UNIVERSITY OF MALAYA

MIDDLE SEMESTER EXAMINATION FOR THE DEGREE OF MASTER OF DATA SCIENCE

ACADEMIC SESSION 2018/2019 : SEMESTER 1

WQD7001 : Principles of Data Science

OCTOBER 2018 Time : 1½ hours

INSTRUCTIONS TO CANDIDATES :

Answer **ALL** questions (50 marks).

**NAME**: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**MATRIC NO.**: **WQD 1\_\_ \_\_ \_\_ \_\_ \_\_**

**ANSWER SCHEME**

(This question paper consists of 5 questions on 6 printed pages)

1. Multiple-choice questions. **Underline** the **ONE** best answer

a) Which of the following is performed by a data scientist?

i. Create reproducible code.

ii. Analyze data for actionable insights.

iii. Challenge results.

iv. All of the functions mentioned.

b) The Command Line Interface (CLI) can help you do the following, EXCEPT?

i. Navigate folders.

ii. Create files, folders, and programs.

iii. Choose an action using menus.

iv. Run computer programs.

c) Suppose you have forked a repository called datascientist on Github but it isn’t on your local computer yet.

Which of the following is the command to bring the directory to your local computer? (For this question assume that your user name is wqd7001)

1. git pull https://github.com/ wqd7001/datascientist.git
2. git clone https://github.com/ wqd7001/datascientist.git
3. git init
4. git pull datascientist master

d) Which of the following big data characteristics are more concerned to data science?

i. Variety ii. Velocity iii. Volume iv. Volatility

e) Point out the **wrong** statement:

i. The big volume indeed represents Big Data.

ii. Big Data is just about lots of data.

iii. The data growth and social media explosion have changed how we look at the data.

iv. All of the Mentioned.

f) Which of the following is characteristic of Raw Data?

i. Data is ready for analysis.

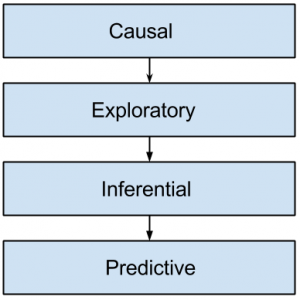
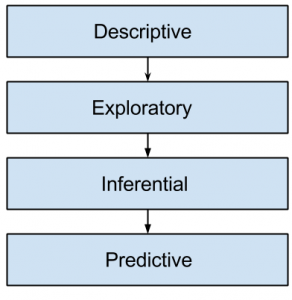
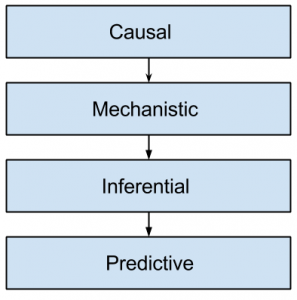
ii. Original version of data.

iii. Easy to use for data analysis.

iv. None of the Mentioned.

g) Which of the following figure correctly shows approximate order of difficulty?

i. ii iii

h) Which of the following is common goal of statistical modelling?

i. Inference ii. Summarizing iii. Subsetting iv. Descriptive

i) All data science investigations start with an existing dataset.

i. TRUE ii. FALSE

j) Most data scientists spend the majority of their time developing new models.

i. TRUE ii. FALSE

[10 marks]

1. (a) Why was it not possible for data science to exist 20 years ago?.

[4 marks]

*Data science was driven by technology change, thus it was impossible to exist 20 years ago (slow computers, low computational power, primitive programming languages, etc.)*

(b) Distinguish between **data analysis**, **data analytics** and **data mining**. Include a diagram in your explanation.

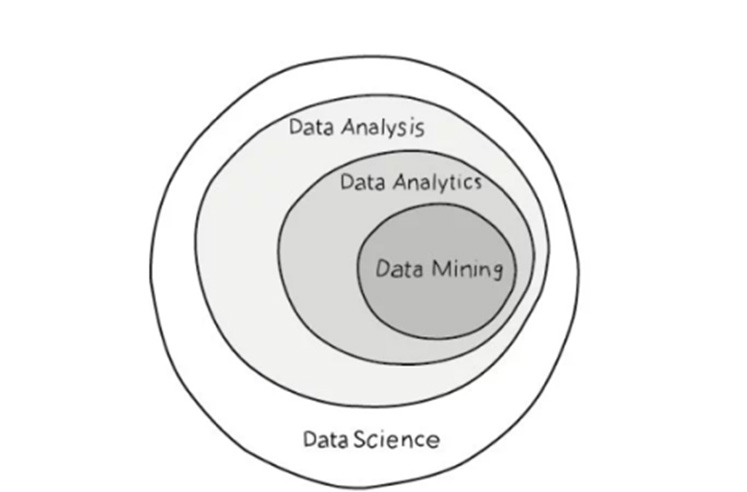
[6 marks]

***Data Analysis*** *involves extracting, cleaning, transforming, modeling and visualization of data with an intention to uncover meaningful and useful information that can help in deriving conclusion and take decisions. Data Analysis as a process has been around since 1960’s.*

***Analytics*** *is about applying a mechanical or algorithmic process to derive the insights for example running through various data sets looking for meaningful correlations between them.*

*Data analysis is a broader term that refers to the process of compiling and analysing data in order to present findings to management to help inform business decision making.* ***Data analytics*** *is a subcomponent of data analysis that involves the use of technical tools and data analysis techniques.*

***Data mining*** *is a systematic and sequential process of identifying and discovering hidden patterns and information in a large dataset. It is also known as Knowledge Discovery in Databases. It has been a buzz word since 1990’s.*



1. (a) Digital disruption happens when advances in technology change our

markets and our societies. Digital disruption has already happened.

