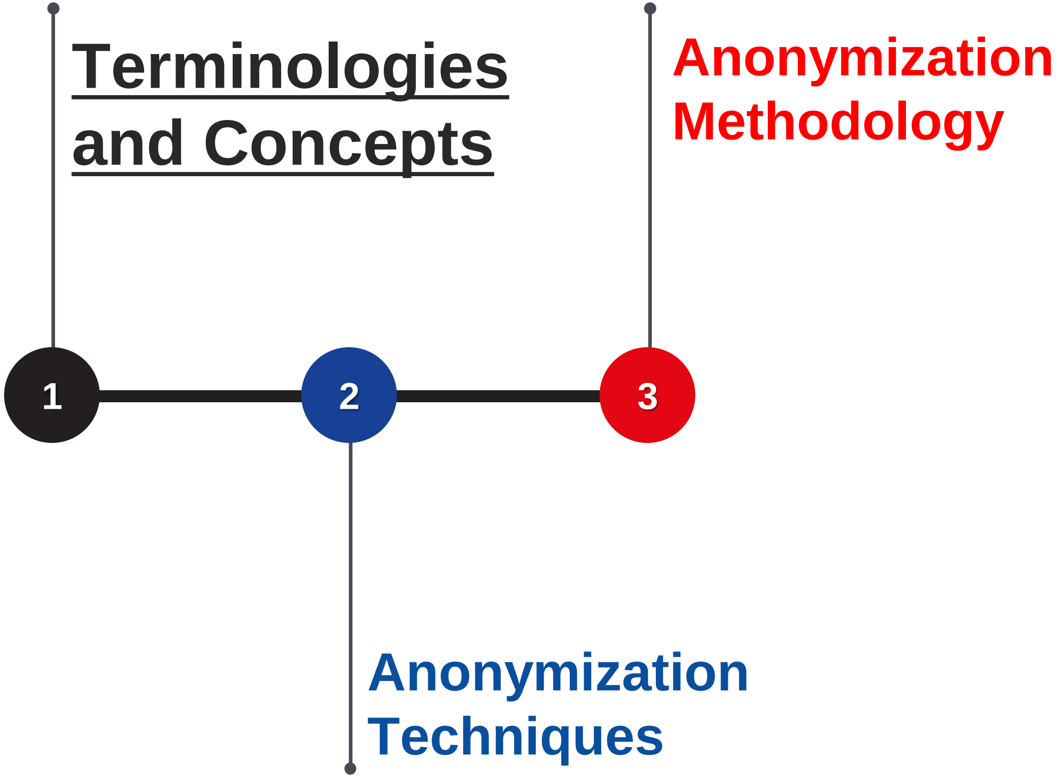
NANYANG
THE INNOVATIVE POLYTECHNIC

Data Privacy and Protection

Topic 3

Data Anonymization

Contents

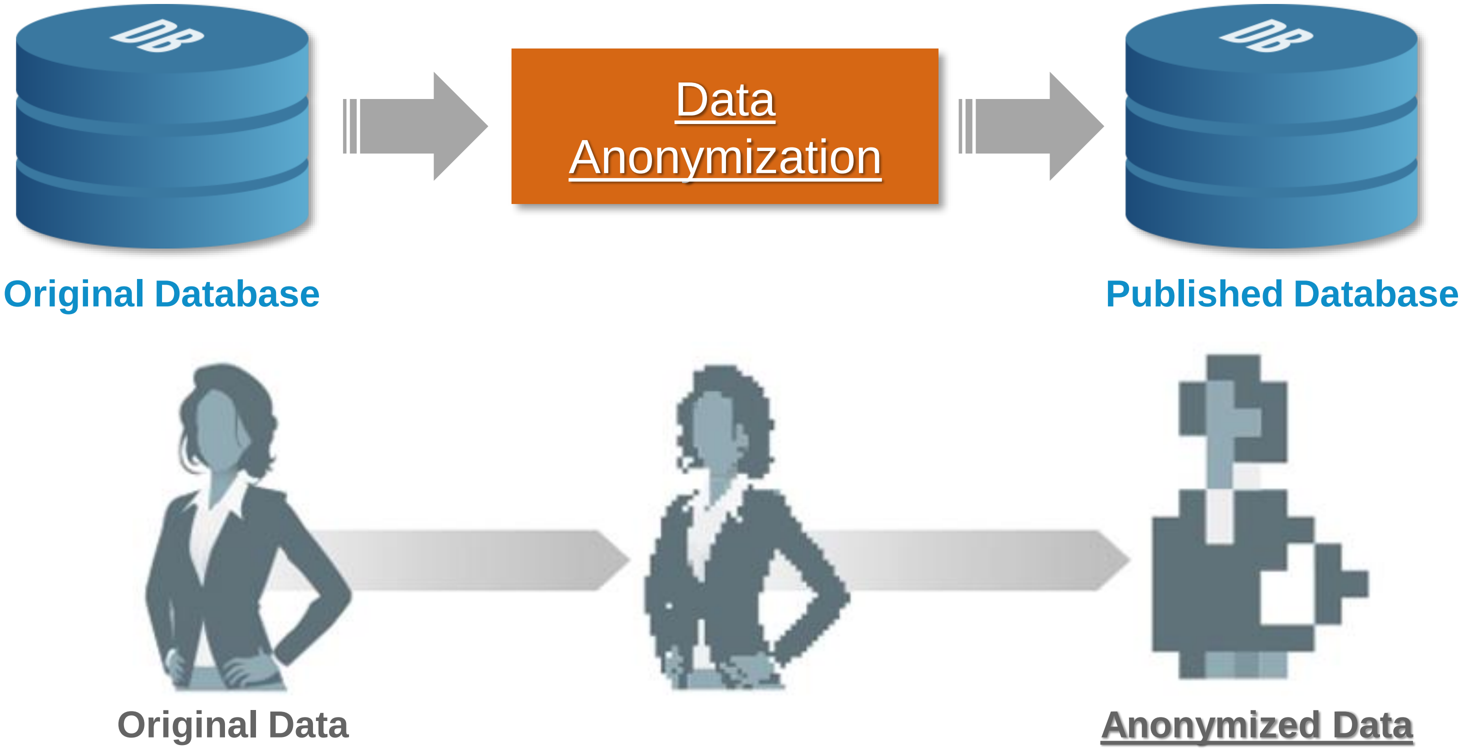


1 Terminologies and Concepts

2 Anonymization Techniques

3 Anonymization Methodology

Terminologies



Original Database

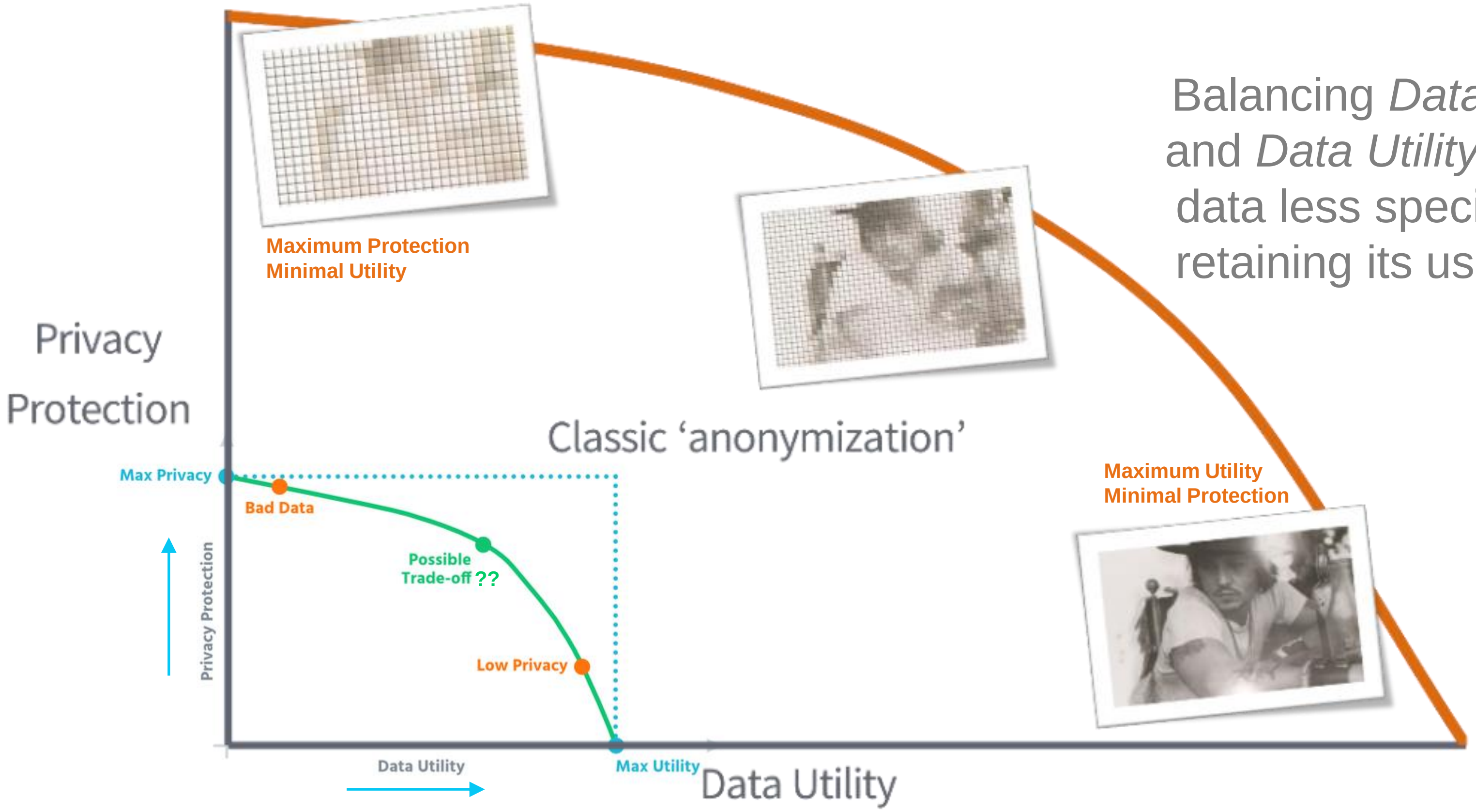
Data Anonymization

Published Database

Original Data

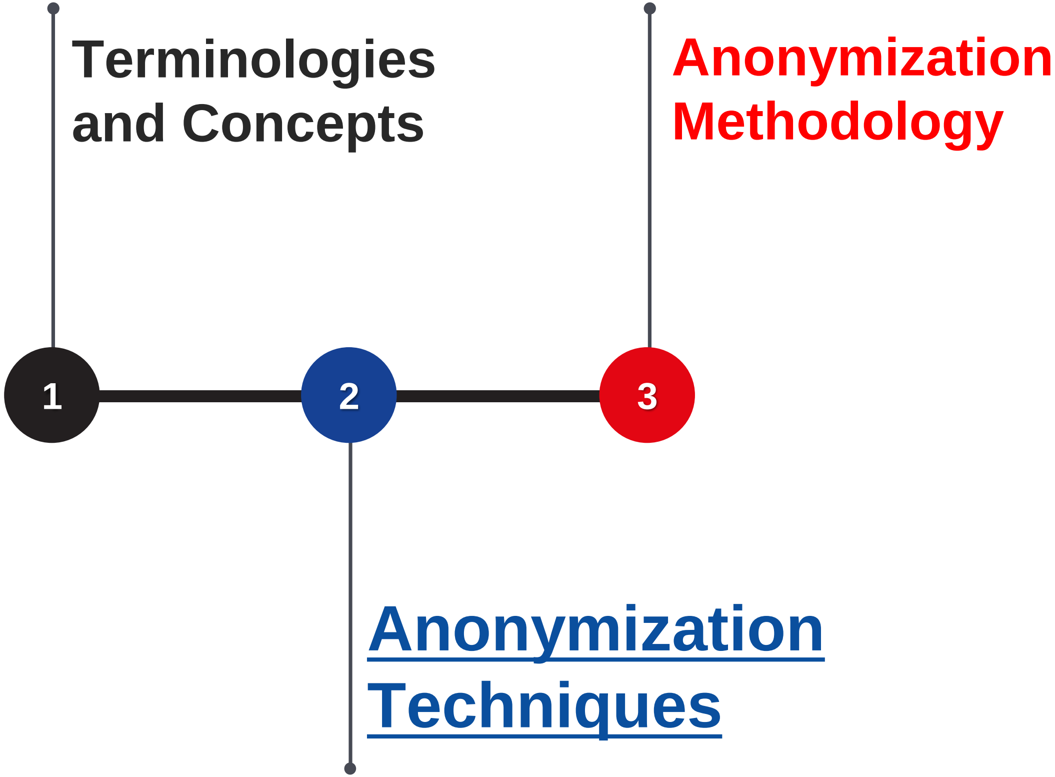
Anonymized Data

Anonymization – Data Privacy vs Data Utility



Balancing *Data Privacy* and *Data Utility* to make data less specific while retaining its usefulness

Contents



1 Terminologies and Concepts

2 Anonymization Techniques

3 Anonymization Methodology

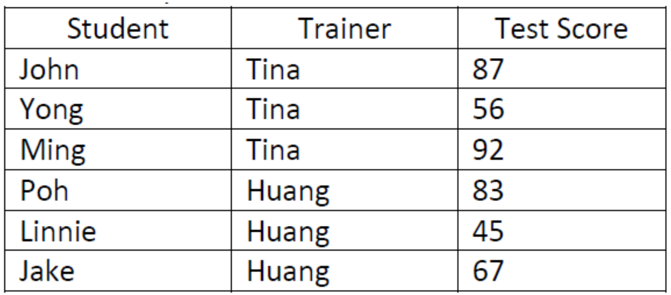
Attribute Suppression

- Removal of an entire part of data (column in databases or spreadsheets) in a dataset

- Used when an attribute is not required in the anonymised dataset

- Strongest type of anonymization technique

Attribute Suppression – Example



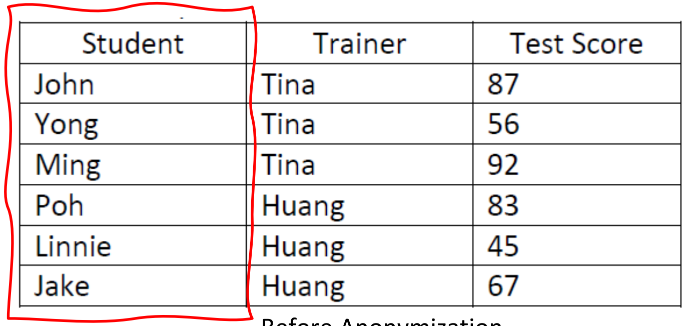
Student	Trainer	Test Score
John	Tina	87
Yong	Tina	56
Ming	Tina	92
Poh	Huang	83
Linnie	Huang	45
