# Workshop 12

COMP20008 Elements of Data Processing

# Learning outcomes

By the end of this class, you should be able to:

- assess whether a data set satisfies k-anonymity and l-diversity
- apply generalisation techniques to anonymise data
- explain basic concepts in differential privacy

# *k*-anonymity

A data set satisfies *k*-anonymity if:

> *"every record is indistinguishable from at least k – 1 other records with respect to the set of quasi-identifiers"*

# Q1: *k*-anonymity

| job | birth | postcode | illness |
|---|---|---|---|
| Cat1 | * | 4350 | HIV |
| Cat1 | * | 4350 | HIV |
| Cat1 | 1955 | 5432 | Flu |
| Cat1 | 1955 | 5432 | Fever |
| Cat2 | 1975 | 4350 | Flu |
| Cat2 | 1975 | 4350 | Fever |

Consider the quasi-identifiers {`job`, `birth`, `postcode`}. What is the highest value of *k* for which this data is *k*-anonymous?

# Q1: *k*-anonymity

| job | birth | postcode | illness |
|---|---|---|---|
| Cat1 | * | 4350 | HIV |
| Cat1 | * | 4350 | HIV |
| Cat1 | 1955 | 5432 | Flu |
| Cat1 | 1955 | 5432 | Fever |
| Cat2 | 1975 | 4350 | Flu |
| Cat2 | 1975 | 4350 | Fever |

Consider the quasi-identifiers {`job`, `birth`, `postcode`}. What is the highest value of *k* for which this data is *k*-anonymous?

Each set of quasi-ids appears at least twice ⇒ data is 2-anonymous

# Q1: *k*-anonymity

| job | birth | postcode | illness |
|---|---|---|---|
| Cat1 | * | 4350 | HIV |
| Cat1 | * | 4350 | HIV |
| Cat1 | 1955 | 5432 | Flu |
| Cat1 | 1955 | 5432 | Fever |
| Cat2 | 1975 | 4350 | Flu |
| Cat2 | 1975 | 4350 | Fever |

Describe one possible privacy attack on this data.

# Q1: *k*-anonymity

| job | birth | postcode | illness |
|---|---|---|---|
| Cat1 | * | 4350 | HIV |
| Cat1 | * | 4350 | HIV |
| Cat1 | 1955 | 5432 | Flu |
| Cat1 | 1955 | 5432 | Fever |
| Cat2 | 1975 | 4350 | Flu |
| Cat2 | 1975 | 4350 | Fever |

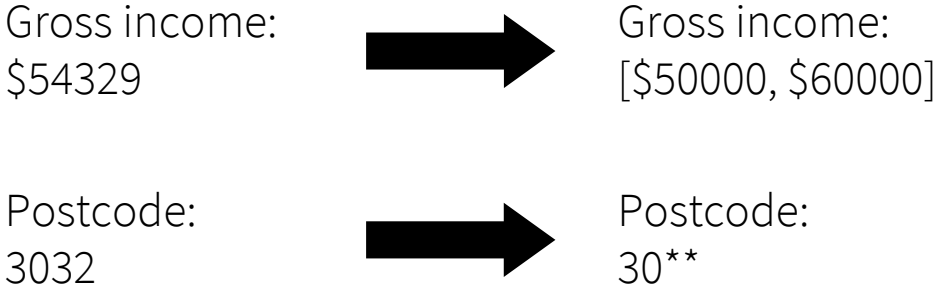
Describe one possible privacy attack on this data.

A *homogeneity attack*: when records with the same quasi-ids have the same value for the sensitive attribute, it is no longer protected.

E.g. we know anyone in this data with `job=Cat1` and `postcode=4350` has HIV.

