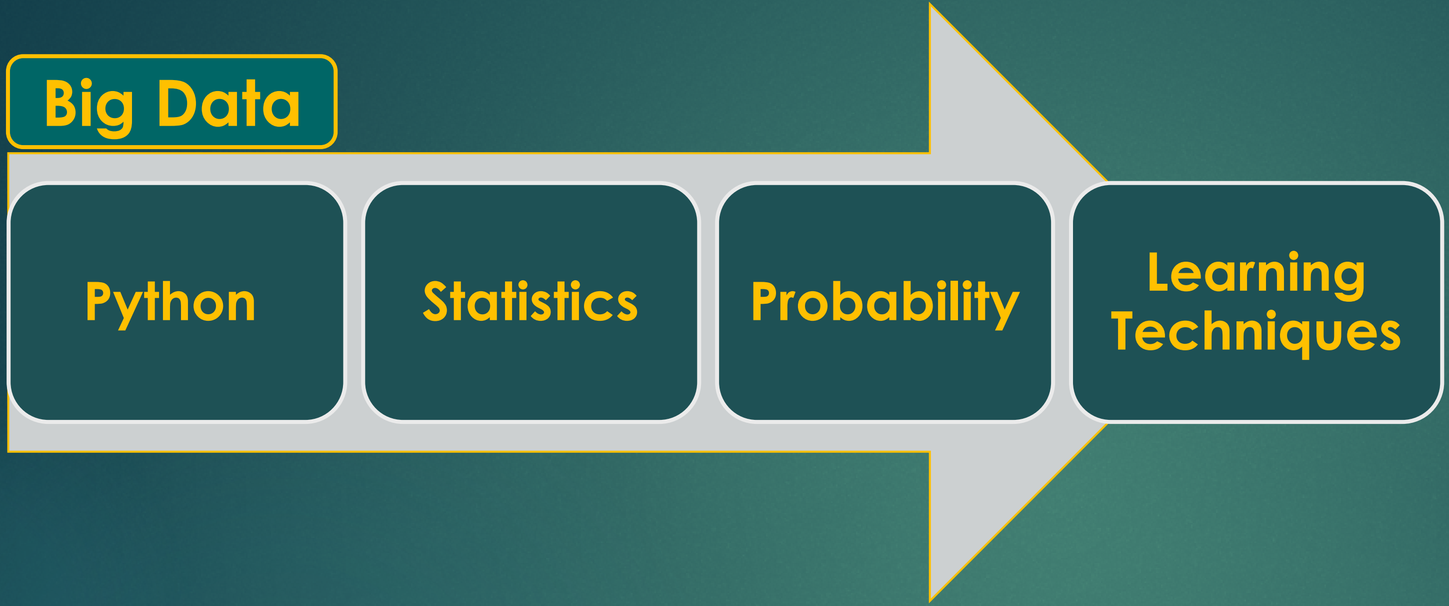
IS 669 Big Data and Information Systems
Fall 2023



Big Data

Python

Statistics

Probability

Learning Techniques

Group Project Briefing

TASK SET 2

Professor HG Locklear
hlocklear@pace.edu

Phase 1

- In Phase 2 of your project, your team will develop a persistent storage of your data.
- This persistent storage will be done using the MySQL RDMS.
- This phase serves several important purposes in the data analysis process:
  - Allows the data set and samples from the data set to be stored so that they can be easily recalled.
  - Allows separation of data into Training, Testing, etc. sets.
  - Allow easy creation of samples that contain only the specified attributes.
  - Increases your understanding of how to manipulate data in a RDMS like MySQL.

Phase 2 Submission and Evaluation

- The Phase 2 portion of your project and the submission of your single Jupyter notebook is due on Friday Nov 24th by 1 PM.
- The submission will be done on BrightSpace by the Group Leader only…no other submissions will be accepted.
- The portion of the project is worth a total of 10 points toward your overall 40 points for the project.
- There is NO extension for this portion and late submissions receive 0 points.
- Your notebook will be evaluated on its correctness and organization…be sure to follow the specifications provided exactly as they are described.
- Code that does not execute, is in the incorrect cell, or provides incorrect output receives 0 points.

Tasks for Phase 2

- In Phase 2 of your project, your team should create a single highly-organized Jupyter notebook which accomplishes the persistent storage task listed.
- Organize your Jupyter notebook as shown below.