Jake	Huang	67

Before Anonymization

Example: Data consists of test scores

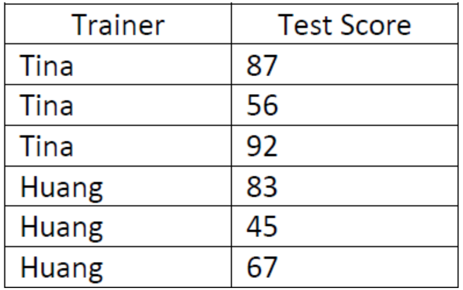
- Recipient only needs to analyse test scores with respect to trainers

Attribute Suppression – Example



Student	Trainer	Test Score
John	Tina	87
Yong	Tina	56
Ming	Tina	92
Poh	Huang	83
Linnie	Huang	45
Jake	Huang	67

Before Anonymization



Trainer	Test Score
Tina	87
Tina	56
Tina	92
Huang	83
Huang	45
Huang	67

After Suppression

Example: Data consists of test scores

- Recipient only needs to analyse test scores with respect to trainers
- The “*student*” attribute is removed

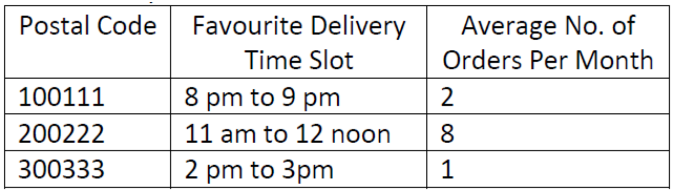
Character Masking

- Characters of a data value is masked by using a symbol, e.g. “*” or “x”

- Used when hiding part of a string of characters, is sufficient to provide the anonymity required

- Depending on attribute type, mask to replace a fixed number of characters, or a variable number of characters

Character Masking – Example

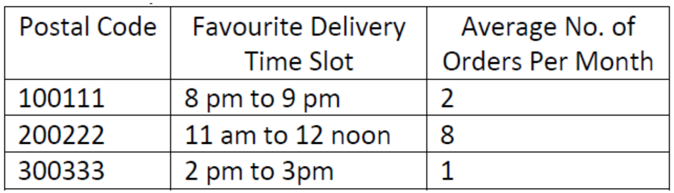


Postal Code	Favourite Delivery Time Slot	Average No. of Orders Per Month
100111	8 pm to 9 pm	2
200222	11 am to 12 noon	8
300333	2 pm to 3pm	1

