UNIVERSITY OF MALAYA

MIDDLE SEMESTER EXAMINATION FOR THE DEGREE OF MASTER OF DATA SCIENCE

ACADEMIC SESSION 2019/2020 : SEMESTER 1

WQD7001 : Principles of Data Science

NOVEMBER 2019 Time : 1½ hours

INSTRUCTIONS TO CANDIDATES :

\_\_\_\_\_\_ / 50

20% - \_\_\_\_\_

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

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

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

**ANSWER SCHEME**

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

1. (a) What gave birth to data science?

[4 marks]

***ANSWER:***

*Data is the new oil for all the industries and data science is the electricity that powers the industry.*

*In today’s world data has become super-abundant and is going to increase exponentially for the next two decades. Before two or three decades, the data which we had with us was tiny, structured, and most of a single format and then the analytics performed was quite simple.*

*But with the rise of technology, this data started to explode, multiple sources started to generate huge amounts of unstructured data of different formats.*

*In other words, we had a lot of data with us, but we were not able to find out any insights from it. The need to understand and analyze data to make better decisions is what gave birth to Data Science.*

(b) Getting a ride from Grab is easy. Relate where data science is applied in this apps.

[3 marks]

***ANSWER:***

*Getting a ride from Grab is easy. Simply you open the app, set your pickup and drop location, book a taxi, get picked up, and pay with your phone.*

*Whenever you book a taxi through Grab you get an estimated fare and time to cover the specific distance. How these apps are able to show all the information? The answer is data science. Data Science predictive analysis is helping Grab to give pickup, drop location and time for arriving.*

(c) Referring to the three statements below, identify what is **X**, **Y** and **Z**. Assuming you have a students dataset, give **ONE** related example for each X,Y and Z.

Something that is apparent from the data / data set is called **X**.

A conclusion drawn from X is known as **Y**.

An action taken/to be taken from the Y is referred as **Z**.

[5 marks]

|  |  |
| --- | --- |
| **X** is ***findings*** | **Example of X:** *Customers are not watching the entire video to its full length. They are watching 90–95%* |
| **Y** is ***insights*** | **Example of Y:** *The parts they are not watching are the title roll and the end credits.* |
| **Z** is ***recommendations*** | **Example of Z:** *Introduce ‘Skip Intro’ at the beginning of title rolls and ‘Watch Next’ at the beginning of end credits. Benchmark 90–95% watched content as completed and measure if customers move to the next video in the series.* |

1. (a) We need to rebuild our education system to support data-driven education. Discuss what data-driven education means, including its prerequisites.

[4 marks]

***ANSWER:***

*Data driven means that progress in an activity is compelled by data rather than by intuition or personal experience. An organization must be collecting data. Data must be accessible and queryable. People with skills to use the data, extract the right data and use that data to inform next steps.*

(b) Waze is *datafying* our driving. What does that mean?

[4 marks]

***ANSWER:***

-------------------------------------------------------------------------------------------------------

-------------------------------------------------------------------------------------------------------

-------------------------------------------------------------------------------------------------------

-------------------------------------------------------------------------------------------------------

-------------------------------------------------------------------------------------------------------

(c) The hallmark of a data-driven organization is an effective “analytics value chain. Draw the Analytics Value Chain.

[4 marks]

***ANSWER:***



*The analytics value chain. In a data-driven organization, the data feeds reports, stimulating deeper analysis. Analysis is placed in the hands of the decision makers who incorporate them into their decision-making process, influencing the direction that the company takes and providing value and impact.*

1. (a) The 5 C’s is a framework for implementing the golden rule for data and the guidelines that help data scientists think about building data products.

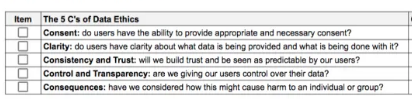
Explain all the **FIVE (5)** elements of the framework.

*Write your answer on the back page of the paper*.

[10 marks]

***ANSWER:***

*5 C’s (Consent, Clarity, Consistency, Control and Consequences)*



(b) What is the main lesson that you had learnt from Viktor Mayer Schönberger video on “The Right to be Forgotten”?

[2 marks]

***ANSWER:***

*You are ultimately responsible for protecting your image and personal information in the world of social networking.*

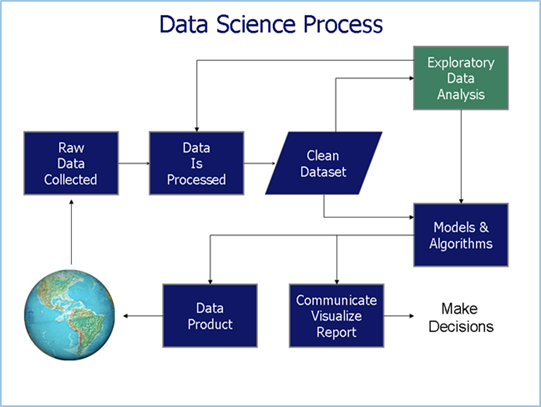
*“The Right to Be Forgotten” – an argument that asks “Do people have the right to remove damaging information about themselves on the Internet so the information can be forgotten?”*

1. You’ve been assigned a project to solve a data problem. As a future data scientist, you need to translate the task into a concrete problem, figure out how to solve it and present the solution to your examiners. Illustrate the data science process that you will apply for the project using a diagram.

*Write your answer on the back page of the paper*.

[10 marks]

***ANSWER:***

****

5. Solve the cross-word puzzle.

[14 marks]

**END**

**Crossword Puzzle**

1. Solution to replace Unicorn data scientist is have a data science team.
2. Data democratization is the idea that digital information should be accessible and understandable to the average end user as a basis for decision-making.
3. You may need to make FOI requests when you want to get more detail on the projects that government money is funding.
4. Dark data is data which is acquired through various computer network operations but not used in any manner to derive insights or for decision making.
5. Twitter data is a great resource for performing the tasks like opinion mining, sentiment analysis, and can be accessed using the Twitter API.
6. Data scientists spend 80% of their time cleaning and manipulating data and only 20% of their time actually analyzing it.
7. Data cleaning is the process of identifying and fixing or removing inaccurate information from your data ecosystem. Data cleaning is closely related to data quality.
8. An outlier is a data point that is considered extremely far from other points
9. Missing data can add BIAS to a model which result in overestimate or underestimate value.
10. Data tidying is the process of structuring your data in a way that makes it easy to analyze and use.
11. A codebook describes the contents, structure, and layout of a dataset collection.
12. Garbage-In-Garbage-Out is the idea that the output of an algorithm, or any computer function for that matter, is only as good as the quality of the input that it receives.
13. Web scraping is the process of pulling data from a website’s source code.
14. Hadoop is a tool that regulate the storage of massive datasets.