# Privacy through generalisation

Make the quasi-identifiers less specific



# Q2: Generalisation

Consider the quasi-identifiers {name, gender, date_of_birth, zip_code}. Apply generalisation to make the following table 3-anonymous.

| name | gender | date_of_birth | zip_code | disease |
|---|---|---|---|---|
| Alice | F | 01/01/1981 | 11111 | Flu |
| Anne | F | 02/02/1981 | 11122 | Flu |
| Sonia | F | 12/03/1981 | 11133 | Flu |
| Bob | M | 12/01/1982 | 33311 | Heart disease |
| Shunsuke | M | 10/04/1982 | 33322 | Cold |
| Carl | M | 02/03/1982 | 33333 | Flu |

# Q2: Generalisation

Consider the quasi-identifiers {name, gender, date_of_birth, zip_code}. Apply generalisation to make the following table 3-anonymous.

| gender | date_of_birth | zip_code | disease |
|---|---|---|---|
| F | 1981 | 111** | Flu |
| F | 1981 | 111** | Flu |
| F | 1981 | 111** | Flu |
| M | 1982 | 333** | Heart disease |
| M | 1982 | 333** | Cold |
| M | 1982 | 333** | Flu |

# $\ell$-diversity

A data set satisfies $\ell$-diversity if:

> *"for each group of records with the same quasi-identifiers, there are at least $\ell$ distinct values for the sensitive attribute"*

# Q3: $\ell$-diversity

| age | zip | diagnosis |
|---|---|---|
| [21-28] | 9**** | Measles |
| [21-28] | 9**** | Flu |
| [21-28] | 9**** | Flu |
| [48-55] | 92*** | Cancer |
| [48-55] | 92*** | Obesity |
| [48-55] | 92*** | Obesity |

Consider the quasi-identifiers {age, zip}. What is the highest *k* for which this data is *k*-anonymous?

# Q3: $\ell$-diversity

| age | zip | diagnosis |
|---|---|---|
| [21-28] | 9**** | Measles |
| [21-28] | 9**** | Flu |
| [21-28] | 9**** | Flu |
| [48-55] | 92*** | Cancer |
| [48-55] | 92*** | Obesity |
| [48-55] | 92*** | Obesity |

Consider the quasi-identifiers {age, zip}. What is the highest *k* for which this data is *k*-anonymous?

Each set of quasi-ids appears at least 3 times
⇒ data is 3-anonymous

# Q3: $\ell$-diversity

| age | zip | diagnosis |
|---|---|---|
| [21-28] | 9**** | Measles |
| [21-28] | 9**** | Flu |
| [21-28] | 9**** | Flu |
| [48-55] | 92*** | Cancer |
| [48-55] | 92*** | Obesity |
| [48-55] | 92*** | Obesity |

Consider the quasi-identifiers {age, zip}. What is the highest $\ell$ for which this data is $\ell$-diverse?

# Q3: $\ell$-diversity

| age | zip | diagnosis |
|---|---|---|
| [21-28] | 9**** | Measles |
| [21-28] | 9**** | Flu |
| [21-28] | 9**** | Flu |
| [48-55] | 92*** | Cancer |
| [48-55] | 92*** | Obesity |
| [48-55] | 92*** | Obesity |

Consider the quasi-identifiers {age, zip}. What is the highest $\ell$ for which this data is $\ell$-diverse?

Within each set of equivalent records, there are at most 2 distinct values for the sensitive attribute.

Hence the data is 2-diverse.

# Q3: $\ell$-diversity

Describe one possible privacy attack on this data.

| age | zip | diagnosis |
|---|---|---|
| [21-28] | 9**** | Measles |
| [21-28] | 9**** | Flu |
| [21-28] | 9**** | Flu |
| [48-55] | 92*** | Cancer |
| [48-55] | 92*** | Obesity |
| [48-55] | 92*** | Obesity |