Before Anonymization

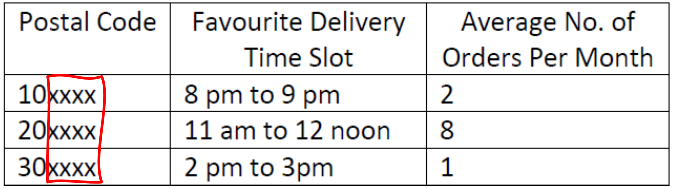
Example: online grocery store conducting a study of its delivery demand from historical data

Character Masking – Example



Postal Code	Favourite Delivery Time Slot	Average No. of Orders Per Month
100111	8 pm to 9 pm	2
200222	11 am to 12 noon	8
300333	2 pm to 3pm	1

Before Anonymization



Postal Code	Favourite Delivery Time Slot	Average No. of Orders Per Month
10xxxx	8 pm to 9 pm	2
20xxxx	11 am to 12 noon	8
30xxxx	2 pm to 3pm	1

After Partial Masking

Example: online grocery store conducting a study of its delivery demand from historical data

- last 4 digits of the postal codes is masked
- leaving the first 2 digits, which correspond to the “sector code”

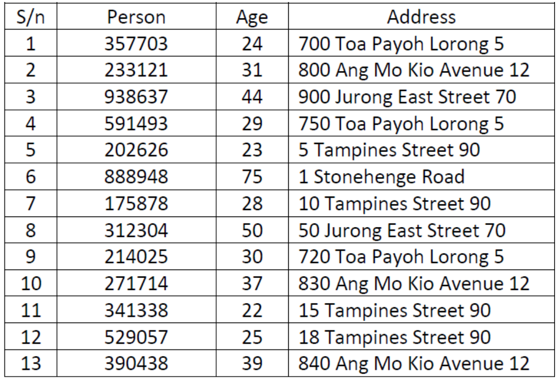
Generalisation

- Reduction in the precision of data, e.g., converting a person’s age into a range of values

- Used where values can be generalised into a range, and still be useful

- Data ranges that are too large may mean too much modification, data ranges too small may be too easy to re-identify individuals

Generalisation – Example

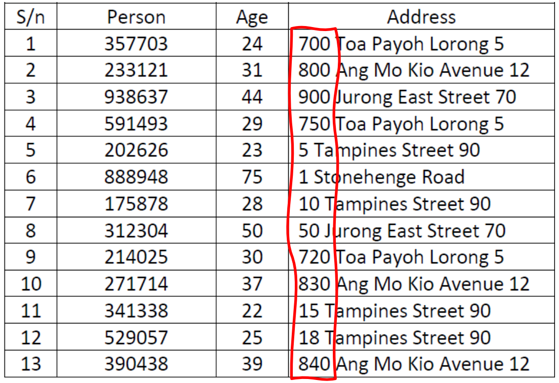


S/n	Person	Age	Address
1	357703	24	700 Toa Payoh Lorong 5
2	233121	31	800 Ang Mo Kio Avenue 12
3	938637	44	900 Jurong East Street 70
4	591493	29	750 Toa Payoh Lorong 5
5	202626	23	5 Tampines Street 90
6	888948	75	1 Stonehenge Road
7	175878	28	10 Tampines Street 90
8	312304	50	50 Jurong East Street 70
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

Pseudonymized Dataset

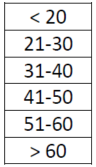
Example: Dataset contains person name, age in years, and residential address

Generalisation – Example



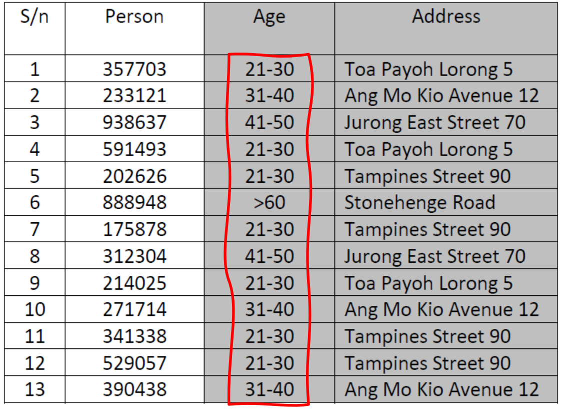
S/n	Person	Age	Address
1	357703	24	700 Toa Payoh Lorong 5
2	233121	31	800 Ang Mo Kio Avenue 12
3	938637	44	900 Jurong East Street 70
4	591493	29	750 Toa Payoh Lorong 5
5	202626	23	5 Tampines Street 90
6	888948	75	1 Stonehenge Road
7	175878	28	10 Tampines Street 90
8	312304	50	50 Jurong East Street 70
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

Pseudonymized Dataset



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

Age Range



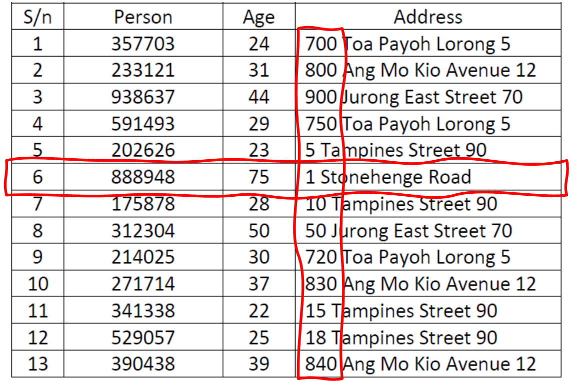
S/n	Person	Age	Address
1	357703	21-30	Toa Payoh Lorong 5
2	233121	31-40	Ang Mo Kio Avenue 12
3	938637	41-50	Jurong East Street 70
4	591493	21-30	Toa Payoh Lorong 5
5	202626	21-30	Tampines Street 90
6	888948	>60	Stonehenge Road
7	175878	21-30	Tampines Street 90
8	312304	41-50	Jurong East Street 70
9	214025	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

After Generalisation

Example: Dataset contains person name, age in years, and residential address

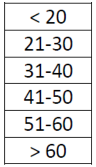
- Age ranges of 10 years, starting with a range <20 years, and ending with range >60 years
- Remove the block/house number and retain only the road name in Address

Generalisation – Example



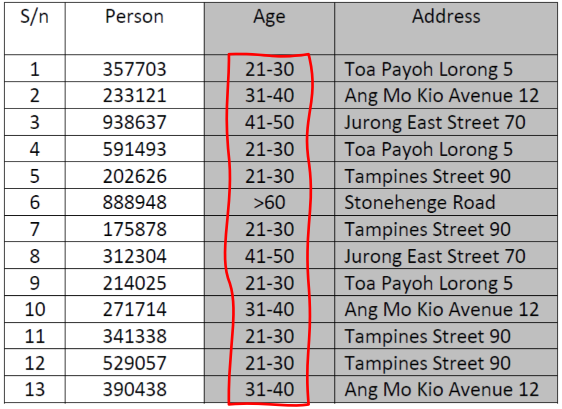
S/n	Person	Age	Address
1	357703	24	700 Toa Payoh Lorong 5
2	233121	31	800 Ang Mo Kio Avenue 12
3	938637	44	900 Jurong East Street 70
4	591493	29	750 Toa Payoh Lorong 5
5	202626	23	5 Tampines Street 90
6	888948	75	1 Stonehenge Road
7	175878	28	10 Tampines Street 90
8	312304	50	50 Jurong East Street 70
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

Pseudonymized Dataset



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

Age Range



S/n	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

After Generalisation

Example: Dataset contains person name, age in years, and residential address

- Age ranges of 10 years, starting with a range <20 years, and ending with range >60 years
- Remove the block/house number and retain only the road name in Address
- Only 1 residential unit on Stonehenge Road – too unique
  - Remove record number 6
  - Generalised address to a greater extent

