VCE APPLIED COMPUTING · UNIT 1 OUTCOME 1 · RELATIONAL DATABASES · LESSONS 1 AND 3

Is it really de-identified?

A case study on personal information, re-identification, test data and collecting only what you need.

Part A: The Westbrook traits survey (fictional)

Westbrook College asked students to send a survey about inherited traits to friends and family. Before the spreadsheet was shared with a class to practise cleaning data, **the email addresses were changed** to random-looking codes, and the file was described as de-identified. These are its columns, with one invented example row.

Column	Example value (invented)	Your classification
RespondentEmail (changed)	qxyz1234@example.com	
Timestamp	14/03/2026 19:42	
StudentEmail of the sender (changed)	abcd0001@westbrook.example	
Connection to the student	Cousin	
Age band	16–20	
Sex	Female	
Height (cm)	165	
Distance from Melbourne CBD	Within 40 km	
Highest education	Currently completing school	
Ethnicity (family background)	(free text)	
Ear lobes / tongue rolling / handedness …	Attached / Yes / Left	

Classify each column: D = direct identifier · C = could identify someone when combined with other columns · S = sensitive information · N = unlikely to identify anyone.

1. **Motivated intruder.** You know the students in your class. Which columns would let you work out whose relative a row describes? Which *combination* is enough?
2. **Verdict.** The Office of the Australian Information Commissioner (OAIC) says information is de-identified only when the risk of re-identifying someone is very low *in the setting where it is shared*, and that removing direct identifiers is not always enough. Did changing the emails make this file de-identified? Give two reasons.
3. **Repair.** What else would you remove, group or drop before sharing it with a class?

Part B: Where should test data come from?

There are several ways to create test data for a database. Rate each one for the Ridgeview database, then choose one and justify it.

Source of data	Privacy risk (L/M/H)	Realistic? (L/M/H)	Bias or accuracy problems?	Use it for Ridgeview? Why?
Scrape data from the web				
Characters from films, books or anime				
Friends' and family's information				
A program that generates random data				
Generative AI (for example, Copilot)				
Random name and address generators				

Part C: Ridgeview, collect, transform or drop

The office has a wishlist for the new database. Its purpose is to **print activity rolls, email students and know which teacher runs each activity**. Under Australian Privacy Principle 3, an organisation should collect personal information only when it is *reasonably necessary* for its purpose. Decide for each field.

Field on the wishlist	Collect	Transform	Drop	Reason (link it to the purpose)
StudentID				
First and last name				
School email				
Year level				
Date of birth				
Home address				
Parent mobile number				
Medical conditions				
Student photo				
Activity and registration date				

Discuss: Medical conditions matter for a sports excursion. Does that mean this database should store them? Who should be able to see them?

Part D: Three levels of the same question

Level	Question	Your answer
Personal	Would you be comfortable if a classmate could find your relative's row in a 'de-identified' file?	
Interpersonal	When you build a test database, why should you not use your friends' real names and emails, even with permission?	
Societal	Large organisations combine datasets. Why does that make 'we changed the emails' a weaker promise every year?	

Teacher notes

- **Part A.** Changed emails are pseudonyms: they remove the direct identifier, but each sender keeps one code, so every row stays linked to one student. With the connection, age band, sex and height, a classmate can often work out whose relative a row describes. Ethnicity is sensitive information under the Privacy Act 1988 (Cth). Verdict: not de-identified in a classroom setting (OAIC, 2018). Repair: drop the sender code, the timestamp and any column the question does not need; group height and age.
- **Part B.** Friends' and family's information is real personal information, so it carries the highest risk and needs consent the class cannot easily get. Scraped data can contain personal information and breach a site's terms. Generative AI may reproduce real people or biased name patterns, so its output must be checked. Characters and random generators are safest. Ridgeview's data was generated by a program, with invented names and a reserved .example email domain.
- **Part C.** Collect: StudentID, name, school email, year level, activity and date. Drop: home address, parent mobile, photo and medical conditions (another school system holds these, with access controls). Date of birth: drop, since year level meets the purpose. Accept a well-argued 'transform' answer.
- **Why it is in the unit.** It makes the unit's ethics thread active: Lesson 1 asks why synthetic data is preferable, and Lesson 3 asks students to justify every field they create. It also previews Unit 2 Outcome 2 (cyber security) and Data Analytics Unit 4 Outcome 2 (data security).
- **Keep it safe.** Use only the invented example. Do not use a real school dataset as the case, and do not ask students to share their own traits or backgrounds.

Resource by Arvin Hay, part of the Ridgeview database unit. All values are invented. Sources: Office of the Australian Information Commissioner (2018), De-identification and the Privacy Act; OAIC (2026), Australian Privacy Principles guidelines.