# Data and Data Preprocessing

**Problem 1: Types of Attributes (14 points)**

Classify the following attributes as nominal, ordinal, interval, ratio. **Explain why.**

1. Rating of an Amazon product by a person on a scale of 1 to 5

**Answer: Ordinal.** There is a natural order. i.e. (5 > 4 > 3 > 2 > 1).

1. The Internet Speed

**Answer: Ratio.** It is measured on a continuous scale with a true zero. 100Mbps is twice of 50Mbps so the ratios are meaningful.

1. Number of customers in a store.

**Answer: Ratio.** It is a count variable where zero means the absence of customers. Doubling the number of customers is meaningful.

1. UCF Student ID

**Answer: Nominal.** Just an identifier label with no inherent order or arithmetic use.

1. Distance

**Answer: Ratio.** It is continuous and has an absolute zero and ratios. For example, 2km is twice of 1km.

1. Letter grade (A, B, C, D)

**Answer: Ordinal.** Order from the best to worst.

1. The temperature at Orlando

**Answer: Interval.** Values are ordered with equal spacing, but zero is arbitrary. If measured in kelvin it would be ratio because it has a true zero.

# Problem 2: Exploring Data Preprocessing Techniques (26 points)

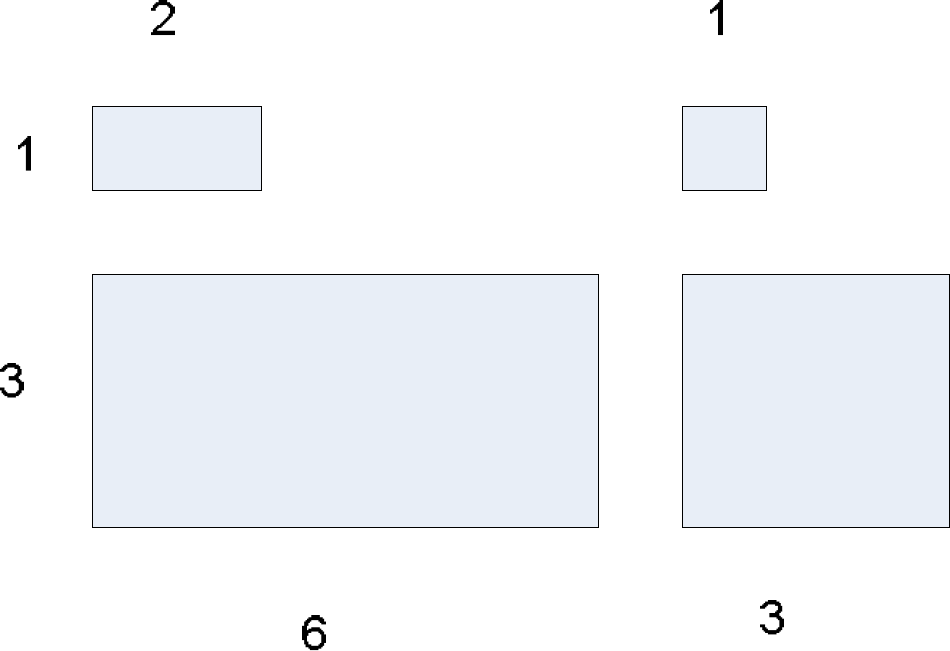
Read the solution post of the Kaggle Titanic Dataset: https://[www.kaggle.com/code/preejababu/titanic-data-science-solutions.](http://www.kaggle.com/code/preejababu/titanic-data-science-solutions) Run the code and reproduce the data preprocessing and classification modeling steps.

Q1 (Reproduce): Please read, understand, run the code and reproduce the model accuracies. Please briefly explain whether you can reproduce the classification accuracies of 'Support Vector Machines', 'KNN', 'Logistic Regression', 'Random Forest', 'Naive Bayes', 'Perceptron', 'Stochastic Gradient Decent', 'Linear SVC', 'Decision Tree'. (10 points)

Q2 (Improve): Is the data preprocessing process proposed in the Kaggle post the best preprocessing solution? If yes, please explain why. If not, can you leverage what you learned in the class and your previous experiences to improve data processing, to obtain better accuracies for all these classification models? Describe what is your improved data preprocessing, and what are your improved accuracies? (16 points)

# Problem 3: Distance/Similarity Measures (10 points)

Given the four boxes shown in the following figure, answer the following questions. In the diagram, numbers indicate the lengths and widths and you can consider each box to be a vector of two real numbers, length and width. For example, the top left box would be (2,1), while the bottom right box would be (3,3). Restrict your choices of similarity/distance measure to Euclidean distance and correlation. **Please explain your choice.**

****

Which proximity measure would you use to group the boxes based on their shapes (length-width ratio)?

Which proximity measure would you use to group the boxes based on their size?

**Please submit a PDF report. In your report, please answer each question with your explanations, plots, results in brief. DO NOT paste your code or snapshot into the PDF. At the end of your PDF, please include a website address (e.g., Github, Dropbox, OneDrive, GoogleDrive) that can allow the TA to read your code if any.**