

COMP8400 – Tutorial 1 questions
Data cleaning, pre-processing, integration, quality

Q1: What are the requirements of a data cleaning approach?

Q2: Can you imagine single-source data cleaning problems that do not occur in multi-source data cleaning situations?

Q3: In section 3 of the *Rahm and Do* paper, the *Backflow of cleaned data* is described. What can become a problem with this approach?

Q4: Expansion and correction of abbreviations is an important part of data standardisation. However, such standardisation can introduce new problems. Can you give an example?

Q5: At the top of page 8 in the *Rahm and Do* paper the *Soundex* method is mentioned. Can you describe what this method is and how it can be useful. (You might have to do some research to find more details about the *Soundex* method.)

Q6: There are many different data cleaning tools available. What is one of the main drawbacks of all of these tools (for example compared to database systems)?

Questions / issues you like to have discussed in this tutorial: