# HUM 4441
# ENGINEERING ETHICS

Dr. Mohammad Rezwanul Huq

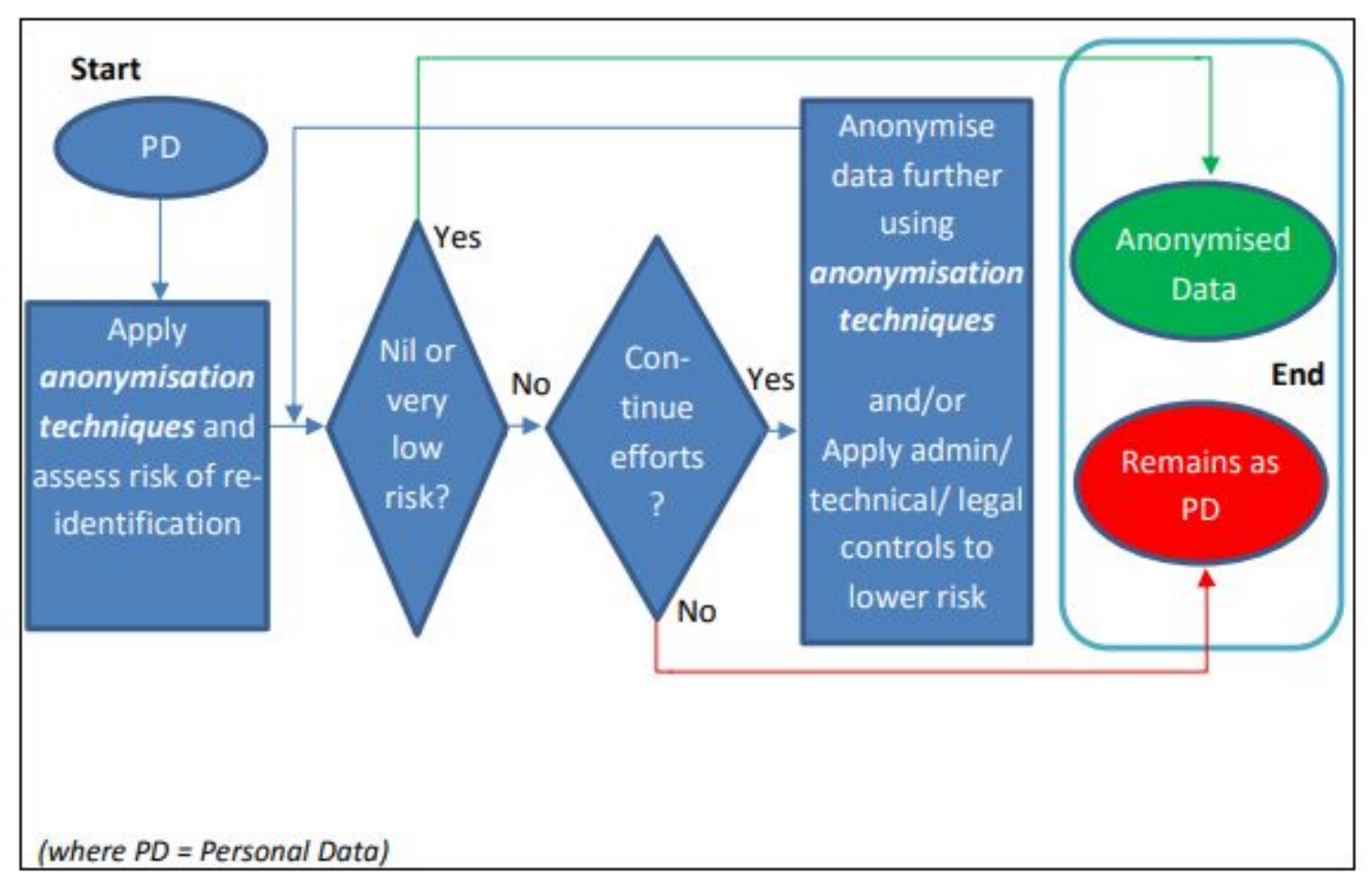
Adjunct Faculty, IUT

# BASIC DATA ANONYMIZATION TECHNIQUES

# DATA ANONYMIZATION

- “Data anonymization” refers to the conversion of personal data into “anonymized data” by applying a range of “anonymization techniques”.
- The process of data anonymization would be **irreversible.**
- However, there may be cases where the organization applying the anonymization retains the ability to recreate the original data from the anonymized data; In such cases, the anonymization process is **reversible**.

