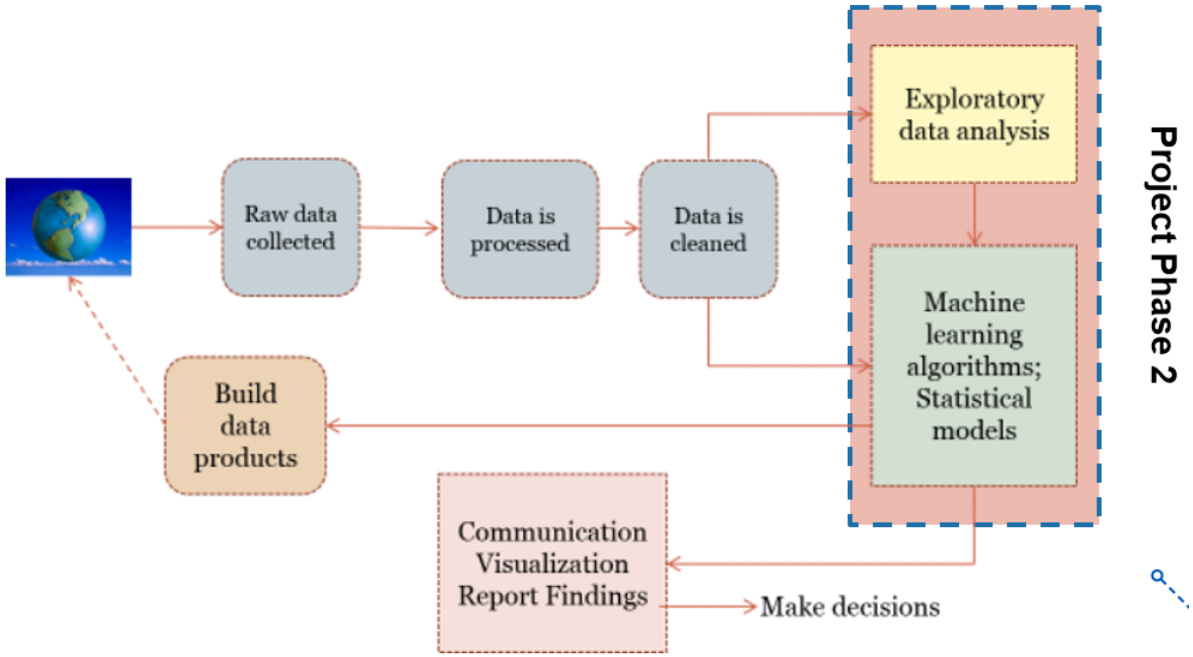
Project 1 Phase 2 – EDA – Data Modeling

Due: 10th Nov



Learning Outcomes for Phase 2:

1. Understand the basic characteristics of the data by performing John Tukey's exploratory data analysis (EDA) [2].
2. Identify suitable ML, MR, and/or statistical modeling algorithms. Model and apply algorithms to get insights into the behavior of the data. It could be classification, regression, clustering, etc.
3. Understand and explain the differences in each of the algorithms used.
4. Visualize the analytics using appropriate charts and graphs. You can use Seaborn or any other plotting tool.

Deliverables:

1. **Exploratory Data Analysis (EDA) [10 points]:**

- Perform exploratory data analysis as defined in the NIST publication [2] and as originally described by John Tukey [3,4].

- Record the outcomes and what you learned and how you will use this information. For example, in choosing features (columns) and dropping columns, and in short feature engineering.
- You need to demonstrate 5 different, significant and relevant EDA operations and describe how you used these to process the data sets further to provision them for downstream modeling and analytics. Figures and tables should be included where relevant.

2. Algorithms/Visualizations [25 points]:

- Apply 3 different significant and relevant algorithms (ML, MR, and/or statistical models) to your data and create visualizations for the results.
- At least 1 of the 3 algorithms must be one that was not discussed in class.
- These outside algorithms can come from the class textbooks, or other sources. Cite the appropriate sources for each outside algorithm you choose to apply.

- https://scikit-learn.org/stable/
- https://plotly.com/python/

3. Explanation and Analysis [15 points]:

- For each of the 3 above algorithms, provide justification for why you chose the particular algorithm, and discuss the effectiveness of the algorithm when applied to your data to answer questions related to your problem statement.
- This should include discussion of any relevant metrics for demonstrating model effectiveness, as well as any intelligence you were able to gain from application of the algorithm to your data

References:

1. C. O'Neill and R. Schutt. Doing Data Science., O'Reilly. 2013.
2. NIST on EDA, https://www.itl.nist.gov/div898/handbook/eda/section1/eda11.htm
3. John Tukey Biography, https://mathshistory.st-andrews.ac.uk/Biographies/Tukey/
4. J. Tukey, http://theta.edu.pl/wp-content/uploads/2012/10/exploratorydataanalysis_tukey.pdf