Digital disruption is threatening the survival of many businesses and industries.

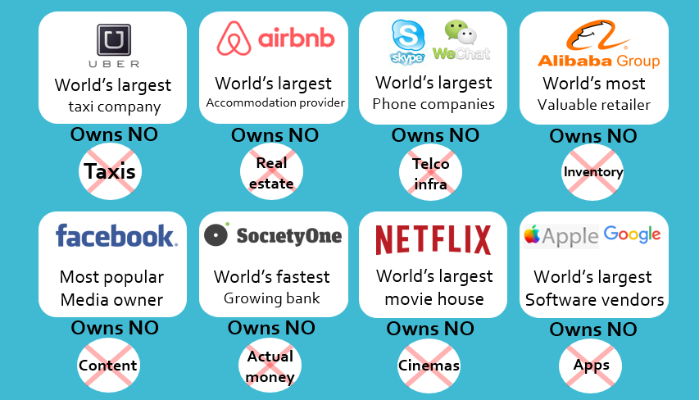
Identify **TWO** industries most vulnerable to digital disruption and show **THREE** successful digital disruption examples.

[6 marks]

*Industries most vulnerable to digital disruption includes:*

* *Media and Entertainment*
* *Technology products and services*
* *Financial services*
* *Retail*
* *Communications*
* *Education*

*Successful digital disruption examples:*



(b) Why is datafication important? Provide **TWO** examples of datafication.

[4 marks]

*Datafication is a modern technological trend turning many aspects of our life into computerised data and transforming this information into new forms of value.*

*The transition of the world into ever increasing usage and applications of digitization has created the need to create a system that can effectively handle all the information that is flowing around the globe. Therefore, datafication has evolved as a necessity of increasing digitization, aimed at creating value for businesses and individuals.*

***Some examples of datafication*** *are:*

* *Facebook datafies our friendships and posts*
* *Twitter datafies our followers, following, Tweets, time of day, and interactions*
* *LinkedIn datafies our professional contacts, locations, likes, posts*
* *Fitbit datafies our physical activities to derive useful information*
* *GPS devices on smartphones, such as Google maps, are able to track where we are at certain times of the day.*

1. Data science is an ocean of information. This is one of the major challenges faced by many aspiring data scientists. As a self-made data scientist, you are invited to give enlightenment on this subject.

(a) What are **TWO** useful learning resources a data scientist can refer to?

[4 marks]

*Answers varies ...*

* ***Data Science E-Books*** *- Journey to Data Scientist: Interviews with More Than Twenty Amazing Data Scientists, The Data Science Handbook, Art of Data Science, etc.*
* ***Data Science Blogs*** *- Data Science 101, Data Science Central, FiveThirtyEight, Dataconomy, etc.*
* ***Data Science******Community*** *– Data Science Community, Kdnuggets, Towards Data science, Analytics Vidhya, R-bloggers, Data Tau, etc.*
* ***Newsletter*** *- Data Science Weekly, Data Elixir, Mode Analytics, etc.*
* ***Online Courses And Tutorials*** *– Udemy, Coursera, Udacity, DataCamp, EdX, Dataquest, etc.*

(b) Who is an inspiring data scientist that you would like to recommend people to follow? State his/her organization. (e.g. DJ Patil from Devoted Health, San Francisco)

[2 marks]

*Answers varies ...*

* *Kenneth Cukier – Author - Big Data: A Revolution That Will Transform How We Live, Work, and Think, Artificial Intelligence*
* *Bernard Marr - founder and CEO of the Advanced Performance Institute*
* *Hilary Mason - founder of Fast Forward Labs, was chief scientist at Bitly, Inc.*
* *Andrew Ng - founder and chairman of the board at Coursera*.
* *Lillian Pierson - founder and chief data scientist at Data-Mania*

(c) Where would you recommend the newbies to join free training programs?

[2 marks]

*Free online Data Science courses & MOOCs available at:*

* *Coursera, DataCamp, EdX, Udemy, etc.*
* *IBM, Microsoft, Google*
* *Big Data University*
* *MIT, Stanford, Univ of Washington, John Hopkins University, etc.*

(d) Where can they find datasets to do a data science project?

[2 marks]

* + Open Data from Government and UN/World Bank websites - [https://data.gov.uk/data/](https://data.gov.uk/data/search)
  + Academic websites – e.g. Univ of California, Irvine – <http://archive.ics.uci.edu/ml/datasets.html>
  + Kaggle & Datascience resources - Titanic Survival Analysis – <https://www.kaggle.com/c/titanic>

1. Argue why dropping observations that have missing values and imputing the missing values based on other observations are not recommended. Suggest the solutions on how to handle missing values for categorical and numerical data.

[10 marks]

***Argument Against***

* *Dropping observations that have missing values.*
* *Imputing the missing values based on other observations.*

*Dropping missing values is sub-optimal because when you drop observations, you drop information.*

*The fact that the value was missing may be informative in itself.*

*“Missingness" is almost always informative in itself, and we should tell your algorithm if a value was missing.*

***Suggestion***

*Label as missing -The key is to tell your algorithm that the value was originally missing.*

***Missing categorical data***

*The best way to handle missing data for categorical features is to simply label them as ’Missing’! We’re essentially adding a new class for the feature. This tells the algorithm that the value was missing.*

***Missing numeric data***

*For missing numeric data, you should flag and fill the values.*

*Flag the observation with an indicator variable of missingness.*

*Then, fill the original missing value with 0 just to meet the technical requirement of no missing values.*

*By using this technique of flagging and filling, we are essentially allowing the algorithm to estimate the optimal constant for missingness, instead of just filling it in with the mean.*

**END**