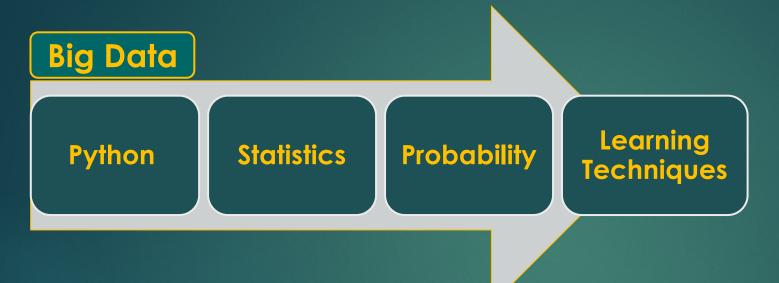
## IS 669 Big Data and Information Systems Fall 2023



# Group Project Briefing

**FINAL TASK SET** 

Professor HG Locklear hlocklear@pace.edu

#### Phase 3

- In Phase 3 of your project, your team will develop a classification systems for of your data.
- ▶ This classification system will involve the creation of a classifier that uses:
  - ▶ Probabilistic Techniques
  - Distance Metrics
  - Similarity Metric
- This phase serves several important purposes in the data analysis process:
  - Gives insight into the variability of the data.
  - Provides a predictive system to analyze new data.
  - Allows for data set feature reduction which reduces complications in data acquisition.

#### Phase 3 Submission and Evaluation

- ► The Phase 3 portion of your project and the submission of your <u>single Jupyter notebook</u> is due on Tuesday Dec 5th by 1 PM.
- ► The submission will be done on <u>BrightSpace by the Group Leader only</u>...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 <u>evaluated on its correctness and organization</u>...be sure to follow the specifications provided <u>exactly</u> as they are described.
- Code that does not execute, is in the incorrect cell, or provides incorrect output receives 0 points.

#### Tasks for Phase 3

- In Phase 3 of your project, your team should create a single highly-organized Jupyter notebook which accomplishes the classification task listed.
- Organize your Jupyter notebook as <u>shown below</u>.

Cell	Content
1	Project # and all member of your group
2	All import statements
3	Probabilistic Classifiier Code
4	Probabilistic Classification Results
5	Euclidean Distance Classifier
6	Euclidean Distance Classification Results
7	Cosine Similarity Classifier
8	Cosine Similarity Classification Results

### Classification Tasks

- Create a probabilistic classification system based on the distribution which you feel most represents your data set.
- Create a distance classification system using Euclidean distance.
- Create a similarity classification system using Cosine similarity.
- ▶ Design your classification such that there are only two classes (binary classifier)...you may have to reconsider the labelling of your data points to accomplish this. Just provide a brief explanation of your reasoning for the two data labels.
- Provide an explanation of your training and testing data set construction. You may have different techniques for different classification methods.
- ▶ Provide a confusion matrix that <u>visually demonstrates</u> the results for each of your classifiers.
- Provide a narrative that explains your results and why you think they came out that way for each classifier.