BIN2023R01 - **INTRODUCTION TO DATAMINING & MACHINE LEARNING FOR BIOINFORMATICS**

Lab Exercise 1 - Exploratory summary of datasets and graphical techniques for visualizing datasets

Aim: To explore various data visualization techniques/methods by implementing via Python code.

Using the data below, write Python codes for
Data = 15, 20, 30,70, 80, 50, 20, 15, 10, 10, 15, 8, 12, 18, 14, 20, 22, 25, 16, 10, 14, 18, 22, 26, 30

1. **Calculating Mean, Median, Mode, Variance, Standard Deviation**
2. **Plot a histogram**
3. **Plot boxplot**
4. **Plot violin plot**
5. **Plot 3D bar plot**

biological_data_A = {'Gene1', 'Gene2', 'Gene3', 'Gene4', 'Gene5'}
biological_data_B = {'Gene3', 'Gene4', 'Gene5', 'Gene6', 'Gene7'}

6. **Create a Venn diagram**
7. **Create an upset plot**

# Assuming you have different species and genes flowing between them
species = ["Species A", "Species B", "Species C"]
genes = ["Gene A1", "Gene A2", "Gene B1", "Gene B2", "Gene C1", "Gene C2"]
# Assuming genes flow between species with corresponding values
Gene B1 flows to Species A
Gene B1 flows to Species B
Gene A2 flows to Species C
Gene A1 flows to Species B
Gene A2 flows to Species A
Gene A1 flows to Species C
Gene B2 flows to Species A
Gene B2 flows to Species C
Gene C2 flows to Species A
Gene C2 flows to Species B
Gene C1 flows to Species C
Gene C2 flows to Species B

8. **Create a sankey plot**

T-SNE plot - https://distill.pub/2016/misread-tsne/

Questions:

1. In the context of omics data (like proteomics or metabolomics), what insights can box plots provide, and how might they be limited in representing the underlying complexity of the data?
2. Time-series data is common in biological studies, such as monitoring gene expression over time under different conditions. What are the key considerations and preferred methods for visually representing this type of dynamic data to capture both temporal patterns and biological relevance effectively?
3. How do violin plots provide more detailed insights about the distribution of data compared to boxplots, particularly in the context of large biological datasets?
4. In what types of biological data analysis are bar graphs most effectively used, and when might another form of visualization be more appropriate?
5. What are the best practices for interpreting outliers in a boxplot, and how can they be indicative of significant biological phenomena?

Soft copy deadline: Jan 29th 11:59PM
Hard copy deadline: Jan 30th 3:15PM