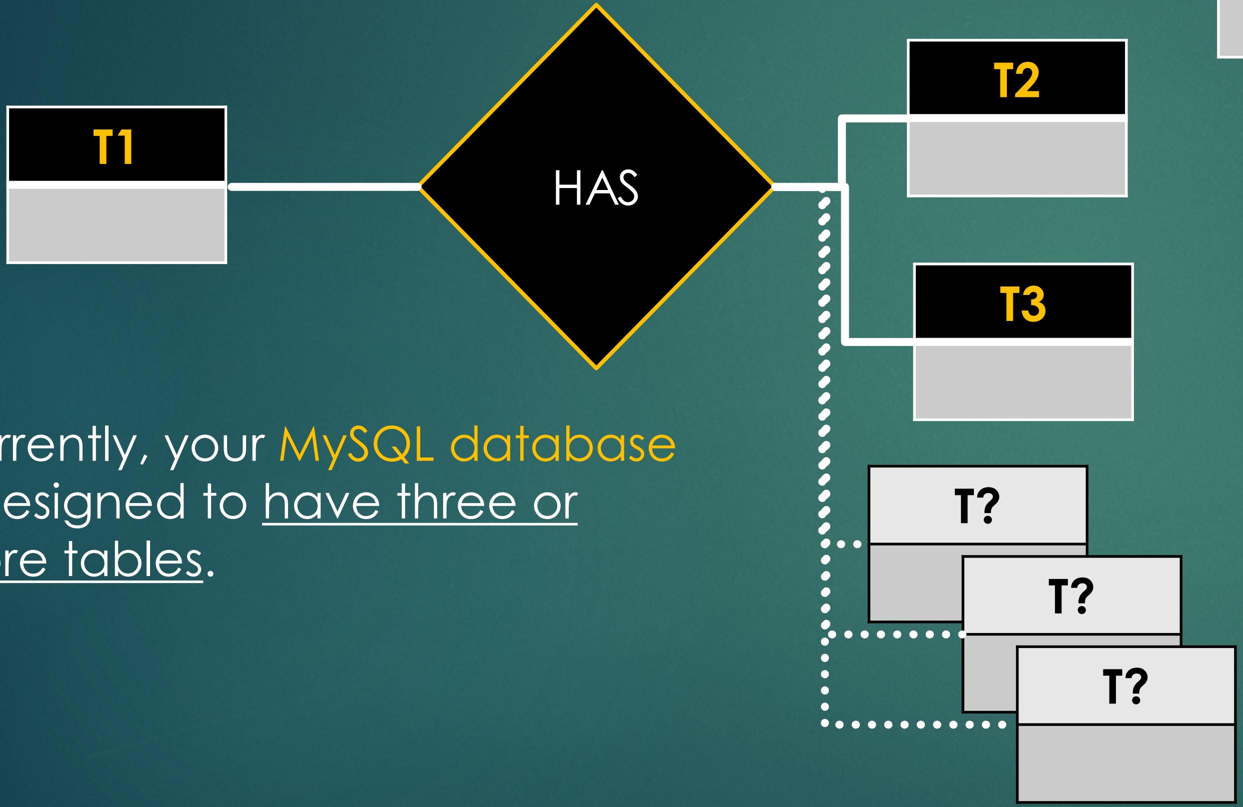
Cell	Content
1	Project # and all member of your group
2	All import statements
3	Database Schema Diagram
4	Database Connection Code
5 - 5n	Database Tables Creation Code **Part 1**
6 - 6n	Table Population Code **Part 2**

All database creation is done through use of the mysqlclient. Population of individual tables will be done using a Python-MySQL connection contained within your function.

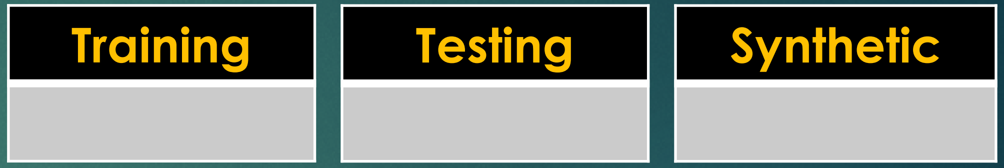
The execution of this notebook (when I select Run All Cells) should create and populate your database.

Tasks for Phase 2 Part 1

- Your Database Schema should be organized as shown.



Currently, your MySQL database is designed to have three or more tables.



In addition to the tables, you already have you will create two additional tables which include all the attributes of your data records

Each of those tables will contain half of the total number of records in your original data set.

Finally, you will create another table which contains each data record that you generated synthetically.

Tasks for Phase 2 Part 2

- Create a single function for each of the tables in your database that reads your original csv file and stores the appropriate data in that table. These functions must use the Python-MySQL connection to populate the tables.
- Call each of these functions (in their own cell) to populate your database.