you need to manage re-identification and disclosure risks for de-identified or anonymised datasets?
Internal data sharing (de-identified data)	As only de-identification has been applied in order to retain high data utility, re-identification and disclosure risk for de-identified data is higher. Hence, protection is required for the de-identified dataset. The identity mapping tables, if any, should be secured. In the event of a data breach, application of de-identification techniques, how the de-identified dataset is protected and how the mapping table is secured would all be considered part of the protection mechanisms implemented.
Internal data sharing (anonymised data)	To lower re-identification and disclosure risks, anonymisation should be applied to the data for internal sharing, where necessary, in the following cases. They are (a) where detailed personal data is not required, (b) where sensitive data may be shared or (c) where a large dataset is shared with more than one department. Basic protection is required for the anonymised dataset. The identity mapping tables, if any, should be secured and not be shared with the other internal departments.
External data sharing	Basic protection is required for the anonymised dataset. The identity mapping tables, if any, should be secured and not be shared externally.
Long-term data retention	Basic protection is required for the anonymised dataset. All identity mapping tables are to be securely destroyed.

For the synthetic data use case, re-identification risks are assumed to be minimal when anonymisation is applied heavily to all indirect identifiers and target attributes such that the records do not resemble the original dataset. As such, no further protection of this dataset is required.

Technical and process controls: You should implement technical protection measures to manage the re-identification and disclosure risk of de-identified and anonymised data. Some good practices are suggested in the following table.

You should review these good practices to determine if they are sufficient to protect your de-identified/anonymised data based on the degree of anonymisation applied, sensitivity of the de-identified/anonymised data and the use case. You may refer to the PDPC's *Guide to Data Protection Practices for ICT Systems* for additional protection measures where relevant.

In the table, "*Y*" means you are recommended to adopt the corresponding technical control and "*NA*" means that particular technical control is not applicable to that use case.

	Technical control	Internal data sharing (de-identified data)	Internal data sharing (anonymised data)	External data sharing	Long-term data retention
Access control and passwords	Implement access control at the application level to restrict data access to a user level. Minimum level of password complexity (i.e. minimum 12 alphanumeric characters with a mix of uppercase, lowercase, numeric and special characters).	Y	Y	Y	Y
	Regularly review user accounts to ensure all the accounts are active and the rights assigned are necessary (e.g. remove user accounts when a user has left the organisation or update the user's rights when he or she has changed his or her role within the organisation).	Y	Y	Y	Y

	Technical control	Internal data sharing (de-identified data)	Internal data sharing (anonymised data)	External data sharing	Long-term data retention
Security for storage devices/databases	Protect computers by using password functions. Examples of these include keying in password during boot-up, requiring login to the operating system, locking the screen after a period of inactivity, etc.	Y	Y	Y	Y
	Encrypt the dataset. Review the method of encryption (e.g. algorithm and key length) periodically to ensure that it is recognised by the industry as relevant and secure.	Y *(where the data involved is sensitive in nature or larger datasets is shared with more than one department but anonymisation is not applied to the dataset.)*	NA	NA *(where the re-identification risk is assessed to be low (e.g. k-anonymity is 5 or more), encryption need not be applied to the anonymised dataset.)*	NA
	Encrypt the identity mapping tables. Identity mapping tables should be secured and not be shared in all use cases.	Y	Y	Y	NA *(Identity mapping tables should be removed.)*
	Communicate the decryption key of the dataset separately to the target recipient of the shared/ exported data.	Y	NA	NA	NA

	Process control	Internal data sharing (de-identified data)	Internal data sharing (anonymised data)	External data sharing	Long-term data retention
Incident management	Develop a data breach management plan to respond to data breaches and manage the loss of datasets more effectively. The plan should also include how to manage the loss of identity mapping tables or information that could allow reversing de-identified/anonymised data back to its original form, resulting in the lost data being re-identified. Refer below for more information on incident management.	Y	Y	Y	Y
Internal governance controls	Keep a central registry of all shared de-identified/anonymised data to ensure that the combined shared data will not result in re-identification of the de-identified/anonymised data.	Y	Y	Y	NA
	Periodically conduct re-identification reviews of the de-identified/ anonymised data.	Y	Y	Y	Y
	Ensure that the recipient (individual or department) and the purpose of using the de-identified/anonymised data have been approved by relevant authorities within the organisation.	Y	Y	NA	NA
	Prohibit the authorised recipient (individual or department) from sharing de-identified/anonymised data to any unauthorised parties or attempting to re-identify the data without approval from relevant authorities within the organisation.	Y	Y	NA	NA
	Regularly purge de-identified/ anonymised data within the organisation when its purpose has been fulfilled and there is no longer any need for the data.	Y	Y	NA	NA
	Periodically conduct internal checks/audits to ensure compliance with processes.	Y	Y	Y	Y

Incident Management: Organisations should identify the risks of data breaches[18] involving identity mapping table, de-identified data and anonymised data, and incorporate relevant scenarios into their incident management plans. The following considerations may be relevant for data breach reporting and internal investigations:

Loss of de-identified data and identity mapping table

Breach of both de-identified data and identity management table will be akin to the breach of personal data. In such an event, the organisation must assess whether a data breach is notifiable and notify the affected individuals and/or the Commission, where it is assessed to be notifiable under the Data Breach Notification obligation.

Loss of de-identified data only

If de-identified data has been breached externally, an assessment is necessary. The organisation must assess whether a data breach is notifiable as de-identified data has a higher risk of re-identification. However, the use of de-identification and other safeguards to protect the data and identity mapping table could be considered part of the protection mechanisms implemented by the organisation.

Loss of anonymised data and identity mapping table

Organisations have to assess the risk of re-identification. Where it is determined to be high, organisations must then determine whether a data breach is notifiable and notify the affected individuals and/or the Commission, where it is assessed to be notifiable under the Data Breach Notification obligation.

