# ST563: Introduction to Statistical Learning - Final Project

Wenbin Lu

**Background:** We may guess the background of a newly met friend by their actions and talking to find a comfortable topic that won't offend either of you. Similar things can apply to songs as well: the release year of most songs may have some time-specific "signatures" that are jointly determined by rhythm, dynamics, etc.

**Project Goal:** A data set[1] is provided to demonstrate what we have learned in this course. We have 100K data in total and it is already separated [2] into train/test split so performance among different groups is comparable. Each row has 92 features, and the first 90 features (covariates) are 12 timbre averages followed by their covariance. The remaining two of them are the release year and a corresponding class(our responses) for you to consider both the Regression problem (consider the trend behind the release year as "continuous") and the Classification problem (predict from one class among three) in this project.

**Components:** You are recommended to think about, but not limited by, the following components for this project:

- Data Pre-processing: Did you normalize, scale, or standardize the input variables? Did you consider treating the outliers?
- Model Building: What kind of models did you implement for each target? How did you select hyper-parameters? What are the selected performance metrics?
- Final Result: What is your model performance on the test set? Does your model have good interoperability?

**General Instruction:** This project is group work so every group member should participate, and we will collect evaluation forms for participation from every member. The write-up should be written in manuscript form with complete sentences and paragraphs and carefully labeled figures and tables. Do not submit a markdown document. Include sufficient detail of your model and results so that another group in the class could reproduce your results. It should be a single PDF document. The main text is no more than 10 pages (double-spaced) including tables and figures but excluding code, which should be given as an appendix that does not count towards the 10-page limit. **Have Fun!**

[1] Sampled from https://archive.ics.uci.edu/ml/datasets/YearPredictionMSD

[2] In a way to avoid the "producer effect" by making sure no song from a given artist ends up in both the train and test set.