UNIVERSITY OF MALAYA

MIDDLE SEMESTER EXAMINATION FOR THE DEGREE OF MASTER OF DATA SCIENCE

ACADEMIC SESSION 2017/2018 : SEMESTER I

WQD7001 : Principles of Data Science

Nov 2017 Time : 2 hours

INSTRUCTIONS TO CANDIDATES :

Answer **ALL** questions (60 marks).

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

**MATRIC NO.**: **WQD17**\_\_ \_\_ \_\_ \_\_

**ANSWER SCHEME**

(This question paper consists of 5 questions on 3 printed pages)

1. Multiple-choice questions. Underline the **ONE** best answer.
2. What is the most important thing in Data Science?
   1. The data.
   2. Hacking skills.
   3. Working with large data sets.
   4. **The question you are trying to answer.**
3. What is the reason behind the explosion of interest in big data?
   1. We recently discovered ways to use data to answer scientific and business questions.
   2. **The price and difficulty of collecting and storing data has dramatically dropped**.
   3. We recently discovered ways to use data to make predictions.
   4. We have better experimental design now than previously.
4. Point out the correct statement:
5. **Raw data is original source of data**.
6. Preprocessed data is original source of data.
7. Raw data is the data obtained after processing steps.
8. None of the mentioned.
9. Which of the following is most important language for Data Science ?
10. Java
11. Ruby
12. **R**
13. None of the mentioned.
14. Which of the following is one of the key data science skill ?
15. Statistics
16. Machine learning
17. Data visualization
18. **All of the mentioned**.
19. Point out the correct statement:
20. Nearly 80% of data analysis is spent on wrangling data.
21. Nearly 20% of data analysis is spent on data dredging.
22. **Nearly 80% of data analysis is spent on the cleaning and preparing data**.
23. None of the mentioned.
24. Which of the following is a trait of tidy data ?
25. each variable in one column.
26. **each observation in different row**.
27. one table for each kind of variable.
28. None of the mentioned.
29. Which of the following is used to extract data from HTML code of websites ?
    1. **Web scraping** ii. Web dredging

iii. Web cleaning iv. All of the mentioned

1. Data that is used to structure and constrain other data, typically stable information with a known set of values that rarely change.

i. Metadata ii. Soft data iii. Qualitative data iv. **Reference data**

1. Number of family members in different families in a town is an example of a:

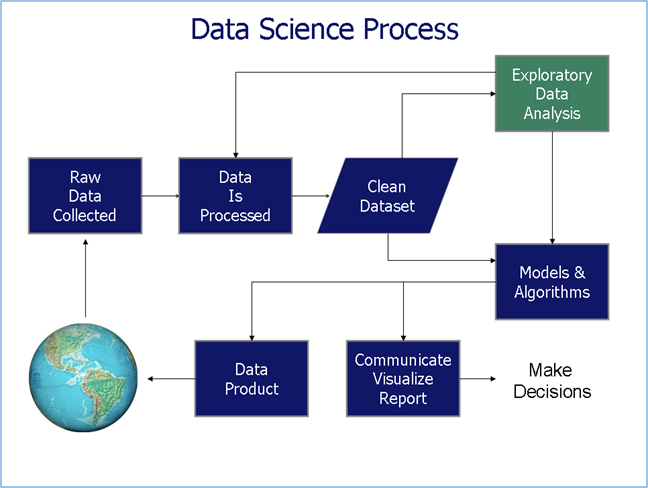
i. **Discrete variable** ii. Continuous variable