Loss of anonymised data only

Where the organisation has applied the anonymisation techniques properly, it need not report the breach as a notifiable breach. However, it should still proceed to investigate the incident to understand the cause to improve its internal safeguards against future data breach incidents.

Loss of identity mapping only

If the datasets that the identity mapping table was used for are still protected, organisations need not report the breach as an identity mapping table on its own is not personal data. However, the organisation should immediately generate new pseudonyms for its datasets and a new identity mapping table. It should also proceed to investigate the incident to understand the cause to improve its internal safeguards against future data breach incidents.

Legal controls: Organisations should protect themselves by ensuring that third-party recipients of their anonymised data incorporate relevant protection to the shared anonymised data to minimise re-identification risks. The good practices in the following table are taken from the PDPC's *Trusted Data Sharing Framework.*

	Legal control	Internal data sharing (de-identified data)	Internal data sharing (anonymised data)	External data sharing	Long-term data retention
Data sharing agreement	Ensure the data is only used for permitted purposes (e.g. no disclosure to unauthorised parties) and liability is allocated for contract breaches.	NA	NA	Y	NA
	Prohibit third-party recipients from attempting to re-identify anonymised datasets that have been shared.	NA	NA	Y	NA
	Ensure third-party recipients comply with the relevant protection on the shared anonymised data as per organisation's internal controls.	NA	NA	Y	NA

ANNEX

ANNEX A: BASIC DATA ANONYMISATION TECHNIQUES

Record Suppression	
Description	Record suppression refers to the removal of an entire record in a dataset. In contrast to most other techniques, this technique affects multiple attributes at the same time.
When to use it	Record suppression is used to remove outlier records which are unique or do not meet other criteria, such as *k*-anonymity, from the anonymised dataset. Outliers can lead to easy re-identification. It can be applied before or after other techniques (e.g. generalisation) have been applied.
How to use it	Delete the entire record. Note that the suppression should be permanent and not just a "hide row"[15] function; similarly, "redacting" may not be sufficient if the underlying data remains accessible.
Other tips	• Refer to the example in the section on generalisation for illustration of how record suppression is used. • Note that removal of records can impact the dataset (e.g. in terms of statistics such as average and median).

Character Masking	
Description	Character masking refers to changing the characters of a data value. This can be done by using a consistent symbol (e.g. "*" or "x"). Masking is typically applied only to some characters in the attribute.
When to use it	Character masking is used when the data value is a string of characters and hiding part of it is sufficient to provide the extent of anonymity required.
How to use it	Depending on the nature of the attribute, replace the appropriate characters with a chosen symbol. Depending on the attribute type, you may decide to replace a fixed number of characters (e.g. for credit card numbers) or a variable number of characters (e.g. for email address).
Other tips	• Note that masking may need to take into account whether the length of the original data provides information about the original data. Subject matter knowledge is critical, especially for partial masking to ensure that the right characters are masked. Special consideration may also apply to checksums within the data; sometimes, a checksum may be used to recover (other parts of) the masked data. As for complete masking, the attribute could alternatively be suppressed unless the length of the data is of some relevance. • The scenario of masking data in such a way that data subjects are meant to recognise their own data is a special one, and does not belong to the usual objectives of data anonymisation. One example of this is the publishing of lucky draw results, where the names and partially masked NRIC numbers of lucky draw winners are typically published for the individuals to recognise themselves as winners. Another example is information such as an individual's credit card number being masked in an app or a statement addressed to the individual. Note that generally, anonymised data should not be recognisable even to the data subject themselves.

15. This refers to using the "hide row" function in your spreadsheet software.

EXAMPLE

This example shows an online grocery store conducting a study of its delivery demand from historical data to improve operational efficiency. The company masked out the last 4 digits of the postal codes, leaving the first 2 digits, which correspond to the "sector code" within Singapore.

Before anonymisation:

Postal code	Favourite delivery time slot	Average number of orders per month
100111	8 pm to 9 pm	2
200222	11 am to 12 noon	8
300333	2 pm to 3pm	1

After partial masking of postal code:

Postal code	Favourite delivery time slot	Average number of orders per month
10xxxx	8 pm to 9 pm	2
20xxxx	11 am to 12 noon	8
30xxxx	2 pm to 3pm	1

Pseudonymisation	
Description	Pseudonymisation refers to the replacement of identifying data with made-up values. It is also referred to as coding. Pseudonyms can be irreversible when the original values are disposed of properly and the pseudonymisation is done in a non-repeatable fashion. They can also be reversible (by the owner of the original data) when the original values are securely kept, but can be retrieved and linked back to the pseudonym should the need arise[16]. Persistent pseudonyms allow linking by using the same pseudonym values to represent the same individual across different datasets. However, different pseudonyms may be used to represent the same individual in different datasets to prevent linking of the different datasets. Pseudonyms can also be randomly or deterministically generated.
When to use it	Pseudonymisation is used when data values need to be uniquely distinguished and no character or any other implied information about the direct identifiers of the original attribute are kept.
How to use it	Replace the respective attribute values with made-up values. One way to do this is to pre-generate a list of made-up values and randomly select from this list to replace each of the original values. The made-up values should be unique and should have no relationship to the original values (such that one can derive the original values from the pseudonyms).

16. For example, in the event that a research study yields results that can provide a useful warning to a data subject.

Other tips

- When allocating pseudonyms, ensure not to re-use pseudonyms that have already been utilised in the same dataset, especially when they are randomly generated. Also, avoid using the exact same pseudonym generator over several attributes without a change (e.g. at least use a different random seed).
  - Persistent pseudonyms usually provide better utility by maintaining referential integrity across datasets.
  - For reversible pseudonyms, the identity mapping table cannot be shared with the recipient; it should be securely kept and can only be used by the organisation where it is necessary to re-identify the individual(s).
- Similarly, if encryption or a hash function is used to pseudonymise the data, the encryption key or hash algorithm and salt value for the hash must be securely protected from unauthorised access. This is because a leak of such information could result in a data breach by enabling the reversal of the encryption or using pre-computed tables to infer the data that was hashed (especially for data that follows pre-determined formats such as in NRICs).

