This question paper consists
of 3 printed pages each
of which is identified by the Code
Number (COMP5122M01)

This is an open book examination.
Any written or printed material is permitted.

School of Computing

January 2018

COMP5122M01

Data Science

Answer all 2 questions

Time allowed: 2 hours

Question 1

(a) The work of a data scientist may be divided into five high-level tasks: discovery, wrangling, profiling, modelling and reporting. As the interviews by Kandel et al. (2012) show, the discovery and wrangling tasks can be a significant bottleneck. Referring to experience you gained during the COMP5122M practical work, describe five reasons why discovery and wrangling are often time-consuming. **[10 marks]**

(b) This part of the question is about data validation rules. Imagine that a dataset has two variables: OPERATION and DATE. The DATE should be in the range 01/01/1930 to today's date (inclusive). However, if OPERATION is missing (indicated by NULL or the value X99) then there should be no value for DATE, so any existing value for DATE is invalid.

In the style that is used for Hospital Episode Statistics (HES) data, write validation rules that set missing dates to the value 01/01/1800 and invalid dates to the value 01/01/1801. **[6 marks]**

(c) Imagine that you work for a retailer called ABC that operates a loyalty card scheme, and has bought a retailer called XYZ that has a different loyalty card scheme. You have been asked to integrate the schemes to produce a single dataset. Both retailers currently store the following data about their customers:

- Loyalty Card Number
- Customer Name
- Address
- Postcode

(i) Describe how you could use an "any two from N" deterministic linkage method to produce a data table that has two variables: the ABC Loyalty Card Number and the XYZ Loyalty Card Number. **[2 marks]**

(ii) State four data quality issues that you could encounter, and would affect the data linkage. **[2 marks]**

[question 1 total: 20 marks]

Question 2

(a) Imagine that you work for a city council and are in charge of analysing data that has been returned by secondary schools in the city. There are 40 schools, with a total of 50,000 pupils. An extract of the data is shown below:

School	Class	Pupil Name	English Grade	Maths Grade
Boston Spa	Year 7	John Smith	A	
Horsforth	Year 8	Clare Jones	B	A
Guiseley	Year 9	David Khan	C	B
Wetherby	Year 10	Maya Wei		C
Pudsey	Year 11	John Stevens	D	D
Morley	Year 12	Ivy Wall		E
Garforth	Year 13	Ian Chasm	E	D
Etc.				

(i) Would you expect the data to be: Open Data, Shared Data, or Closed Data. Justify your answer. **[2 marks]**

(ii) State the type of each variable, choosing from the following terms: categorical, ordinal, numerical, or connection. **[2 marks]**

(iii) Your initial analysis shows that the School, Class and Pupil Name data are complete, but some of the English Grade and Maths Grade values are missing. Explain how you would use hierarchical classification to comprehensively investigate the origin of the missing data. In your answer, remember to explain any additional variables that you would derive. **[6 marks]**

(iv) Your team is overworked, so you need to subcontract the analysis to a specialist data science company, which you have worked with before and trust. Before providing data to the company, what is the most important data governance issue to address and how will you do that? **[2 marks]**

(b) Imagine that an organisation has brought together all of the data about its customers into one data warehouse, so that the data can be used in a multitude of data science projects. The organisation has also just appointed a Chief Data Officer (CDO) who needs to ensure that those projects adhere to a high standard of data governance.

(i) Write a memo from the CDO to all staff, providing guidance about use of the data warehouse in data science projects from the perspective of the collection stage of the data lifecycle. **[4 marks]**

(ii) State two principles of the Data Protection Act that concern the storage stage of the data lifecycle and, for each principle, describe a step that the CDO should take to ensure compliance with the Act. **[4 marks]**

[question 2 total: 20 marks]