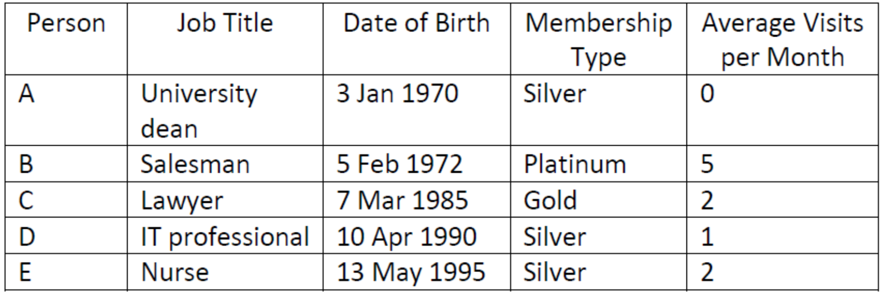
Swapping

- Rearrangement of data in the dataset such that the individual attribute values are represented, but do not correspond to the original records

- Used when subsequent analysis only needs to look at aggregated data, not relationships between attributes

- Not all attributes (columns) need to be swapped, depending on the situation, only attributes containing values that are relatively identifiable need to be swapped

Swapping – Example

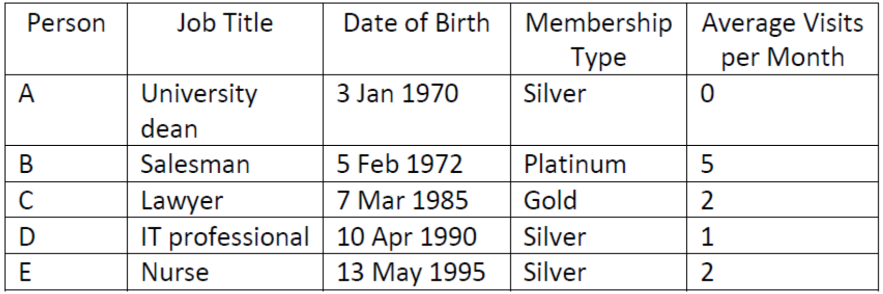


Person	Job Title	Date of Birth	Membership Type	Average Visits per Month
A	University dean	3 Jan 1970	Silver	0
B	Salesman	5 Feb 1972	Platinum	5
C	Lawyer	7 Mar 1985	Gold	2
D	IT professional	10 Apr 1990	Silver	1
E	Nurse	13 May 1995	Silver	2

Before Anonymization

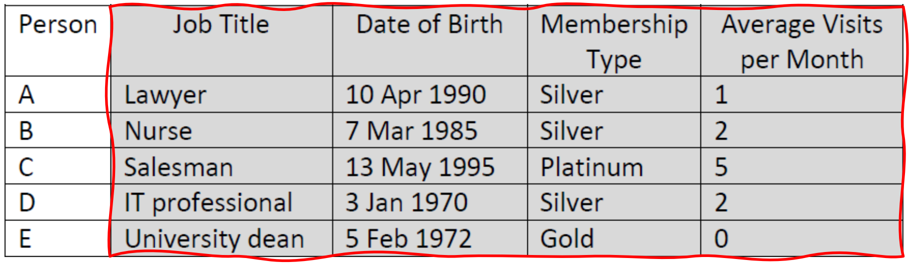
Example: Dataset contains information about customer records for a business organisation

Swapping – Example



Person	Job Title	Date of Birth	Membership Type	Average Visits per Month
A	University dean	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

Before Anonymization



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 dean	5 Feb 1972	Gold	0

After Anonymization

Example: Dataset contains information about customer records for a business organisation

- All values for all attributes have been swapped

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

Data Perturbation

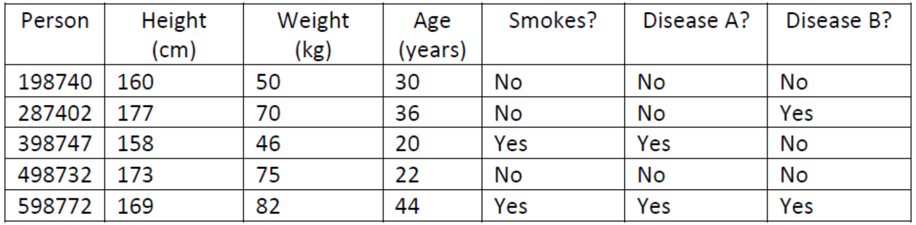
- The values from the original dataset are modified to be slightly different

- This is used for quasi-identifiers and typically for numbers and dates, and should not be used where data accuracy is crucial

- The degree of perturbation should be proportionate, to the range of values, of the attribute

Data Perturbation – Example

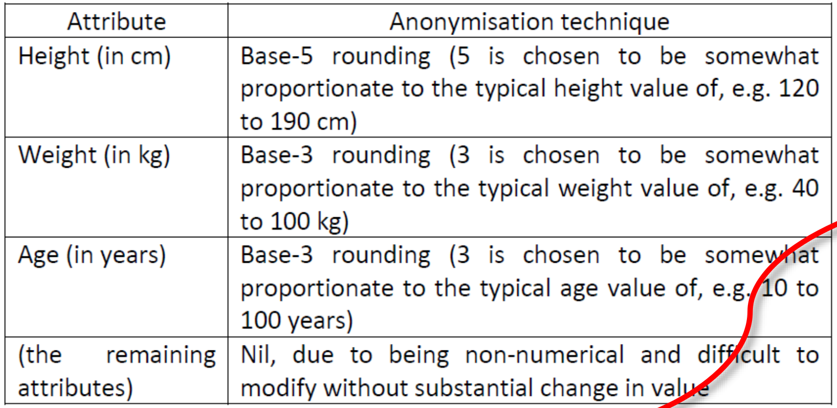
Before Anonymization



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

Example: Information to be used for research on possible linkage between a person’s height, weight, age, whether the person smokes, and whether the person has “disease A” and/or “disease B”. Name has been pseudonymised.

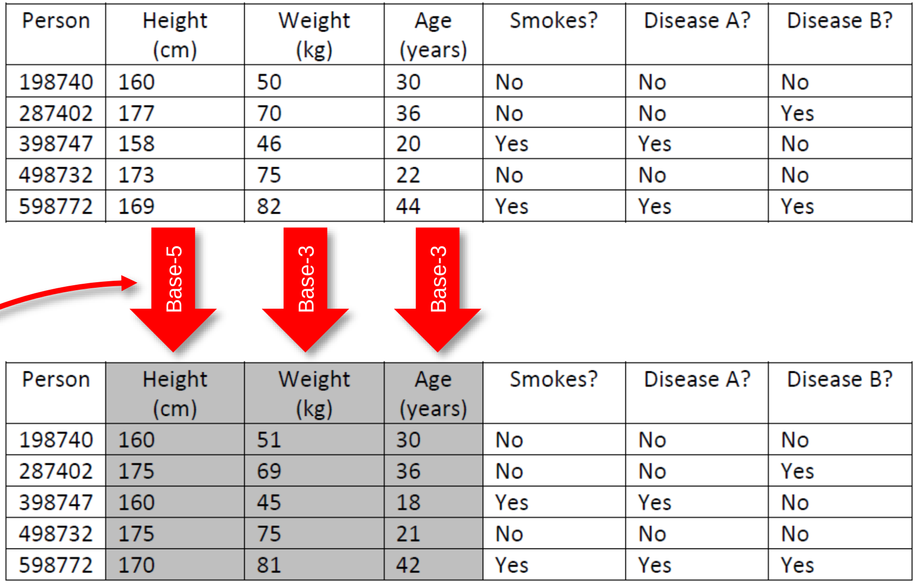
Data Perturbation – Example



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

Rounding to be applied

Before Anonymization


Base-5
Base-3
Base-3

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

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

After Anonymization
(shaded columns represent affected attributes)

Example: Information to be used for research on possible linkage between a person’s height, weight, age, whether the person smokes, and whether the person has “disease A” and/or “disease B”. Name has been pseudonymised.