  The same applies for pseudo-random number generators, which require a seed. The security of any key used must be ensured like with any other type of encryption or reversible process[17]. Organisations should also review the method of encryption (e.g. algorithm and key length) and hash function periodically to ensure that it is recognised by the industry as relevant and secure.
- In some cases, pseudonyms may need to follow the structure or data type of the original value (e.g. for pseudonyms to be usable in software applications); in such cases, special pseudonym generators may be needed to create synthetic datasets or in some cases, so-called "format preserving encryption" can be considered, which creates pseudonyms that have the same format as the original data.

EXAMPLE

This example shows pseudonymisation being applied to the names of persons who obtained their driving licences and some information about them. In this example, the names were replaced with pseudonyms instead of the attribute being suppressed because the organisation wanted to be able to reverse the pseudonymisation if necessary.

Before anonymisation:

Person	Pre-assessment result	Hours of lessons taken before passing
Joe Phang	A	20
Zack Lim	B	26
Eu Cheng San	C	30
Linnie Mok	D	29
Jeslyn Tan	B	32
Chan Siew Lee	A	25

17. Note that relying on a proprietary or "secret" reversal process (with or without a key) has a greater risk of being decoded and broken compared to using a standard key-based encryption or hashing.

After pseudonymising the "Person" attribute:

Person	Pre-assessment result	Hours of lessons taken before passing
416765	A	20
562396	B	26
964825	C	30
873892	D	29
239976	B	32
943145	A	25

For reversible pseudonymisation, the identity mapping table is securely kept in case there is a legitimate future need to re-identify individuals. Security controls (including administrative and technical ones) should also be used to protect the identity mapping table.

Identity mapping table (Single coding):

Pseudonym	Person
416765	Joe Phang
562396	Zack Lim
964825	Eu Cheng San
873892	Linnie Mok
239976	Jeslyn Tan
943145	Chan Siew Lee

For added security regarding the identity mapping table, double coding can be used. Following from the previous example, this example shows the additional linking table, which is placed with a trusted third party. With double coding, the identity of the individuals can only be known when both the trusted third party (who has the linking table) and the organisation (which has the identity mapping table) put their data together.

After anonymisation:

Person	Pre-assessment result	Hours of lessons taken before passing
373666	A	20
594824	B	26
839933	C	30
280074	D	29
746791	B	32
785282	A	25

Linking table (Securely kept by a trusted third party only and even the organisation will remove it eventually. The third party is not given any other information):

Pseudonym	Interim pseudonym
373666	OQCPBL
594824	ALGKTY
839933	CGFFNF
280074	BZMHCP
746791	RTJYGR
785282	RCNVJD

Identity mapping table (Securely kept by the organisation)

Interim pseudonym	Person
OQCPBL	Joe Phang
ALGKTY	Zack Lim
CGFFNF	Eu Cheng San
BZMHCP	Linnie Mok
RTJYGR	Jeslyn Tan
RCNVJD	Chan Siew Lee

Note: In both the linking table and identity mapping table, it is good practice to scramble the order of the records rather than leave it in the same order as the dataset. In this example, the records in both tables are left in the original order for easier visualisation.

Generalisation	
Description	Generalisation is a deliberate reduction in the precision of data. Examples include converting a person's age into an age range or a precise location into a less precise location. This technique is also referred to as recoding.
When to use it	Generalisation is used for values that can be generalised and still be useful for the intended purpose.
How to use it	Design appropriate data categories and rules for translating data. Consider suppressing any records that still stand out after the translation (i.e. generalisation).
Other tips	• Choose an appropriate data range. A data range that is too large may mean significant loss in data utility, while a data range that is too small may mean that the data is hardly modified and therefore, still easy to re-identify. If *k*-anonymity is used, the *k* value chosen will affect the data range as well. Note that the first and the last range may be a larger range to accommodate the typically lower number of records at these ends; this is often referred to as top/bottom coding.

EXAMPLE

In this example, the dataset contains the person's name (which has already been pseudonymised), their age in years and residential address.

Before anonymisation:

Serial number	Person	Age	Address
1	357703	24	700 Toa Payoh Lorong 5
2	233121	31	800 Ang Mo Kio Avenue 12
3	938637	44	900 Jurong East Street 70
4	591493	29	750 Toa Payoh Lorong 5
5	202626	23	5 Tampines Street 90
6	888948	75	1 Stonehenge Road
7	175878	28	10 Tampines Street 90
8	312304	50	50 Jurong East Street 70
9	214025	30	720 Toa Payoh Lorong 5
10	271714	37	830 Ang Mo Kio Avenue 12
11	341338	22	15 Tampines Street 90
12	529057	25	18 Tampines Street 90
13	390438	39	840 Ang Mo Kio Avenue 12

For the "Age" attribute, the approach taken is to generalise into the following age ranges.

< 20	21-30	31-40	41-50	51-60	> 60

For the "Address", one possible approach is to remove the block/house number and retain only the road name.

After generalisation of the "Age" and "Address" attributes:

Serial number	Person	Age	Address
1	357703	21-30	Toa Payoh Lorong 5
2	233121	31-40	Ang Mo Kio Avenue 12
3	938637	41-50	Jurong East Street 70
4	591493	21-30	Toa Payoh Lorong 5
5	202626	21-30	Tampines Street 90
6	888948	> 60	Stonehenge Road
7	175878	21-30	Tampines Street 90
8	312304	41-50	Jurong East Street 70
9	214025	21-30	Toa Payoh Lorong 5
10	271714	31-40	Ang Mo Kio Avenue 12
11	341338	21-30	Tampines Street 90
12	529057	21-30	Tampines Street 90
13	390438	31-40	Ang Mo Kio Avenue 12

As an example, assume there is, in fact, only one residential unit on Stonehenge Road. The exact address can be derived even though the data has gone through generalisation. This could be considered "too unique".

