2017 Fall Data Science Final Assignment

Liu, Wei RIN:661849419

1. Construct a scientific (data-oriented) workflow. Present the workflow in a diagram form with suitable annotations (documentation) for someone else to review. The workflow should be an accurate depiction of the data flow from the chosen project for your specific instance, i.e. it need not be generalized. The diagram should be included in what you submit. Provide a minimum of a 5-6-sentence description of what was required to carry out the construction of the workflow.

(Please refer to the "workflow.pdf" for the diagram)

On the first stage of the workflow, the sensor and participant are required because it was about the raw data collection.

Licence of data copy is required because the data is belongs to the experiment team.

Skill of Python and SPARQL are required in the second stage which is data normalization.

Also, familiarity of R and Tableau is required because they are the primary tools for data analysis and visualization.

Just a little mathematical knowledge are required in this workflow because the analysis is not complicated.

AIP is required for ducumentation of the final result, data frame, metadata as well as the visualization graph.

2. Describe each major stage of the workflow and indicate how well (or poorly) data and information preservation is enabled or accommodated. You may include URLs to information sources, etc., minimum of 3-5 sentences.

The first major stage is data collection. The data was collected directly from the experiment. Raw data was stored in .csv table and has not been made public yet. The second major stage is data normalization by using Python and SPARQL. The third major stage is Data Analysis using R and Tableau. But before that a certain analysis plan should be made. The fourth stage is Documentation. All the normalized data frame, metadata as well as the result of analysis were documented in an AIP.

The data and information preservation is not very good in this project. That is because the data was from experiment and had not been made public yet. Only by asking people who engaged in that experiment can others access the data.

Because of the licence of data, the result of our analysis could be shared, but the raw and normalized data in our analysis may be prevented to be made public. But it depends. After

talking with the principal of the experiment, our result could be upload to public platform such as Github if we get the permit.

3.Describe how the workflow and your assessment of existing documentation such as provenance and other contextual information enabled (or not) the data stewardship that was needed in your project.

In the first stage of the workflow, which is raw data collection, the data is collected in the RPI experiment with CASE. Therefore, the security of data is well supported which means the data is reliable and the data ownership is obviously belongs to the lab or experiment team.

The persistence of data could be long because the nature and kinds of the data (concentration of Co2, human heart rate and physiological stress factor). The data and metadata collected by the lab is well stored and probably will not be updated in the future. However, the data set is obviously under sampling due to the scale of the experiment and the missing value. Therefore, the data is not suitable for analysis which require lots of data. But the knowledge and information discovery should be fine for those analysis doesn't need large scale of data. Another problem is that the data provenance is not that good because of the licence of data.

The normalized data is well stored in several .Rdata file which could be open anytime with R related tools. That guarantees the data preservation well.