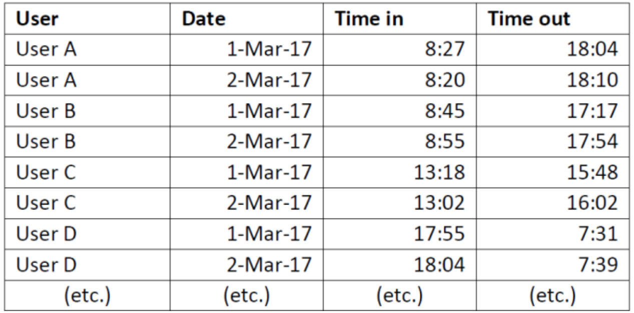
Synthetic Data

- Data that is artificially or programmatically created often with the help of algorithms, rather than being generated by actual events

- Captures the underlying structure and display the same statistical distributions as the original data

- Used for a wide range of activities, including as test data for new products, and in AI model training, yet maintaining data privacy

Synthetic Data – Example



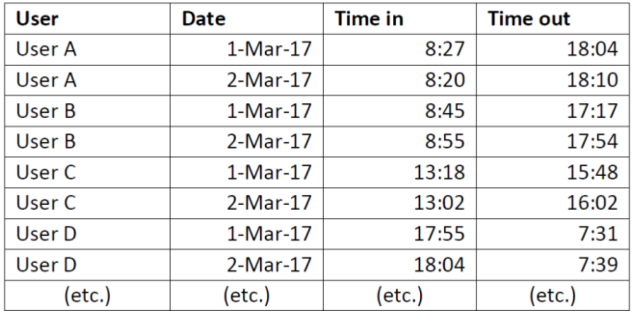
User	Date	Time in	Time out
User A	1-Mar-17	8:27	18:04
User A	2-Mar-17	8:20	18:10
User B	1-Mar-17	8:45	17:17
User B	2-Mar-17	8:55	17:54
User C	1-Mar-17	13:18	15:48
User C	2-Mar-17	13:02	16:02
User D	1-Mar-17	17:55	7:31
User D	2-Mar-17	18:04	7:39
(etc.)	(etc.)	(etc.)	(etc.)

Original Data (actual events)

Example: Office facility, providing “hot-desking” facilities, keep records of the time that users start and end using their facilities.

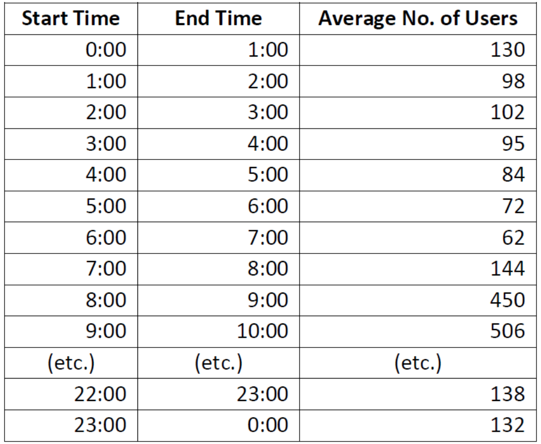
- They would like synthetic data for 1 day, to perform simulation testing on a new facility allocation

Synthetic Data – Example



User	Date	Time in	Time out
User A	1-Mar-17	8:27	18:04
User A	2-Mar-17	8:20	18:10
User B	1-Mar-17	8:45	17:17
User B	2-Mar-17	8:55	17:54
User C	1-Mar-17	13:18	15:48
User C	2-Mar-17	13:02	16:02
User D	1-Mar-17	17:55	7:31
User D	2-Mar-17	18:04	7:39
(etc.)	(etc.)	(etc.)	(etc.)

Original Data (actual events)



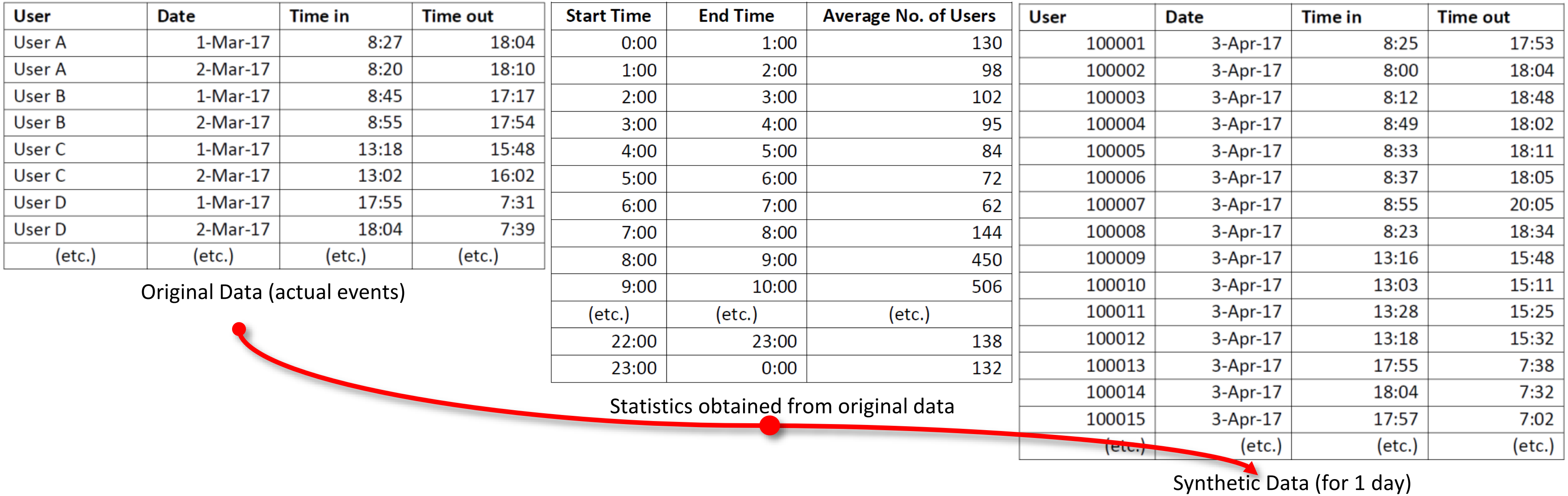
Start Time	End Time	Average No. of Users
0:00	1:00	130
1:00	2:00	98
2:00	3:00	102
3:00	4:00	95
4:00	5:00	84
5:00	6:00	72
6:00	7:00	62
7:00	8:00	144
8:00	9:00	450
9:00	10:00	506
(etc.)	(etc.)	(etc.)
22:00	23:00	138
23:00	0:00	132

Statistics obtained from original data

Example: Office facility, providing “hot-desking” facilities, keep records of the time that users start and end using their facilities.

- They would like synthetic data for 1 day, to perform simulation testing on a new facility allocation

Synthetic Data – Example



User	Date	Time in	Time out
User A	1-Mar-17	8:27	18:04
User A	2-Mar-17	8:20	18:10
User B	1-Mar-17	8:45	17:17
User B	2-Mar-17	8:55	17:54
User C	1-Mar-17	13:18	15:48
User C	2-Mar-17	13:02	16:02
User D	1-Mar-17	17:55	7:31
User D	2-Mar-17	18:04	7:39
(etc.)	(etc.)	(etc.)	(etc.)

Original Data (actual events)

Start Time	End Time	Average No. of Users
0:00	1:00	130
1:00	2:00	98
2:00	3:00	102
3:00	4:00	95
4:00	5:00	84
5:00	6:00	72
6:00	7:00	62
7:00	8:00	144
8:00	9:00	450
9:00	10:00	506
(etc.)	(etc.)	(etc.)
22:00	23:00	138
23:00	0:00	132

Statistics obtained from original data

User	Date	Time in	Time out
100001	3-Apr-17	8:25	17:53
100002	3-Apr-17	8:00	18:04
100003	3-Apr-17	8:12	18:48
100004	3-Apr-17	8:49	18:02
100005	3-Apr-17	8:33	18:11
100006	3-Apr-17	8:37	18:05
100007	3-Apr-17	8:55	20:05
100008	3-Apr-17	8:23	18:34
100009	3-Apr-17	13:16	15:48
100010	3-Apr-17	13:03	15:11
100011	3-Apr-17	13:28	15:25
100012	3-Apr-17	13:18	15:32
100013	3-Apr-17	17:55	7:38
100014	3-Apr-17	18:04	7:32
100015	3-Apr-17	17:57	7:02
(etc.)	(etc.)	(etc.)	(etc.)

Synthetic Data (for 1 day)

Example: Office facility, providing “hot-desking” facilities, keep records of the time that users start and end using their facilities.