Hence, as the next step of generalisation, record 6 could be removed (i.e. using the record suppression technique) as the address is still "too unique" after removing the unit number. Alternatively, all the addresses could be generalised to a greater extent (e.g. town or district) such that suppression is not needed. However, this may affect the utility of the data much more than suppressing a few records from the dataset.

Swapping	
Description	The purpose of swapping is to rearrange data in the dataset such that the values of individual attributes are still represented in the dataset but generally do not correspond to the original records. This technique is also referred to as shuffling and permutation.
When to use it	Swapping is used when subsequent analysis only needs to look at aggregated data or analysis is at the intra-attribute level; in other words, there is no need for analysis of relationships between attributes at the record-level.
How to use it	First, identify which attributes to swap. Then, for each value in the attribute, swap or reassign the value to other records in the dataset.
Other tips	• Assess and decide which attributes (columns) need to be swapped. Depending on the situation, organisations may decide that, for instance, only attributes (columns) containing values that are relatively identifiable need to be swapped.

EXAMPLE

In this example, the dataset contains information about customer records for a business organisation.

Before anonymisation:

Person	Job title	Date of birth	Membership type	Average visits per month
A	University lecturer	3 Jan 1970	Silver	0
B	Salesman	5 Feb 1972	Platinum	5
C	Lawyer	7 Mar 1985	Gold	2
D	IT professional	10 Apr 1990	Silver	1
E	Nurse	13 May 1995	Silver	2

After anonymisation:
In this example, all values for all attributes have been swapped.

Person	Job title	Date of birth	Membership type	Average visits per month
A	Lawyer	10 Apr 1990	Silver	1
B	Nurse	7 Mar 1985	Silver	2
C	Salesman	13 May 1995	Platinum	5
D	IT professional	3 Jan 1970	Silver	2
E	University lecturer	5 Feb 1972	Gold	0

Note: On the other hand, if the purpose of the anonymised dataset is to study the relationships between job profile and consumption patterns, other methods of anonymisation may be more suitable (e.g. generalisation of job titles, which could result in "university lecturer" being modified to become "educator").

Data perturbation	
Description	The values from the original dataset are modified to be slightly different.
When to use it	Data perturbation is used for indirect identifiers (typically numbers and dates), which may potentially be identifiable when combined with other data sources but slight changes in value are acceptable for the attribute. This technique should not be used where data accuracy is crucial.
How to use it	It depends on the exact data perturbation technique used. These include rounding and adding random noise. The example in this section shows base-x rounding.
Other tips	• The degree of perturbation should be proportionate to the range of values of the attribute. If the base is too small, the anonymisation effect will be weaker; on the other hand, if the base is too large, the end values will be too different from the original and utility of the dataset will likely be reduced. • Note that where computation is performed on attribute values that have been perturbed before, the resulting value may experience perturbation to an even larger extent.

EXAMPLE

In this example, the dataset contains information to be used for research on the possible link between a person's height, weight, age, whether the person smokes and whether the person has "disease A" and/or "disease B". The person's name has already been pseudonymised.

The following rounding is then applied:

Attribute	Anonymisation technique
Height (in cm)	Base-5 rounding (5 is chosen, being somewhat proportionate to the typical height value of 120 to 190 cm).
Weight (in kg)	Base-3 rounding (3 is chosen, being somewhat proportionate to the typical weight value of 40 to 100 kg).
Age (in years)	Base-3 rounding (3 is chosen, being somewhat proportionate to the typical age value of 10 to 100 years).
The remaining attributes	Nil, because they are non-numerical and difficult to modify without substantial change in value.

Dataset before anonymisation:

Person	Height (cm)	Weight (kg)	Age (years)	Smokes?	Disease A?	Disease B?
198740	160	50	30	No	No	No
287402	177	70	36	No	No	Yes
398747	158	46	20	Yes	Yes	No
498732	173	75	22	No	No	No
598772	169	82	44	Yes	Yes	Yes

Dataset after anonymisation:

Person	Height (cm)	Weight (kg)	Age (years)	Smokes?	Disease A?	Disease B?
198740	160	51	30	No	No	No
287402	175	69	36	No	No	Yes
398747	160	45	18	Yes	Yes	No
498732	175	75	21	No	No	No
598772	170	81	42	Yes	Yes	Yes

Note: For base-x rounding, the attribute values to be rounded are rounded to the nearest multiple of x.

Data aggregation	
Description	Data aggregation refers to the conversion of a dataset from a list of records to summarised values.
When to use it	It is used when individual records are not required and aggregated data is sufficient for the purpose.
How to use it	A detailed discussion of statistical measures is beyond the scope of this guide, however typical ways include using totals or averages, etc. It may also be also useful to discuss with the data recipient about the expected utility and find a suitable compromise.
Other tips	• Where applicable, watch out for groups having too few records after performing aggregation. In the below example, if the aggregated data includes a single record in any of the categories, it could be easy for someone with some additional knowledge to identify a donor. • Hence, aggregation may need to be applied in combination with suppression. Some attribute may need to be removed, as they contain details that cannot be aggregated and new attributes may need be added (e.g. to contain the newly computed aggregate values).

EXAMPLE

In this example, a charity organisation has records of donations made, as well as some information about the donors.

The charity organisation assessed that aggregated data is sufficient for an external consultant to perform data analysis, hence performed data aggregation on the original dataset.