# ANONYMIZATION PROCESS



# DATA ANONYMIZATION TECHNIQUES

- Attribute Suppression
- Record Suppression
- Character Masking
- Pseudonymization
- Generalization
- Swapping (shuffling and permutation)
- Data Perturbation
- Data Aggregation

# ATTRIBUTE SUPPRESSION

- Attribute suppression refers to **the removal of an entire part of data** (also referred to as "column" in databases and spreadsheets) in a dataset.
- When an attribute is not required in the anonymized dataset, or when the attribute cannot otherwise be suitably anonymized with another technique.
- This technique should be applied at the start of the anonymization process, as it is an easy way to decrease identifiability at this point.

# EXAMPLE: ATTRIBUTE SUPPRESSION

Before anonymisation:

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

After suppressing the "student" attribute:

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

# RECORD SUPPRESSION

- 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.
- To remove outlier records which are unique or do not meet other criteria such as k-anonymity, and not to keep in the anonymized dataset.
- Outliers can lead to easy re-identification. It can be applied before or after other techniques (e.g. generalization) have been applied.

# CHARACTER MASKING

- Character masking is the change of the characters of a data value, e.g. by using a constant symbol (e.g. “*” or “x”). Masking is typically partial, i.e. applied only to some characters in the attribute.
- When the data value is a string of characters and hiding part of it is sufficient to provide the extent of anonymity required.

# EXAMPLE: CHARACTER MASKING

Before anonymisation:

| 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 |

After partial masking of postal code:

| 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 |

# PSEUDONYMIZATION

- The replacement of identifying data with made up values. Pseudonymization is also referred to as coding. Pseudonyms can be irreversible, where the original values are properly disposed and the pseudonymization was done in a non-repeatable fashion, or reversible (by the owner of the original data), where the original values are securely kept but can be retrieved and linked back to the pseudonym, should the need arises.
- When data values need to be uniquely distinguished and where no character or any other implied information of the original attribute shall be kept.

# EXAMPLE: PSEUDONYMIZATION

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 |

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 |

# GENERALIZATION

- A deliberate reduction in the precision of data. E.g. 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.
- Design appropriate data categories and rules for translating data. Consider suppressing any records that still stand out after the translation (i.e. the generalisation).

# EXAMPLE: GENERALIZATION

Before anonymisation:

| 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 |

After generalisation of Age and Address:

| 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 |

# SWAPPING

- The purpose of swapping is to rearrange data in the dataset such that the individual attribute values 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 analysis is required at intra-attribute level only.
- There is no need for analysis of relationships between attributes at the record level.

# EXAMPLE: SWAPPING

Before anonymisation:

| 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 |

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 dean | 5 Feb 1972 | Gold | 0 |

# DATA PERTURBATION

- The values from the original dataset are modified to be slightly different.
- It should be used for quasi-identifiers (typically numbers and dates) which may potentially be identifying when combined with other data sources, and slight changes in value are acceptable. This technique should not be used where data accuracy is crucial.
- Rounding data and adding noise can achieve it.

# EXAMPLE: DATA PERTURBATION

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 (shaded columns represent the affected attributes):

| 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 |

# DATA AGGREGATION

- Converting a dataset from a list of records to summarized values.
- Use aggregate functions such as sum, min, max, count, avg, median etc.
- It should be used when individual records are not required and aggregated data is sufficient for the purpose.

# EXAMPLE: DATA AGRREGATION

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

# SYNTHETIC DATA GENERATION

- This technique is slightly different as compared to the other techniques described in this Guide, as it is mainly used to generate synthetic datasets directly and separately from the original data, instead of modifying the original dataset.
- Typically, when a large amount of data is required for system testing, but the actual data cannot be used and yet the data should be “realistic” in certain aspects, like format, relationship among attributes etc.
- Synthetic dataset is often generated based on the statistics derived from the original dataset.