- They would like synthetic data for 1 day, to perform simulation testing on a new facility allocation
- Synthetic data created, based on the statistics derived from the original data

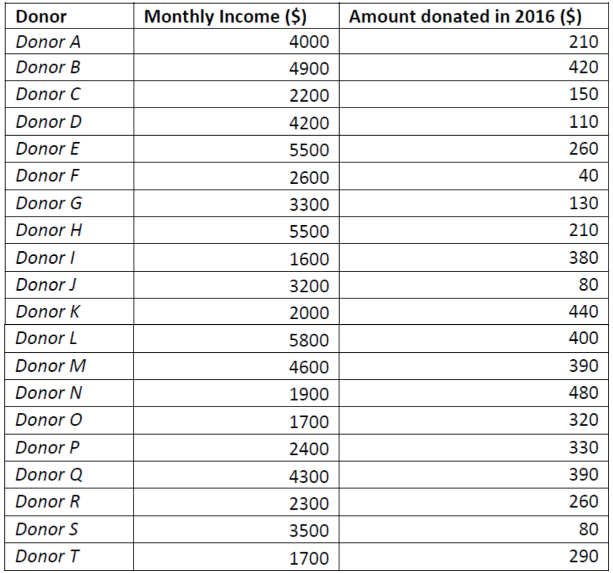
Data Aggregation

- Converting a dataset from a list of records to summarised values

- Used when individual records are not required and aggregated data is sufficient for the purpose

- 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 an individual, hence, aggregation may need to be applied in combination with suppression

Data Aggregation – Example

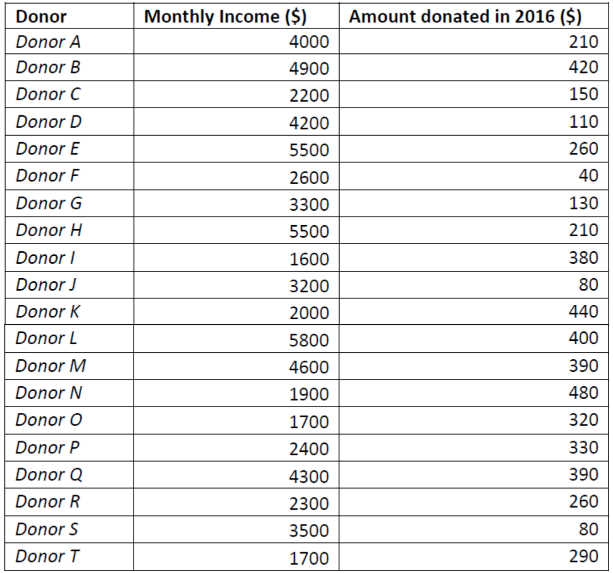


Donor	Monthly Income ($)	Amount donated in 2016 ($)
Donor A	4000	210
Donor B	4900	420
Donor C	2200	150
Donor D	4200	110
Donor E	5500	260
Donor F	2600	40
Donor G	3300	130
Donor H	5500	210
Donor I	1600	380
Donor J	3200	80
Donor K	2000	440
Donor L	5800	400
Donor M	4600	390
Donor N	1900	480
Donor O	1700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

Original Data

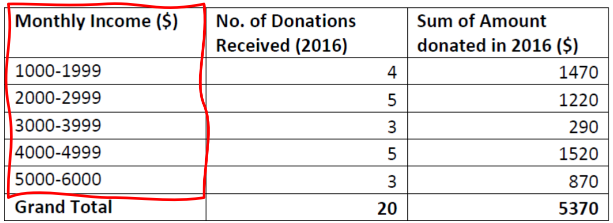
Example: charity organisation has records of the donations made, as well as some information about the donors.

Data Aggregation – Example



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

Original Data



Monthly Income ($)	No. of Donations Received (2016)	Sum of Amount donated in 2016 ($)
1000-1999	4	1470
2000-2999	5	1220
3000-3999	3	290
4000-4999	5	1520
5000-6000	3	870
Grand Total	**20**	**5370**

Anonymized Data

Example: charity organisation has records of the donations made, as well as some information about the donors. Aggregated data is assessed to be sufficient to perform data analysis.

K-anonymity

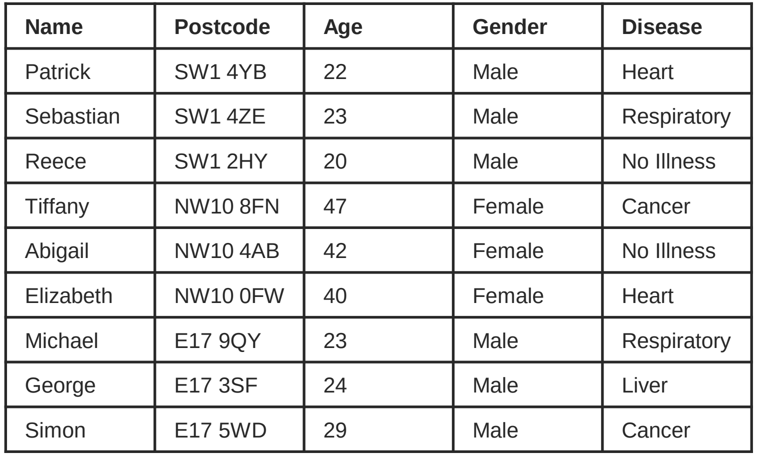
- A property of a dataset that is usually used in order to describe the dataset’s level of anonymity

- Protects against re-identification, and often described as a ‘hiding in the crowd’ guarantee

- k in k-anonymity refers to the number of times each combination of values appears in a dataset

- If k = 3, the data is said to be 3-anonymous, the higher the value of ‘k’, the harder it is for individuals to be identified

K-anonymity – Example



Name	Postcode	Age	Gender	Disease
Patrick	SW1 4YB	22	Male	Heart
Sebastian	SW1 4ZE	23	Male	Respiratory
Reece	SW1 2HY	20	Male	No Illness
Tiffany	NW10 8FN	47	Female	Cancer
Abigail	NW10 4AB	42	Female	No Illness
Elizabeth	NW10 0FW	40	Female	Heart
Michael	E17 9QY	23	Male	Respiratory
George	E17 3SF	24	Male	Liver
Simon	E17 5WD	29	Male	Cancer

Before Anonymization

Example: Research needs to be done on the types of disease

- Name, Postcode, Age, and Gender are attributes that could be used to identify an individual

K-anonymity – Example

Name	Postcode	Age	Gender	Disease
Patrick	SW1 4YB	22	Male	Heart
Sebastian	SW1 4ZE	23	Male	Respiratory
Reece	SW1 2HY	20	Male	No Illness
Tiffany	NW10 8FN	47	Female	Cancer
Abigail	NW10 4AB	42	Female	No Illness
Elizabeth	NW10 0FW	40	Female	Heart
Michael	E17 9QY	23	Male	Respiratory
George	E17 3SF	24	Male	Liver
Simon	E17 5WD	29	Male	Cancer

Before Anonymization

Postcode	Age	Gender	Disease
SW1 *	[20 – 29]	Male	Heart
SW1 *	[20 – 29]	Male	Respiratory
SW1 *	[20 – 29]	Male	No Illness
NW10 *	[40 – 49]	Female	Cancer
NW10 *	[40 – 49]	Female	No Illness
NW10 *	[40 – 49]	Female	Heart
E17 *	[20 – 29]	Male	Respiratory
E17 *	[20 – 29]	Male	Liver
E17 *	[20 – 29]	Male	Cancer

After Anonymization

Example: Research needs to be done on the types of disease

- Name, Postcode, Age, and Gender are attributes that could be used to identify an individual
- Data anonymised to achieve k-anonymity of k = 3, or at least 1/3 chance to identify an individual

Pseudonymization

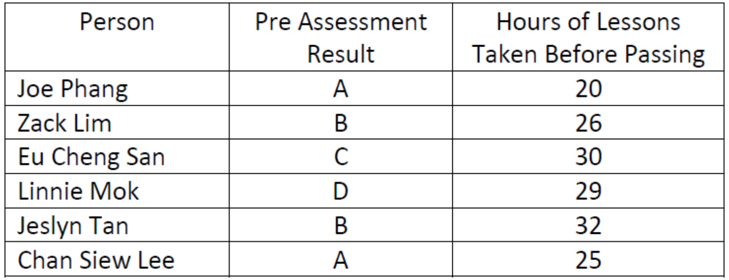
- Replacement of identifying data with made up values, which are unique, and should have no relationship to the original values

- Used when the data values need to be uniquely distinguished

- Persistent pseudonyms allow linkage across other different datasets

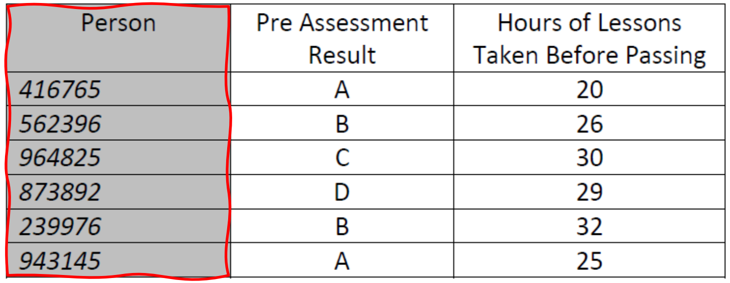
- May need to follow the structure or data type of the original value, simply to look more similar to the original attribute