Original dataset:

Donor	Monthly income ($)	Amount donated in 2016 ($)
Donor A	4000	210
Donor B	4900	420
Donor C	2200	150
Donor D	4200	110
Donor E	5500	260
Donor F	2600	40
Donor G	3300	130
Donor H	5500	210
Donor I	1600	380
Donor J	3200	80
Donor K	2000	440
Donor L	5800	400
Donor M	4600	390
Donor N	1900	480
Donor O	1700	320
Donor P	2400	330
Donor Q	4300	390
Donor R	2300	260
Donor S	3500	80
Donor T	1700	290

Anonymised dataset:

Monthly Income ($)	Number of donations received (2016)	Sum of amount donated in 2016 ($)
1000-1999	4	1470
2000-2999	5	1220
3000-3999	3	290
4000-4999	5	1520
5000-6000	3	870
Grand Total	**20**	**5370**

ANNEX B: COMMON DATA ATTRIBUTES AND SUGGESTED ANONYMISATION TECHNIQUES

Direct identifiers

The following table provides suggestions on anonymisation techniques that can be applied to some common types of direct identifiers. Generally, direct identifiers should be suppressed (removed) or pseudonymised. If assigning of pseudonyms is required, usually one set (i.e. one column) of pseudonyms per dataset is sufficient.

For the synthetic data use case, all direct identifier columns can be retained but must be replaced with pseudonymised values.

Record suppression	Commonly used technique	Example	
		Before	After
• Name • Email address • Mobile phone number • NRIC number • Passport number • Account number • Birth certificate number • Foreign Identification Number (FIN) • Work permit number	**Attribute suppression**	John Tan	(Deleted)
	Assignation of pseudonyms, for example:		
	• Replace direct identifier values with unique random values; or	John Tan	123456
	• Replace direct identifier values with randomly generated values that follow the format of the data.	John.tan@gmail.com S8822311H	123456@abc.com S8512345A

Indirect identifiers

The following table provides suggestions on anonymisation techniques that can be applied to some common types of indirect identifiers. You should choose to apply one or more of the techniques to each indirect identifier (e.g. apply generalisation and swapping to age, based on your use case).

For the synthetic data use case, two useful techniques are data swapping and data perturbation. These apply to all indirect identifiers.

Indirect identifier(s)	Commonly used technique(s)	Example(s)	
		Before	After
• Age • Height • Weight	**Generalisation:** Generalise the age/height/weight to ranges of 5 or 10 years/cm/kg.	Record #1: 24 Record #2: 39 Record #3: 18	**Generalisation** (age range of 5 years): Record #1: 21 to 25 Record #2: 36 to 40 Record #3: 16 to 20
	Data perturbation: Add random values (+/- 5) to the original value.		**Data perturbation:** Record #1: 25 Record #2: 36 Record #3: 17
	Swapping: Randomly switch the age/height/weight associated with each record.		**Swapping:** Record #1: 39 Record #2: 18 Record #3: 24
• Gender	This indirect data attribute typically only has two generic non-identifying values—M or F and thus, it is generally safe to retain as it is. For the synthetic data use case, the following technique may be further applied to this attribute. **Swapping:** Randomly switch the gender within the dataset.	Record #1: M Record #2: M Record #3: F Record #4: M	**Swapping:** Record #1: M Record #2: F Record #3: M Record #4: M
• Race • Marital status	**Generalisation:** Depending on your dataset, you may combine and generalise selected ethnic groups or marital statuses into a category labelled "Others". This is to be done if there are unique ethnic groups/ marital statuses or too few of the same ethnic groups/marital statuses within your dataset.	Record #1: Indian Record #2: Chinese Record #3: Chinese Record #4: Malay Record #5: Eurasian	**Generalisation:** Record #1: Others Record #2: Chinese Record #3: Chinese Record #4: Others Record #5: Others
	Swapping: Randomly switch the race or marital status within the dataset.		**Swapping:** Record #1: Malay Record #2: Chinese Record #3: Indian Record #4: Eurasian Record #5: Chinese

• Date of Birth	**Generalisation:** Generalise the date of birth to year, or month and year.	Record #1: 1 Feb 2003 Record #2: 15 Aug 1990 Record #3: 30 Dec 1998	**Generalisation** (month and year): Record #1: Feb 2003 Record #2: Aug 1990 Record #3: Dec 1998
	Data perturbation: Randomly modify the date (e.g. +/- 30 days from the original date).		**Data perturbation:** Record #1: 20 Jan 2003 Record #2: 18 Aug 1990 Record #3: 6 Jan 1999
	Swapping: Randomly switch the dates within the dataset.		**Swapping:** Record #1: 30 Dec 1998 Record #2: 1 Feb 2003 Record #3: 15 Aug 1990
• Address	**Generalisation:** Generalise the address to pre-defined zones (e.g. with reference to the Urban Redevelopment Authority's (URA) Master Plan[18]).	71 Punggol Central, Singapore 828755	**Generalisation:** Punggol
	Swapping: Randomly switch addresses within the dataset. *Note: For addresses, unit numbers may be identifying. Where not required, unit numbers should be removed from the dataset.*	Record #1: 71 Punggol Central, #10-1122, Singapore 828755 Record #2: 35 Mandalay Road, #13-37 Singapore 208215	**Swapping:** Record #1: 35 Mandalay Road, #13-37 Singapore 208215 Record #2: 71 Punggol Central, #10-1122, Singapore 828755
• Postal code	**Character masking:** Mask the last four digits of the postal code. (Singapore has 80 postal districts).	117438	**Character masking:** 11xxxx
	Swapping: Randomly switch the postal codes within the dataset.	Record #1: 117438 Record #2: 828755	**Swapping:** Record#1: 828755 Record#2: 117438
• Job title	**Generalisation:** There is no easy way to anonymise job titles in an automated way because job titles are non-standard, and organisations can invent their own. One way is to generalise job titles to a pre-defined taxonomy of job natures and/or job levels. However, the mapping likely has to be done manually.	Chief Executive Officer Team Lead, Software Development	**Generalisation:** C-level Officer IT Manager
	Swapping: Randomly switch the job titles within the dataset.	Record #1: CEO Record #2: Director Record #3: Manager	**Swapping:** Record #1: Manager Record #2: CEO Record #3: Director

[18] https://www.ura.gov.sg/maps/?service=MP.

