Assignment 1: Data Science 2025 ITWS/CSCI/ERTH 4350/6350 (10% of overall score)

Due: September 24th, 2025 (by 10:00 pm ET)

Submission method: email eleisa2@rpi.edu / LMS

Document naming: DataScience_2025_Assignment1_YOUR_NAME.ext (e.g. txt, pdf, doc, zip).

Late submission policy: first time – no penalty, otherwise 20% of score deducted each late day.

Note: Your report for this assignment should be the result of your own individual work. Take care to avoid plagiarism ("copying"), including all web resources, texts, and class presentations. You may discuss the problems with other students but do not share your written solutions.

Use the numbering below when completing your responses to this assignment.

General assignment: propose **two different data collection exercises (label them A and B)** and perform a survey of data formats, metadata and application support for data management suitable for the data you will collect in the future as a part of the Assignment 2. This is the planning process of data collection. You must plan properly, and in details before you collect. This is a data collection planning exercise.

You MUST complete the following questions for the BOTH collections (both collection-A and collection-B)

Note the overall modes of data collection:

- Observation
- Measurement
- Generation

Driven by

- Questions
- Research idea
- Exploration

Include all citations and sources of information you use especially for Q 2, 3 and 4.

1. Data collection – propose two data collection options - 4%

a. State the details of the mode of each collection and what the data collection need is being driven by. Suggested minimum response is 6-7 sentences. Include any drawings/diagrams/sketches as needed (these are not mandatory, feel free to include them if you create them).

b. Describe a management plan for the data and metadata acquisition, and initial curation using the 9 headings under Data Management in the lecture slides from week 2. Minimum 4-6 sentences per category.

2. Survey of data storage/ formats - 3%

a. Based on Q1 (for both proposed collections), research and describe existing suitable data formats that could be used. Provide references (URLs) for uncommon data formats (e.g. FASTA for bioinformatics, KML for geospatial). If no suitable format matches your data needs exactly, consider how you may restructure the data to fit existing formats and describe these choices. Minimum 5-6 sentences.

3. Survey of metadata conventions, standards - 3% (4000-level), 2% (6000-level)

a. Based on Q1 (for both proposed collections), research and describe existing metadata conventions/ standards that could be used. If none are suitable, describe the metadata needs and how you may record these metadata without a standard. Minimum 4-5 sentences.

4. 6000-Level course work student question - 1%

a. Describe the provenance information you plan to collect (for both proposed collections) and how that may support the overall data collection and goals of the investigations. Minimum 4-5 sentences.