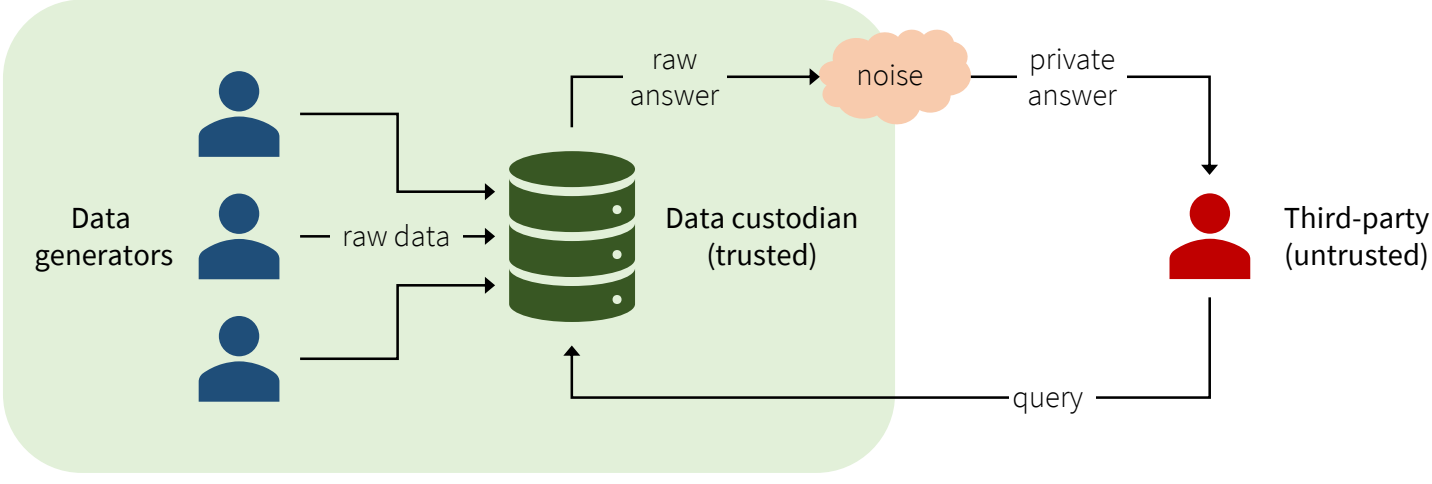
# Q3: $\ell$-diversity

| age | zip | diagnosis |
|---|---|---|
| [21-28] | 9**** | Measles |
| [21-28] | 9**** | Flu |
| [21-28] | 9**** | Flu |
| [48-55] | 92*** | Cancer |
| [48-55] | 92*** | Obesity |
| [48-55] | 92*** | Obesity |

Describe one possible privacy attack on this data.

Background attack: if I know a 50 year old with zip code 92031 is recorded in the data and they aren't obese, I can conclude that they have cancer.

# (Global) differential privacy



# Q4: Differential privacy

What's the global sensitivity $G$?

What's the privacy loss $k$?

How does the ratio $\frac{G}{k}$ affect the noise level?

# Q4: Differential privacy

What's the global sensitivity $G$?

The largest possible change that one record can make on the result of a query.

What's the privacy loss $k$?

Related to the confidence gained about the presence of a record in the data, based on the result of a query. Lower $k$ implies better privacy.

How does the ratio $\frac{G}{k}$ affect the noise level?

This is the amount of noise added when using the Laplace mechanism

# Q5: Global sensitivity

Consider a survey that collects marital status and sex from respondents.

What is the *global sensitivity* of the query: (CountFemale, CountMarried)?

| | Marital status | Sex |
|---|---|---|
| 1 | Married | Female |
| 2 | Single | Male |
| 3 | Single | Male |
| 4 | Single | Female |
| 5 | Married | Female |
| ⋮ | ⋮ | ⋮ |

# Q5: Global sensitivity

Consider a survey that collects marital status and sex from respondents.

What is the *global sensitivity* of the query: (CountFemale, CountMarried)?

Consider change in output after adding a record:

| | | | | |
|---|---|---|---|---|
| n | Married | Female | → | (1, 1) |
| n | Married | Male | → | (0, 1) |
| n | Single | Female | → | (1, 0) |
| n | Single | Male | → | (0, 0) |

Global sensitivity is 2 (using the L1 norm)

| | Marital status | Sex |
|---|---|---|
| 1 | Married | Female |
| 2 | Single | Male |
| 3 | Single | Male |
| 4 | Single | Female |
| 5 | Married | Female |
| ⋮ | ⋮ | ⋮ |

# Q5: Global sensitivity

What is the *global sensitivity* of the query: (CountMaleMarried, CountMaleSingle, CountFemaleMarried, CountFemaleSingle)?

| | Marital status | Sex |
|---|---|---|
| 1 | Married | Female |
| 2 | Single | Male |
| 3 | Single | Male |
| 4 | Single | Female |
| 5 | Married | Female |
| ⋮ | ⋮ | ⋮ |

# Q5: Global sensitivity

What is the *global sensitivity* of the query: (CountMaleMarried, CountMaleSingle, CountFemaleMarried, CountFemaleSingle)?