• Company name	**Generalisation:** Generalise the company name to industry sector (e.g. with reference to the Singapore Standard Industrial Classification (SSIC))[19].	Speedy Taxi Ltd	**Generalisation:** Transportation and Storage
	Swapping: Randomly switch the company names within the dataset.	Record #1: Speedy Taxi Ltd Record #2: Best Food Ltd Record #3: No. 1 Cold Wear Pte Ltd	**Swapping:** Record #1: Best Food Ltd Record #2: No. 1 Cold Wear Pte Ltd Record #3: Speedy Taxi Ltd
• IP address	**Character masking:** Mask the last two octets[20] of IPv4 IP addresses and the last 80 bits of IPv6 IP addresses. *Note: Swapping may be applied in addition to character masking.*	IPv4: 12.120.210.88 IPv6: 2001:0db8:85a3:0000:0000:8a2e:0370:7334	**Character masking:** IPv4: 12.120.xxx.xxx IPv6: 2001:0db8:85a3:xxxx-:xxxx:xxxx:xxxx:xxxx
• Vehicle license plate number	**Character masking:** Mask the last four characters of the vehicle license plate number. *Note: Swapping may be applied in addition to character masking.*	SMF1234A	**Character masking:** SMF1xxxx
• In-vehicle unit (IU) number	**Character masking:** Mask the last three digits of the IU number. *Note: Swapping may be applied in addition to character masking.*	1234567890	**Character masking:** 1234567xxx

19. https://www.singstat.gov.sg/standards/standards-and-classifications/ssic.
20. We suggest masking the last two octets regardless of the network address class (A/B/C) to prevent the masked address from being identified as belonging to a Class B or C subnet. It also makes it harder to group individuals residing on the same subnet.

• Global Positioning System (GPS) location	**Generalisation:** *Round the GPS coordinates (in decimal degrees) to the nearest two decimal places (equivalent to accuracy of 1.11 km) or three decimal place (equivalent to accuracy of 111 m).*	1.27434, 103.79967	**Generalisation:** 1.274, 103.800 (decimal degrees rounded to three decimal places)
	Data perturbation: *Add random values between 0.005 and -0.005 or between 0.0005 and -0.0005.*		**Data perturbation:** 1.27834, 103.79767
	Swapping: *Randomly switch the GPS location values within the dataset.*	*Record #1: 1.27434, 103.79967* *Record #2: 1.26421, 103.80405* *Record #3: 1.26463, 103.82226*	**Swapping:** Record #1: 1.26463, 103.82226 Record #2: 1.27434, 103.79967 Record #3: 1.26421, 103.80405

Target attributes

Target attributes are proprietary information that is important to preserve for data utility. Hence, for most of the use cases, anonymisation techniques are not applied to target attributes. However, for the synthetic data use case, as the record-level data is typically used in development and testing environments which may not be properly secured, it is recommended that one or more anonymisation techniques are applied to the target attributes to ensure no re-identification will occur in the event of a data breach.

It is important to check and ensure that, after applying the anonymisation techniques, no record in the synthetic dataset resembles any record in the original dataset.

Target attributes	Commonly used technique(s)	Example(s)	
		Before	After
• Transactions • Salary • Credit rating • Insurance policy • Medical diagnosis • Vaccination status	**Data perturbation:** Randomly modify the numerical data (e.g. adding or subtracting random values from the original data). Data perturbation is not possible for alphanumerical or unstructured textual data.	Purchase value: $38.05 Salary: $6,200	**Data perturbation:** Purchase value: $42 Salary: $7,500
	Swapping: Randomly switch data within the dataset. *Note: Swapping may be applied in addition to data perturbation.*	Vaccination status: Record#1: Vaccinated Record#2: First dose Record#3: Unvaccinated	**Swapping:** Vaccination status: Record#1: First dose Record#2: Unvaccinated Record#3: Vaccinated

ANNEX C: *k*-ANONYMITY

k-anonymity (and similar extensions to it like *l*-diversity and *t*-closeness) is a measure used to ensure that the risk threshold has not been surpassed, as part of the anonymisation methodology.

k-anonymity is not the only measure available, nor is it without its limitations, but it is relatively well understood and easy to apply. *k*-anonymity may not be suitable for all types of datasets or other complex use cases. Other approaches and/or tools such as Special Uniques Detection Algorithm (SUDA) and μ-Argus may be more suitable for assessing the risk of large datasets. Alternative methods, such as differential privacy[21], have also emerged over the past few years.

***k*-anonymity**	
Description	The *k*-anonymity model is used as a guideline before anonymisation techniques (e.g. generalisation) have been applied, and for verification after as well, to ensure that any record's indirect identifiers are shared by at least *k*-1 other records. This is the key protection provided by *k*-anonymity against linking attacks, because *k* records (or at least different indirect identifiers) are identical in their identifying attributes and thus, create an equivalence class[22] with *k* members. Therefore, it is not possible to link or single out an individual's record since there are always *k* identical attributes. An anonymised dataset may have different *k*-anonymity levels for different sets of indirect identifiers, but for "maximum risk" protection against linking, the lowest *k* is used as a representative value for comparison against the threshold.
When to use it	*k*-anonymity is used to confirm that the anonymisation measures put in place achieve the desired threshold against linking attacks.
How to use it	First, decide on a value for *k* (that is equal to or higher than the inverse of the equivalence class size), which provides the lowest *k* to be achieved among all equivalence classes. Generally, the higher the value of *k*, the harder it is for data subjects to be identified; however, utility may become lower as *k* increases and more records may need to be suppressed. After anonymisation techniques have been applied, check that each record has at least *k*-1 other records with the same attributes addressed by the *k*-anonymisation. Records in equivalence classes with less than *k* records should be considered for suppression; alternatively, the dataset can be anonymised further.
Other tips	• Besides generalisation and suppression, synthetic data can also be created to achieve *k*-anonymity. These techniques (and others) can sometimes be used in combination, but do note that the specific method chosen can affect data utility. Consider the trade-offs between dropping the outliers or inserting synthetic data. • *k*-anonymity assumes that each record relates to a different individual. If the same individual has multiple records (e.g. visiting the hospital on several occasions), then *k*-anonymity will need to be higher than the repeat records, otherwise the records may not only be linkable, but *may* also be re-identifiable, despite seemingly fulfilling "*k* equivalence classes".