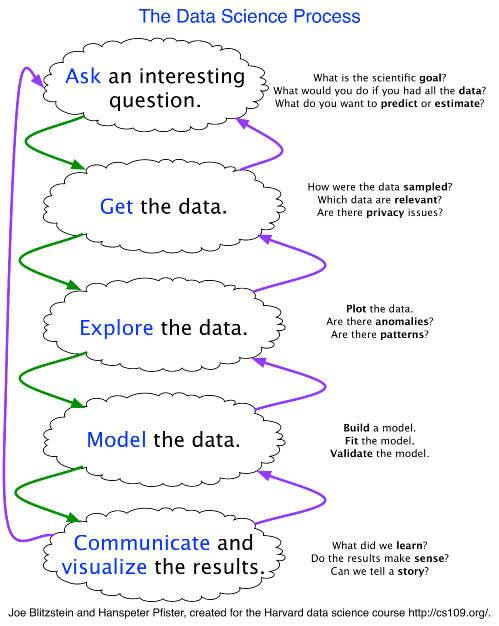
iii. Dependent variable iv. Qualitative variable

[10 marks]

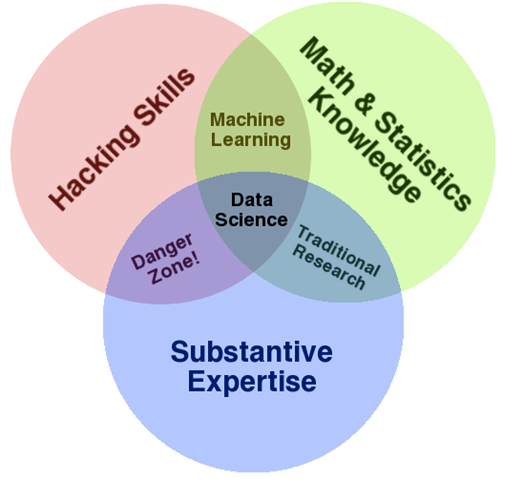
1. You’ve just been hired for your first job as a data scientist. Your boss wants you to solve a data problem. It is up to you, as the data scientist, to translate the task into a concrete problem, figure out how to solve it and present the solution back to all of your stakeholders. Illustrate **using a diagram** the “Data Science Process.”

[10 marks]





1. How to read the Data Science Venn Diagram shown below? Provide detail interpretation including each intersections.



[10 marks]

The primary skills expected for data scientist: hacking skills, math and stats knowledge, and substantive expertise.

**Hacking skills** - Being able to manipulate text files at the command-line, understanding vectorized operations, thinking algorithmically

.

Once data have been acquired and cleaned, the next step is to actually extract insight from it. In order to do this, DS need to apply appropriate **math and statistics** methods, which requires at least a baseline familiarity with these tools. For example, knowing what an ordinary least squares regression is and how to interpret it.

1. What makes a good data science culture? Express **THREE** principles that you think is important.

[6 marks]

1. **Data accessibility**

Data needs to be accessible:

* provide access to the real-raw data, don’t hide it behind abstraction layers and permission handling. Liberate it and let all employees see the data as is.
* table names, column names and data values should be explicit. Have some kind of a schema explorer / detection scripts to help you figure out where you need to look and which data set you need to query to get the results you are looking for.

1. **Cross domain availability**

Some of the most interesting and profitable analysis i’ve ever made came from joining multiple data sources. For example - join sales data from a CRM with behavioral data and marketing data can really increase marketing ROI

1. **Commitment and data integrity**

Now that you have built it - it’s up to you to keep it alive - show graphs on company updates. Celebrate new data sets and achieved goals. Share your decision making process with your colleagues, encourage youe emploees to go on a data quest from time to time, and above all - keep the data real.

No one will use data they can not trust - so put your money where your mouth is and allocate resources - you can hire a data engineer or an ETL developer or use a SaaS ETL tool (like Alooma, this is where I work) to maintain a trust relationship between your people and their data.

BTW - if from time to time you make decisions that can not be backed by data - share this too. It’s important for a good data science company culture to keep things transparent and mostly to avoid analysis paralysis.

1. Data can be acquired for a data science project by many different means. Summarize **THREE** ways to acquire data.

[4 marks]

1. You were asked to give a presentation on big data to the audience who have no clue on what big data is all about. Prepare **THREE** points to teach the audience.

[6 marks]

1. What are the **TWO** options dealing with missing data? Discuss the options.

[4 marks]

1. See cross-word puzzle.

**END**