Consider change in output after adding a record:

| | | | | |
|---|---|---|---|---|
| n | Married | Female | → | (0, 0, 1, 0) |
| n | Married | Male | → | (1, 0, 0, 0) |
| n | Single | Female | → | (0, 0, 0, 1) |
| n | Single | Male | → | (0, 1, 0, 0) |

Global sensitivity is 1 (using the L1 norm)

| | Marital status | Sex |
|---|---|---|
| 1 | Married | Female |
| 2 | Single | Male |
| 3 | Single | Male |
| 4 | Single | Female |
| 5 | Married | Female |
| ⋮ | ⋮ | ⋮ |

# Revision

# 2018 Exam Q4

University X is planning to build a recommender system for its students. Based on subjects they have enrolled in, the system will recommend new subjects they might consider studying in the future.

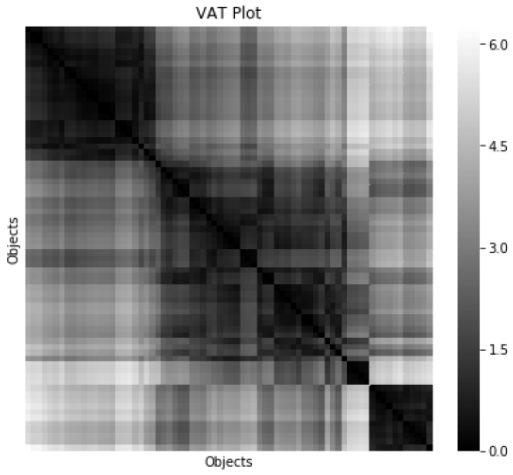
A table showing a fragment of the data input to the system is below. The columns correspond to the codes of all the subjects in the University handbook. Rows correspond to students and whether they have enrolled in a subject ("Yes" if they have previously enrolled and "-" otherwise). The dataset covers the period 2010-2017, with 100,000 students and 3000 subjects. It is proposed that subject recommendations should be made using user based collaborative filtering.

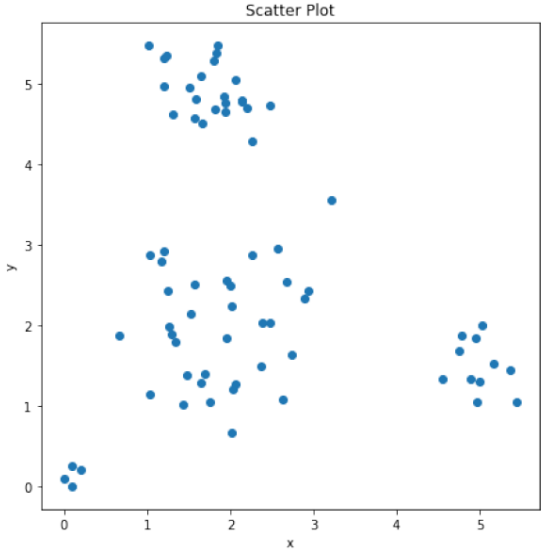
| PersonName | Subject1 | Subject2 | Subject3 | Subject4 | Subject5 | ... |
|---|---|---|---|---|---|---|
| Alice | Yes | - | - | - | - | ... |
| Bob | Yes | Yes | - | Yes | - | ... |
| Margaret | Yes | Yes | - | - | - | ... |
| ... | ... | ... | ... | ... | ... | ... |

Explain three challenges for making this recommendation approach effective.

# 2018 Exam Q4

# 2019 Exam Q3




(a) (1.5 mark) Are these likely to have come from the same dataset? Give two reasons why or why not.

(b) (1.5 marks) Consider a dataset with 10000 rows and 500 features. Give three reasons why we might want to apply PCA while analysing the dataset.

# 2019 Exam Q3

# 2019 Exam Q2b

(3 marks) Consider the following regular expression:

```
int[1-9]+=[A-Z][a-z]*
```

Which of the following terms would matched by this regular expression (write down all that apply):

```
int=Abc
int2=A
int3=Abc
int44=Azz
int5Abc
int6=abc
```

# 2019 Exam Q2b