

| | | |
|--|--|--|
| | | |
|--|--|--|

**Semester Two 2018
Examination Period**

Faculty of Information Technology

EXAM CODES: FIT1043

TITLE OF PAPER: Introduction to Data Science – PAPER 1

EXAM DURATION: 2 hours writing time

READING TIME: 10 minutes

THIS PAPER IS FOR STUDENTS STUDYING AT: (tick where applicable)

- | | | | | |
|------------------------------------|---|--|--|--|
| <input type="checkbox"/> Berwick | <input checked="" type="checkbox"/> Clayton | <input checked="" type="checkbox"/> Malaysia | <input type="checkbox"/> Off Campus Learning | <input type="checkbox"/> Open Learning |
| <input type="checkbox"/> Caulfield | <input type="checkbox"/> Gippsland | <input type="checkbox"/> Peninsula | <input type="checkbox"/> Monash Extension | <input type="checkbox"/> Sth Africa |
| <input type="checkbox"/> Parkville | <input type="checkbox"/> Other (specify) | | | |

During an exam, you must not have in your possession any item/material that has not been authorised for your exam. This includes books, notes, paper, electronic device/s, mobile phone, smart watch/device, calculator, pencil case, or writing on any part of your body. Any authorised items are listed below. Items/materials on your desk, chair, in your clothing or otherwise on your person will be deemed to be in your possession.

No examination materials are to be removed from the room. This includes retaining, copying, memorising or noting down content of exam material for personal use or to share with any other person by any means following your exam.

Failure to comply with the above instructions, or attempting to cheat or cheating in an exam is a discipline offence under Part 7 of the Monash University (Council) Regulations.

AUTHORISED MATERIALS

| | | |
|---|------------------------------|--|
| OPEN BOOK | <input type="checkbox"/> YES | <input checked="" type="checkbox"/> NO |
| CALCULATORS | <input type="checkbox"/> YES | <input checked="" type="checkbox"/> NO |
| SPECIFICALLY PERMITTED ITEMS if yes, items permitted are: | <input type="checkbox"/> YES | <input checked="" type="checkbox"/> NO |

Candidates must complete this section if required to write answers within this paper

STUDENT ID: _____

DESK NUMBER: _____

This page is intentionally left blank

Instructions

You must write all your answers in the Script-book and clearly indicate which question you are answering.

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 65 marks.

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

Part 1 (15 marks in total)

Multiple Choice Questions: This section is worth 15 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 feasible, 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. **Please Note in the final exam you will have 15 multiple choice questions.**

Marking Scheme for Multiple Choice Questions:

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

QUESTION 1.1: Value chains (1 mark)

Which of the following represents a typical value chain in data science?

- A. Collect > Wrangle > Analyse > Present
- B. Collect > Analyse > Wrangle > Present
- C. Wrangle > Collect > Analyse > Present
- D. Wrangle > Collect > Present > Analyse

QUESTION 1.2: Privacy (1 mark)

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 1.3: Data Science software and tools (1 mark)

Which of the following options consist of operating system, programming language, database and visualization tool respectively? **R is for visual ?**

- A. Window, R, SQL, Spark
- B. Unix, Java, MySQL, matplotlib
- C. Mac OS, Hadoop, Oracle, Visual Basic
- D. All of the options

QUESTION 1.4: Machine Learning

Machine learning is useful when:

- A. human expertise is not available
- B. ALL of the other cases**
- C. humans cannot explain their expertise (as a set of rules)
- D. humans are expensive to use for the work

QUESTION 1.5: Tasks

Which of these tasks might a data scientist typically perform?

- A. Pitching project ideas.
- B. Collecting and cleaning data.
- C. ALL of the three other options.**
- D. Integrating and Interpreting data.

QUESTION 1.6: Python

Which of the following statements about Python is TRUE?

- A. The first element of an array in Python has the index 1.
- B. Python is a scripting language.**
- C. Python was designed by statisticians.
- D. Python is a proprietary programming language.

