**DSCI-549: Introduction to**

**USC** Viterbi School of Engineering

**Computational Thinking and**

**Data Science**

**Term: Fall 2021**

**Project – Part 2**

**Date due: December 15th, 2021 - 5PM PT**

**Goals**

The purpose of the project is to familiarize yourself with the steps involved in most data science projects, at least in the conceptualization phase.

**Overview**

Choose a data science project of interest and answer the questions in the detailed assignment section below. Answer each question individually. **DO NOT** submit an essay or report-like assignment (you will be marked down).

Remember that **ALL ANSWERS MUST BE JUSTIFIED and MUST APPLY TO YOUR DATASET.** e.g., when describing the algorithm, ensure that you explain which columns/image… in the dataset the analysis will be applied to. Even though you are not writing code, you are expected to provide a close recipe.

You are allowed to work in groups. **HOWEVER,** this is an individual assignment, and your data science question/dataset/analysis must be unique. See the Course Overview slides for examples.

This is the second part of the assignment. You are allowed to change you question/dataset from what you have done in part 1; in which case you will have to redo this assignment for context. Note that it will not be regraded.

**Assignment**

Part 1: Parallel processing (25 points)

1. Can any steps be run in parallel? If so, which ones? Explain why you can/cannot do so.

Part 2: Provenance (75 points)

1. Enter the PROV ontology in Protégé. (30 points)
2. Create instances of entities, activities, use and generation, and agents for the analysis you are proposing to conduct. (35 points)
3. Upload the .owl file with your homework.

Part 3: Data stewardship (75 points)

1. Describe **in details** (include relevant software/website as we have done in class) the steps you would take to document and share
   1. The data (25 points)
   2. The software (25 points)
   3. The workflow (25 points)

Part 4: Communicating data science (75 points)

1. Record a five minute presentation highlighting your problem, its relevance, and anticipated results. Your presentation should follow best practice as discussed in class. Grading will follow two criteria:
   1. Execution (organization, speech, timing) (30 points)
   2. Content (Context/importance of the study; hypothesis; method; anticipated results). (35 points)

***IMPORTANT NOTES***

*Plagiarism – presenting someone else’s ideas as your own, either verbatim or recast in your own words – is a serious academic offense with serious consequences.  Please familiarize yourself with the discussion of plagiarism in SCampus in Section 11, Behavior Violating University Standards* [*https://scampus.usc.edu/1100-behavior-violating-university-standards-and-appropriate-sanctions*](https://scampus.usc.edu/1100-behavior-violating-university-standards-and-appropriate-sanctions/)*.  Other forms of academic dishonesty are equally unacceptable.  See additional information in SCampus and university policies on scientific misconduct,* [*http://policy.usc.edu/scientific-misconduct*](http://policy.usc.edu/scientific-misconduct/)*.*

*A number of USC’s schools provide support for students who need help with scholarly writing.  Check with your advisor or program staff to find out more.  Students whose primary language is not English should check with the American Language Institute* [*http://dornsife.usc.edu/ali*](http://dornsife.usc.edu/ali)*, which sponsors courses and workshops specifically for international graduate students.*

*For more information, see the class syllabus and the USC web site.*