Pseudonymization – Example



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

Before Anonymization



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

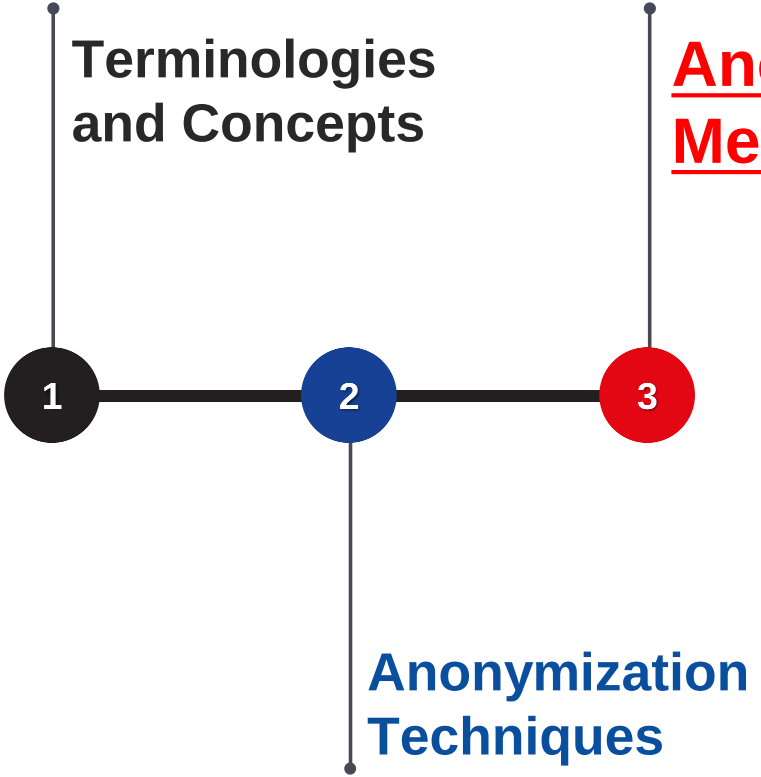
After Pseudonymization

Example: names of persons who obtained their driving licenses and other information

- the names were replaced with pseudonyms

Useful for cross dataset linking and where original data structure is needed, but does not comply with personal data protection regulations, if applied specifically on explicit identifiers

Contents



1 Terminologies and Concepts

2 Anonymization Techniques

3 Anonymization Methodology

Anonymization Methodology

- Refers to how the anonymised dataset will be released
- Public or Non-Public

Determine release model

1

Anonymization Methodology

- Refers to how the anonymised dataset will be released
- Public or Non-Public

Determine release model

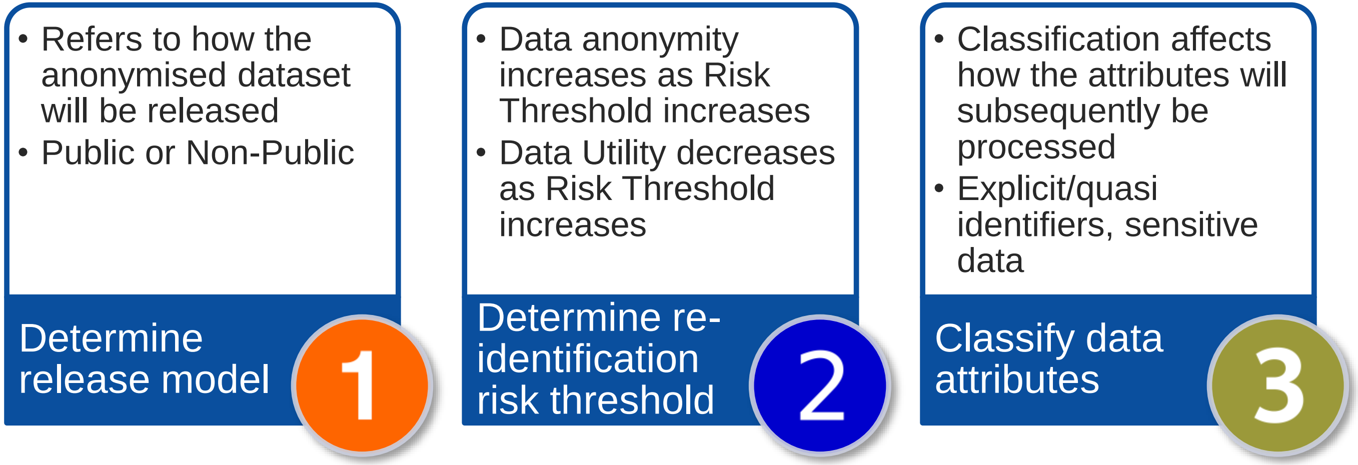
1

- Data anonymity increases as Risk Threshold increases
- Data Utility decreases as Risk Threshold increases

Determine re-identification risk threshold

2

Anonymization Methodology



- Refers to how the anonymised dataset will be released
- Public or Non-Public

Determine release model

1

- Data anonymity increases as Risk Threshold increases
- Data Utility decreases as Risk Threshold increases

Determine re-identification risk threshold

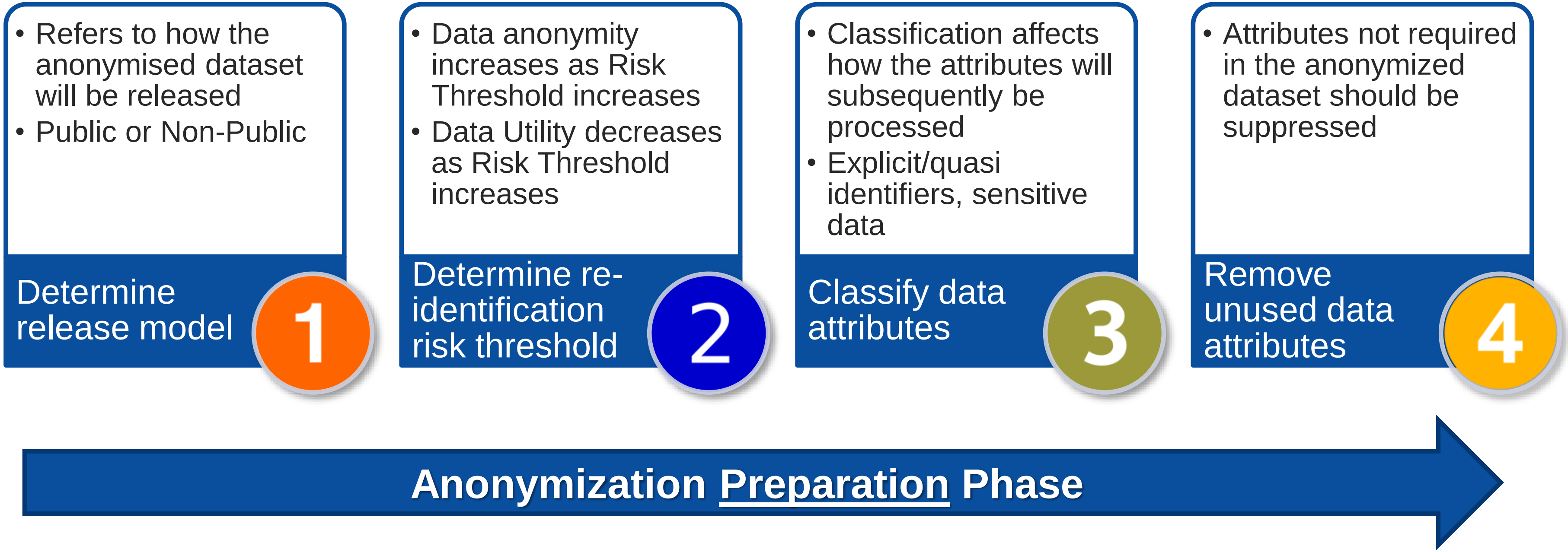
2

- Classification affects how the attributes will subsequently be processed
- Explicit/quasi identifiers, sensitive data

Classify data attributes

3

Anonymization Methodology



- Refers to how the anonymised dataset will be released
- Public or Non-Public

Determine release model

1

- Data anonymity increases as Risk Threshold increases
- Data Utility decreases as Risk Threshold increases

Determine re-identification risk threshold

2

- Classification affects how the attributes will subsequently be processed
- Explicit/quasi identifiers, sensitive data

Classify data attributes

3

- Attributes not required in the anonymized dataset should be suppressed

Remove unused data attributes

4

Anonymization Preparation Phase

Anonymization Methodology

- Apply relevant anonymization techniques
- Different techniques are applicable for types of identifiers