QUESTION 1.7: Shell commands

Unix shell commands like “less” and “grep”:

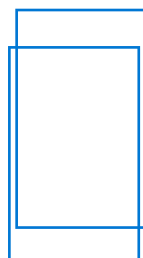
- A. can be used to manipulate large data files easily**
- B. are poorly documented
- C. are examples of technology that is too old to be useful to a modern data scientist
- D. are used to fit regression tree models

QUESTION 1.8: Disks

Over the years, disk capacity is generally growing:

- A. quadratically
- B. logarithmically
- C. linearly
- D. exponentially**

dont understand the word but know the graph



Part 2 (50 marks in total)

Short Answer Questions: This section is worth 50 marks and each question is worth 2 marks. Your answer 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. **Please Note in the final exam you will have 25 multiple choice questions.**

QUESTION 2.1: Data scientist role (2 marks)

Name two different styles (roles) of data scientists and their responsibilities.

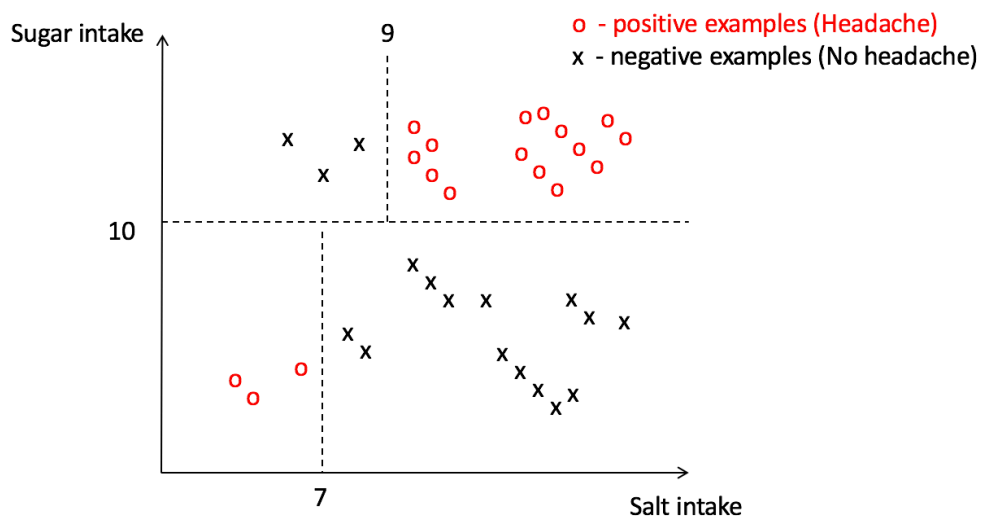
data analysis and chief data scientist.

chief: data engineering and ____ (monitoring)

analysis: collect, wangling, analysing,

QUESTION 2.2: Decision trees (2 marks)

Consider we are given the sugar intake and salt intake of 35 people and also the information about whether they have a headache or not. The goal is to build a decision tree to predict whether they have headache or not. Also, suppose that the feature space has been recursively partitioned according to the following figure:



Build the relevant decision tree for the above partitioning.

dont understand it

QUESTION 2.3: Metadata

List some types of metadata might be associated with an image.

integer, and the status of people, if they feel headache or not

QUESTION 2.4: Predictive model

Assuming you are collecting data about traffic accidents in Melbourne in order to develop a predictive model. Would it be better to collect “more data” (e.g. the locations of accidents over many years) or “more types of data” (e.g. the types of vehicles involved, the weather conditions, etc.)? Give a brief justification.

more data can make the model get more accuracy when it make prediction. but we also need to make sure the data is enough for modelling. If there are not so much data about location then we shouldnt include it in the collection

QUESTION 2.5: Implicit data

Give an example of implicit data that reveals personal information about a user. Describe the regular data that lies behind the implicit data and then describe the implicit data and why it is implicit.

: Data inferred by the use of other data, e.g. facebook like, if u like football post in facebook, facebook will push more about football post to user

QUESTION 2.6: Confidentiality [review it](#)

Would you consider users' emails to be sensitive information? Why or why not?

user email: the user email can get the phone number and user account, and it can be use in scam, and it

QUESTION 2.7: Unix Shell

Why are pipes and redirects in the Unix Shell useful for dealing with big data?

because we can use pipe to

pipe is kind of buffer,

END OF EXAM