Office Use Only

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

MONASH University

# 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)***

| | | | | |
|---|---|---|---|---|
| ☐ Berwick | ✓ Clayton | ✓ Malaysia | ☐ Off Campus Learning | ☐ Open Learning |
| ☐ Caulfield | ☐ Gippsland | ☐ Peninsula | ☐ Monash Extension | ☐ Sth Africa |
| ☐ Parkville | ☐ 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** | ☐ **YES** | ✓**NO** |
| **CALCULATORS** | ☐ **YES** | ✓ **NO** |
| **SPECIFICALLY PERMITTED ITEMS**<br>**if yes, items permitted are:** | ☐ **YES** | ✓ **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?

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

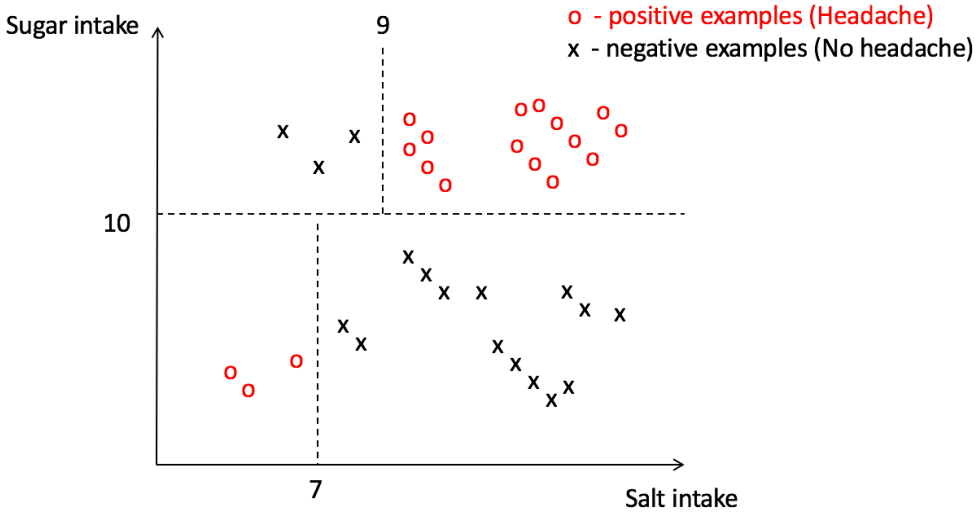
# Part 2 (50 marks in total)

**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.

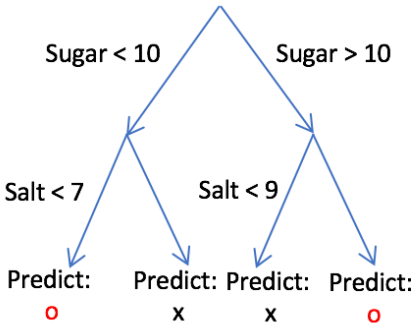
**QUESTION 2.1: Data scientist role (2 marks)**
Name two different styles (roles) of data scientists and their responsibilities.
Answer. Any of the following two:

- Data developers: people focused on the technical problem of managing data--how to get it, store it, and learn from it
- Data researchers: people with an academic research background, using their training to "understand complex processes"
- Data businesspersons: most focused on the organization and how data projects yield profit
- Data creatives: the broadest of data scientists, those who excel at applying a wide range of tools and technologies to a problem

**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.
Answer:




**QUESTION 2.3: Metadata**
List some types of metadata might be associated with an image.

Answer: location (lat/long coordinates), time/date, image information (size, encoding, etc.), camera information (make, focal length, etc.

**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.

Answer: Usually more types of data helps a predictive model more than just collecting more data.
(Assumes that there is sufficient data for building a predictive model to start with.)

**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.

Answer: Data inferred by the use of other data. A basic example is a users "Likes" on Facebook used to (statistically) infer political preference where political preference was not considered in the "Likes" dataset. Or, if a Facebook user has liked a number of pages to do with Adelaide sports teams, then the user's home city is Adelaide represents implicit data.

**QUESTION 2.6: Confidentiality**
Would you consider users' emails to be sensitive information? Why or why not?

Answer: Yes. - They contain all sorts of private information: addresses, credit card
numbers, phone numbers, etc.

**QUESTION 2.7: Unix Shell**
Why are pipes and redirects in the Unix Shell useful for dealing with big data?

Answer: Pipes are buffered, so they don't load the whole le into memory. Can process the file one line at a time (as a stream).

END OF EXAM