Instructions

Answer each question in the space provided. You can write in pen or pencil. Marks are indicated next to each question. This exam paper consists of 2 parts and the total marks for the exam are 100.

This is a sample exam. It is not complete, in that it does nnot have the full complement of questions.

Part 1 (50 marks in total)

Multiple Choice Questions: This section is worth 50 marks. Each question is worth 1 mark. Identify the choice that best completes the statement or answers the question. There is only one best answer for each question. Sometimes two answers may appear feasable, but you are to pick the one you believe is the best. Mark your selection by placing a tick or a cross through your selected answer. If you change your selection during the review of your paper, prior to the end of the Examination, make sure that the alteration is clear.

Marking Scheme for Multiple Choice Questions:

- 1 mark for a correct answer
- 0 marks for a wrong or more than one answer
- 0 marks for no answer

QUESTION 1.1: Data sets
Important characteristics for data analysis of a data set are:

A. the software for processing it

B. its supporting metadata

C. the 3 Vs, velocity, volume and variety

D. all options are correct

QUESTION 1.2: Data separation
Separating data so individual departments manage their own:

A. is cheaper

B. allows Hadoop-style processing to be done more easily

C. causes problems because of inconsistencies across departments

D. is the preferred solution to managing volume and variety in large organisations

QUESTION 1.3: Database types
How does a graph database differ from a relational database?

A. graph databases are used for money transfers

B. graph databases are better at storing and analysing data interaction patterns

C. graph databases are used for storing graphics

D. they are the same

QUESTION 1.4: Database issues
Distributed databases, in-memory databases and RDBMSs are specifically designed to address the following issue:

A. the need for scalable systems

B. the need for cheaper systems

C. the need to handle semi-structured data

D. the need for security

QUESTION 1.5: Volume
Volume in the big data definitions is best:

A. is relative, as it varies with the kind of data and current processing capacity, and the task being performed

B. measured in terabytes back in the year 2001

C. measured in units relative to typical hard-drive capacity of the time

D. measured in yettabytes

QUESTION 1.6: Privacy
What is the technological reason for the continued increase in lack of privacy?

A. the flow of technology makes surveillance easier unless particular measures are set in place.

B. the increase in cybercrime and terrorism makes it a necessity.

C. the open internet and the cloud removes privacy.

D. it follows from Koomey's Law.

QUESTION 1.7: Spark
Spark was built into the Hadoop platform because:

A. it is easier to build on top of the Hadoop infrastructure

B. it is implemented on top of the basic Map-Reduce mechanism of Hadoop

C. to gain from the Hadoop brand-name

D. the same core programmer team did the initial development

QUESTION 1.8: The Vs importance
The 3Vs of big data are important because:

A. they are an industry standard

B. they are the basis for the development of more Vs (e.g. Value)

C. they are used to describe in what way a dataset may be too big to handle

D. they are from the influential Gartner Inc

QUESTION 1.9: Text data
Text data can also be:

A. structural metadata

B. a digital container

C. image data

D. markup language

QUESTION 1.10: NoSQL
The growth of NoSQL databases occurred because:

A. they were better suited for distributed implementation

B. variety, volume and specific processing demands of some classes of data challenges RDBMSs

C. they were more easily integrated with web client applications

D. enterprising database developers expanded in the niche markets of NoSQL

QUESTION 1.11: Significance testing errors
Which if the following is a correct use of significance testing. (Some may be corrected with a proper adjustment to significance levels, but ignore this.)

A. When results are negative, repeat the test in order to confirm the result.

B. Performing a second study afterwards to enlarge the result set.

C. Testing for many different affects in parallel in order to speed up experimental work.

D. Having many different institutions perform the testing to ensure reproducability.

QUESTION 1.12: Useful visualisation
Which of the following is visualisation not useful for:

A. to perform discovery visually

B. to perform significance tests

C. to explore the data during the initial cleaning process.

D. to explain results

QUESTION 1.13: Scientific method in medicine
Medicine has many aspects of Data Science. Which of the following is not a cause for poor results shared by the two:

A. Industry/career motivational bias.

B. Misuse of significance testing.

C. Correlation does not imply causation.

D. Computer simulation of a correct model to generate data.

QUESTION 1.14: Clinical trials
A clinical trial is primarily designed to:

A. apply the principle of intervention to test cause.

B. test correlation between treatments and outcomes.

C. stop scientists from cheating.

D. isolate the causes of outcomes.

QUESTION 1.15: Evaluating the results of learning
When evaluating and presenting the results of learning, it is not important to

A. keep a separate data set for unbiased testing.

B. use the standard significance level of 0.01.

C. work with a domain expert to understand the proper costs and benefits of different outcomes and errors.

D. record the processing steps for background and reproducibility.

Part 2 (50 marks in total)

Short Answer Questions: This section is worth 50 marks. Your answers should be written in clear, simple English and should be complete enough in addressing the question. Extensive prose is not required. Structured bullet points are acceptable.

Question 2.1 (2 marks)
What is "Linked Open Data"? Why is it called "linked" and why "open"? What sort of format can it be in?

Question 2.2 (2 marks)
Why is metadata important for data analysis?

Question 2.3 (2 marks)
SAS, DataWrangler and Python can all be used for data wrangling. Describe some characteristics of these tools that could be used to choose between them.

Question 2.4 (2 marks)
What is the Predictive Model Markup Language and what is it used for?

Question 2.5 (2 marks)
They say "correlation does not imply causation". Give an example of variables that are correlated but not causal.

END OF EXAM

Blank page for additional answers if needed.

Blank page for additional answers if needed.