21. Differential privacy involves several concepts, including answering queries instead of providing the anonymised dataset, adding randomised noise to the protect individual records, providing mathematical guarantees that the pre-defined "privacy budget" is not exceeded, etc.

22. "Equivalence class" refers to records in a dataset that share the same values within certain attributes, typically indirect identifiers.

EXAMPLE

In this example, the dataset contains information about people taking taxis.

$k = 5$ is used (i.e. each record should eventually share the same attributes with four other records after anonymisation).

The following anonymisation techniques are used in combination. The level of granularity is one approach to achieving the required k level.

Attribute	Anonymisation technique
Age	Generalisation (10-year intervals)
Occupation	Generalisation (e.g. both "database administrator" and "programmer" are generalised to "IT")
Record suppression	Records that do not meet the 5-anonymity criteria after anonymisation techniques have been applied (in this case, generalisation) are removed. For example, the banker's record is removed as it is the only such value under "Occupation".

Dataset before anonymisation:

Serial number	Age	Gender	Occupation	Average number of trips per week
1	21	Female	Assistant Data Protection Officer	15
2	38	Male	Lead IT Consultant	2
3	25	Female	Banker	8
4	34	Male	Database Administrator	3
5	30	Female	Chief Privacy Officer	1
6	29	Female	Regional Data Protection Officer	5
7	38	Male	Programmer	3
8	32	Male	IT Analyst	4
9	25	Female	Deputy Data Protection Officer	2
10	23	Female	Manager, DPO Office	11
11	31	Male	UX Designer	0

Dataset becoming 5-anonymous after the anonymisation of age and occupation, and suppression of the outlier. (The respective equivalence classes are highlighted in different colours):

Serial number	Age	Gender	Occupation	Average number of trips per week
1	21 to 30	Female	Data Protection Officer	15
2	31 to 40	Male	IT	2
3	~~21 to 30~~	~~Female~~	~~Banker~~	~~8~~
4	31 to 40	Male	IT	3
5	21 to 30	Female	Data Protection Officer	1
6	21 to 30	Female	Data Protection Officer	5
7	31 to 40	Male	IT	3
8	31 to 40	Male	IT	4
9	21 to 30	Female	Data Protection Officer	2
10	21 to 30	Female	Data Protection Officer	11
11	31 to 40	Male	IT	0

Note: The average number of trips per week is taken here as an example for a target attribute, without a need to further anonymise this attribute.

ANNEX D: ASSESSING THE RISK OF RE-IDENTIFICATION

There are various ways to assess the risk of re-identification, and these may require rather complex calculations involving computation of probabilities.

This section describes a simplified model, using *k*-anonymity[27], and makes the following assumptions:

1. The release model is non-public;
2. The attacker is motivated to link an individual to the anonymised dataset; and
3. The content of the anonymised data is not taken into consideration and that the risk calculated is independent of the kind of information the attacker actually has available.

First, the risk threshold should be established. This value, reflecting a probability, ranges between 0 and 1. It reflects the risk level that the organisation is willing to accept. The main factors affecting the risk threshold should include the harm that could be caused to the data subject, as well as the harm to the organisation, if re-identification takes place; but, it also takes into consideration what other controls have been put in place to mitigate any residual risks. The higher the potential harm, the higher the risk threshold should be. There are no hard and fast rules as to what risk threshold values should be used; the following are just examples.

Potential harm	Risk threshold
Low	0.2
Medium	0.1
High	0.01

23. The calculations would be different if done using differential privacy or traditional statistical disclosure controls, for example.

In computing the re-identification risk, this guide uses the "Prosecutor Risk", which assumes that the attacker knows a specific person in the dataset and is trying to establish which record in the dataset refers to that person.

The simple rule for calculating the probability of re-identification for a single record in a dataset, is to take the inverse of the record's equivalence class size:

P (link individual to a single record) = 1 / record's equivalence class size

To compute the probability of re-identification of any record in the entire dataset, given that there is a re-identification attempt, a conservative approach would be to equate it to the maximum probability of re-identification among all records in the dataset.

P (re-ID any record in dataset) = 1 / Min. equivalence class size in dataset

Note: If the dataset has been k-anonymised, P (re-ID any record in dataset) <= 1 / k

We can consider three motivated intruder attack scenarios:

1. the deliberate insider attack;
2. the inadvertent recognition by an acquaintance; and
3. a data breach.

P (re-ID) = P (re-ID | re-ID attempt) x P (re-ID attempt)

where P (re-ID | re-ID attempt) refers to the probability of successful re-identification, given there is a re-identification attempt. As discussed earlier, we can take P (re-ID | re-ID attempt) to be (1 / Min. equivalence class size in dataset)

Therefore, P (re-ID) = (1 / Min. equivalence class size in dataset) x P (re-ID attempt)

For Scenario #1, the deliberate insider attack, we assume a party receiving the dataset attempts re-identification. To estimate P (re-ID attempt): the probability of a re-identification attempt, factors to consider include the extent of mitigating controls put in place as well as the motives and resources of the attacker. The following table presents example values; again, it is for the party anonymising the dataset to decide on suitable values to use.

