# DAT16 SF: HOMEWORK 6 "MIDTERM"

**Assigned:** Wednesday, September 2nd, 2015
**Due:** Wednesday, September 9th before class
**Submission Method:** Submit via Slack directly to Lead Instructor. Don't share your results with general channel. Work independently on this midterm assignment.

This homework is a formally assessed "midterm" assignment. It will be evaluated on a 0-100 point scale.

The purpose of this homework is to gain experience implementing logistic regression, as well as to pull together much of what we have learned to date. This includes data wrangling, imputation, model evaluation (e.g. cross validation, ROC curves), visualization (e.g. matplotlib), and thoughtful interpretation of results.

## DATA & CONTEXT

In this assignment, we will use the passenger list of the Titanic, as provided in a well-known Kaggle competition.

The dataset is a list of passengers. The second column of the dataset is a "label" for each person indicating whether that person survived (1) or did not survive (0). Here is the Kaggle page with more information on the dataset:

http://www.kaggle.com/c/titanic-gettingStarted/data

For your convenience the dataset is copied to the midterm folder.

## DELIVERABLES

You are expected to deliver 2 files: one with the code and one with a written PDF report.

## CODE

You may work in iPython notebook, or via Python scripting, or both. Note that iPython notebook provides the ability to export Python scripts. Filenames for Python scripts should end in ".py". To execute a Python script from the command line, simply type:

```
python <filename.py>
```

Please make sure to specify if any additional library is necessary to run your code. Please submit your work via Slack.

## REPORT

In addition to your code, please submit a 1-3 page report discussing your approach, results, conclusions, and next steps. Include visualizations. Report must be in PDF format. It can be either in the form of a written document or in the form of a slides presentation.

## MIDTERM QUESTIONS

Please address the following specific questions:

1. Several passengers are missing data points for age. Impute the missing values so that there are no "NaN" values for age as inputs to your model. Explain what value or values you used for the missing age data and why you used those values. (Hint: Think statistically here. Explain your thinking and choices.)

2. Create and run a logistic regression on the Titanic data to predict the survival of passengers.
    a. Make sure your model results are reproducible by the instructor team, given any random aspects to your code / steps.
    b. Show your model output. Include coefficient values.
    c. Explain which features are predictive for this logistic regression and, intuitively, why you think this may be so. Describe your thinking / logic in words. It is not sufficient to cite output statistics.
    d. Make sure that you can push new data points through your model to get regression output. We will be providing a test set for this purpose.

3. Implement cross-validation for your logistic regression. Pick the number of folds. Explain your choice.

4. Create an ROC curve for your logistic regression by varying the threshold value for survival.
    a. Plot this ROC curve visually.
    b. What is the AUC for your model?
    c. Explain why the model has achieved this level of accuracy/precision.
    d. Explain how you could improve these metrics, as potential next steps.
    e. Armed with this knowledge, what threshold value would you use? Why? Describe your thinking.