Anonymise identifiers

5

Anonymization Methodology

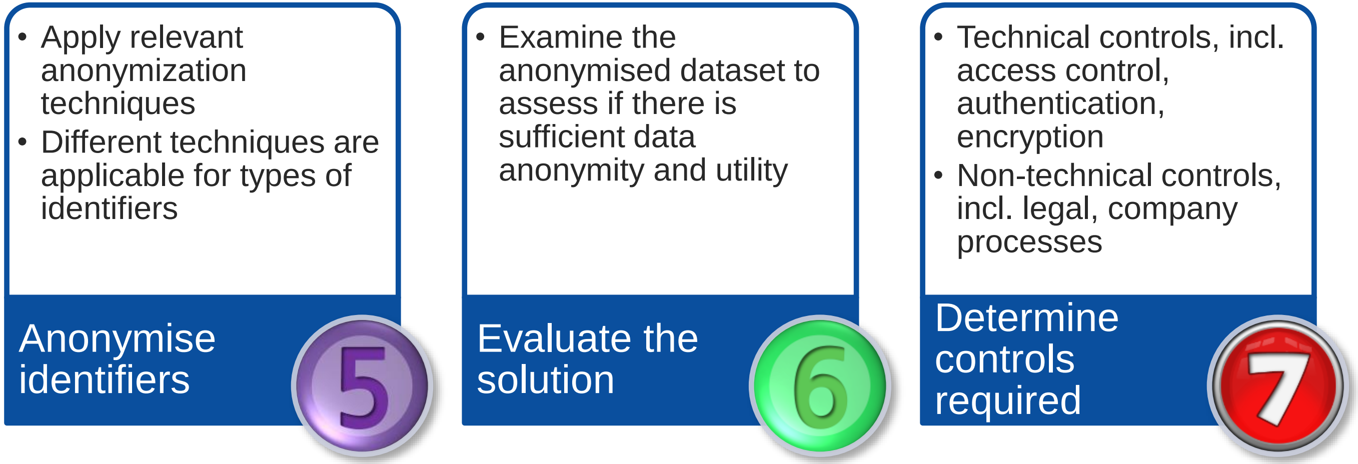
- Apply relevant anonymization techniques
- Different techniques are applicable for types of identifiers

Anonymise identifiers 5

- Examine the anonymised dataset to assess if there is sufficient data anonymity and utility

Evaluate the solution 6

Anonymization Methodology



- Apply relevant anonymization techniques
- Different techniques are applicable for types of identifiers

Anonymise identifiers

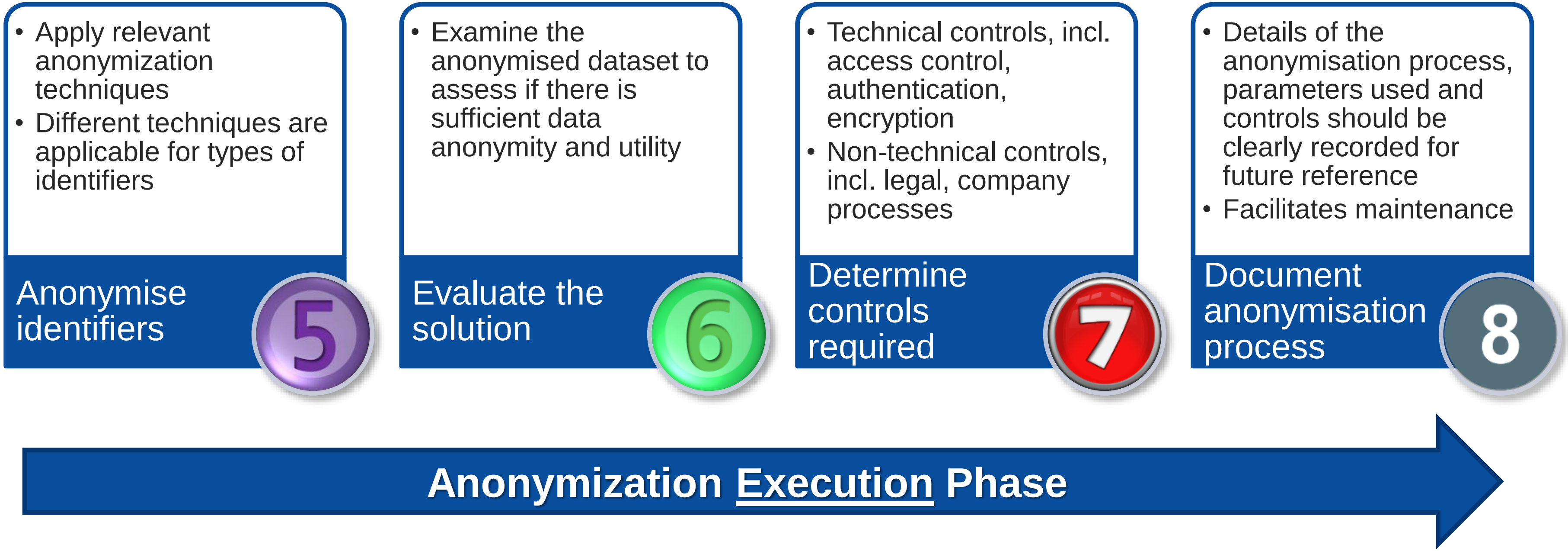
- Examine the anonymised dataset to assess if there is sufficient data anonymity and utility

Evaluate the solution

- Technical controls, incl. access control, authentication, encryption
- Non-technical controls, incl. legal, company processes

Determine controls required

Anonymization Methodology



- Apply relevant anonymization techniques
- Different techniques are applicable for types of identifiers

Anonymise identifiers 5

- Examine the anonymised dataset to assess if there is sufficient data anonymity and utility

Evaluate the solution 6

- Technical controls, incl. access control, authentication, encryption
- Non-technical controls, incl. legal, company processes

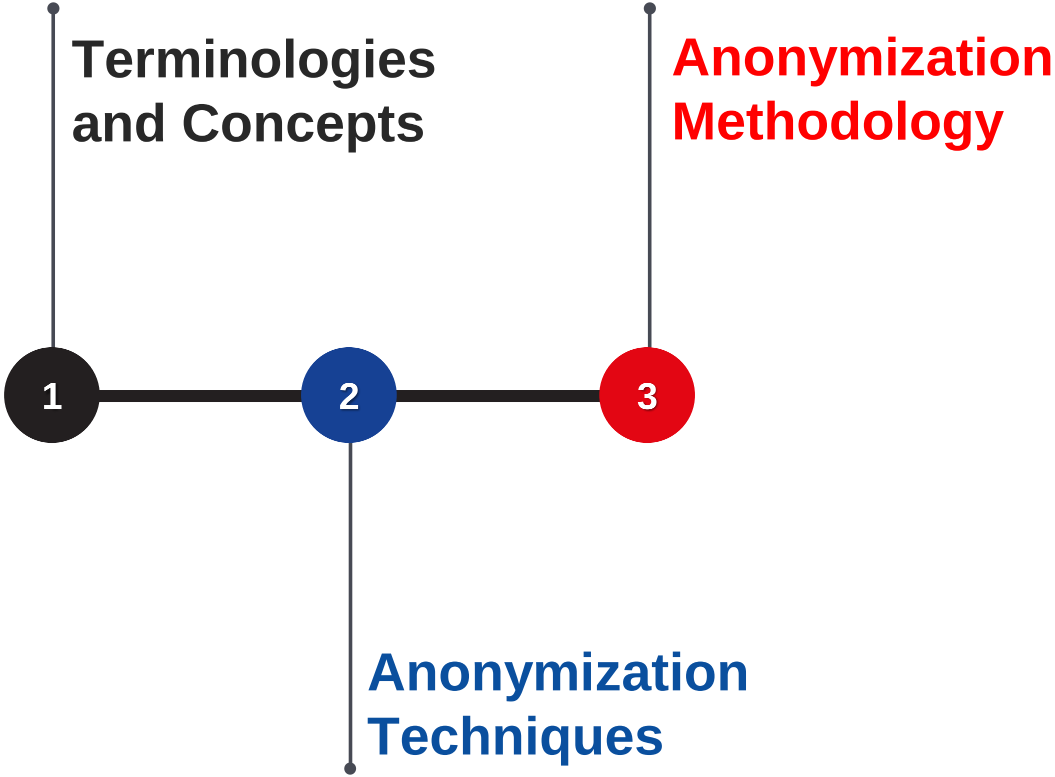
Determine controls required 7

- Details of the anonymisation process, parameters used and controls should be clearly recorded for future reference
- Facilitates maintenance

Document anonymisation process 8

Anonymization Execution Phase

Contents



Terminologies and Concepts
1
2
3
Anonymization Methodology
Anonymization Techniques