Scenario #1—the deliberate insider attack P (re-ID attempt) = P(insider attack)		Motivation and resources of attacker		
		Low	Medium	High
Extent of mitigating controls	High	0.03	0.05	0.1
	Medium	0.2	0.25	0.3
	Low	0.4	0.5	0.6
	None	1.0	1.0	1.0

Factors affecting the motivation and resources of the attacker may include:

1. Willingness to violate contract (assuming contract preventing re-identification is in place)
2. Financial and time constraints
3. Inclusion of high-profile personalities (e.g. celebrities) or sensitive data (e.g. credit information) in the dataset
4. Ease of access to "linkable" data or information, whether publicly available or privately owned, that may enable re-identification of the anonymised dataset

Factors affecting the extent of mitigating controls include:

1. Organisational structures
2. Administrative/legal controls (e.g. contracts)
3. Technical and process controls

For Scenario #2, the inadvertent recognition by an acquaintance, we assume a party receiving the dataset inadvertently re-identifies a data subject while examining the dataset. This is possible because the party has some additional knowledge about the data subject due to their relationship (e.g. friend, neighbour, relative, colleague, etc.). To estimate P (re-ID attempt): the probability of a re-identification attempt, the main factor to consider is the likelihood that the data recipient knows someone in the dataset.

Scenario #2—inadvertent recognition by an acquaintance
P (re-ID attempt) = P (data recipient knowing a person in the dataset)

For Scenario #3, the probability of a data breach occurring at the data recipient's ICT system can be estimated based on available statistics about the prevalence of data breaches in the data recipient's industry. This is based on the assumption that the attackers who obtained the dataset will attempt re-identification.

Scenario #3—a data breach
P (re-ID attempt) = P (data breach in data recipient's industry)

The highest probability among the three scenarios should be used as P (re-ID attempt).

P (re-ID attempt) = Max (P(insider attack), P(data recipient knowing a person inside the dataset), P(data breach in data recipient's industry))

To put everything together,

P (re-ID) = (1 / Min. equivalence class size in dataset) x P (re-ID attempt)
= (1 / k) x P (re-ID attempt) ***for k-anonymised dataset***

where P (re-ID attempt) = Max (P (insider attack),
P (data recipient knowing a person in the dataset), P(data breach in data recipient's industry))

ANNEX E: ANONYMISATION TOOLS

The following is a list of some commercial or open-source anonymisation tools.

Tool	Description	URL
Amnesia	Amnesia anonymisation tool is a software used locally to anonymise personal and sensitive data. It currently supports k-anonymity and km-anonymity guarantees.	https://amnesia.openaire.eu/
Arcad DOT-Anonymizer	DOT-Anonymizer is a tool that maintains the confidentiality of test data by concealing personal information. It works by anonymising personal data while preserving its format and type.	https://www.arcadsoftware.com/dot/data-masking/dot-anonymizer/
ARGUS	ARGUS stands for "Anti Re-identification General Utility System". The tool uses a wide range of different statistical anonymisation methods such as global recoding (grouping of categories), local suppression, randomisation, adding noise, microaggregation, top- and bottom coding. It can also be used to generate synthetic data.	https://research.cbs.nl/casc/mu.htm
ARX	ARX is an open-source software for anonymising sensitive personal data.	https://arx.deidentifier.org/
Eclipse	Eclipse is a suite of tools from Privacy Analytics that facilitates anonymisation of health data.	https://privacy-analytics.com/health-data-privacy/health-data-software/
sdcMicro	sdcMicro is used to generate anonymised microdata such as public and scientific use files. It supports different risk estimation methods.	https://cran.r-project.org/web/packages/sdcMicro/index.html
UTD Anonymisation Toolbox	UT Dallas Data Security and Privacy Lab compiled various anonymisation techniques into a toolbox for public use.	http://cs.utdallas.edu/dspl/cgi-bin/toolbox/index.php?go=home

ACKNOWLEDGEMENTS

ACKNOWLEDGEMENTS

The PDPC and Infocomm Media Development Authority (IMDA) express their sincere appreciation to the following organisations for their valuable feedback in the development of this publication.

- AsiaDPO
- BetterData Pte Ltd
- ISACA (Singapore Chapter)—Data Protection SIG
- Law Society of Singapore—Cybersecurity and Data Protection Committee (CSDPC)
- Ministry of Health (MOH)
- Replica Analytics
- Privitar Ltd
- SGTech
- Singapore Business Federation (SBF)—Digitalisation Committee
- Singapore Corporate Counsel Association (SCCA)—Data Protection, Privacy and Cybersecurity (DPPC) Chapter
- Singapore Department of Statistics (DOS)
- Smart Nation and Digital Government Group (SNDGO)

The following guides were referenced in this guide.

- UKAN. *The Anonymisation Decision Making Framework 2nd Edition: European Practitioners' Guide, by* Mark Elliot, Elaine Mackey and Kieron O'Hara, 2020.
- CSIRO and OAIC. *The De-Identification Decision-Making Framework*, by Christine M O'Keefe, Stephanie Otorepec, Mark Elliot, Elaine Mackey and Kieron O'Hara, 18 September 2017.
- IPC. *De-identification Guidelines for Structured Data*, June 2016, https://www.ipc.on.ca/wp-content/uploads/2016/08/Deidentification-Guidelines-for-Structured-Data.pdf.
- El Emam, K. *Guide to the De-Identification of Personal Health Information*, CRC Press, 2013.
- Article 29 Data Protection Working Party (European Commission). "Opinion 05/2014 on Anonymisation Techniques". 10 April 2014, http://ec.europa.eu/justice/data-protection/article-29/documentation/opinion-recommendation/files/2014/wp216_en.pdf.
- NIST. *NISTIR 8053: De-Identification of Personal Information*, by S L Garfinkel, October 2015, http://nvlpubs.nist.gov/nistpubs/ir/2015/NIST.IR.8053.pdf.

#SGDIGITAL

Singapore Digital (SG:D) gives Singapore's digitalisation efforts a face, identifying our digital programmes and initiatives with one set of visuals, and speaking to our local and international audiences in the same language.

The SG:D logo is made up of rounded fonts that evolve from the expressive dot that is red. SG stands for Singapore and :D refers to our digital economy. The :D smiley face icon also signifies the optimism of Singaporeans moving into a digital economy. As we progress into the digital economy, it's all about the people — empathy and assurance will be at the heart of all that we do.

BROUGHT TO YOU BY

pdpc PERSONAL DATA PROTECTION COMMISSION SINGAPORE