Overview  This course aims to provide an introduction to some of the fundamental methodologies in classical multivariate analysis and modern statistical learning for upper level undergraduates in statistics and data science. A particular focus is on applying these methods to analyze real data. Through the process, you will be able to understand and speak the basic language of statistics and appreciate both strengths and limitations of each method and formulate conclusions accordingly. Finally, and most importantly, you will learn how to use open source software R for an effective data analysis.

Required background  All students should have (at least) taken 180A or 183 or equivalent courses in basic probability and statistics. For example, the following concepts will be frequently encountered throughout this course: random variable, mean, variance, probability density function, cumulative distribution function, independence, covariance, correlation, response variable and covariate, etc. The default programming language for this course is R (https://www.r-project.org/).

Structure  This course tries to cover the following topics:

- Visualize and summarize multivariate data sets
- Inference on multivariate means
- Multivariate analysis of variance
- Linear regression: simple and multiple regression
- Classification: linear discriminate analysis, K-nearest neighbors, support vector machines
- Tree-based methods: decision trees, random forest, boosting
- Unsupervised learning: principal component analysis, clustering methods

Course book  This is not a traditional course that follows textbooks chapter by chapter. The lecture notes will be self-contained, and serve as the primary course materials. Also, the R codes that replicate the numerical experiments in the lecture notes will be provided on a weekly basis. Having said that, we do have the following books listed as references:

- An Introduction to Statistical Learning with Applications in R (by G. James, D. Witten, T. Hastie & R. Tibshirani) [Link]

Students do not need to purchase the above references. Keeping an electronic version will suffice.
Grading  Homework (80%) and Final Project (20%)

Homework  Homework will be assigned weekly, and must be turned in (via Gradescope) by the end of every Friday, starting Week 2. Students will be asked to form study groups, each consists of 3 members (or less). Details of the rules will be given during the lecture, Week 1. For each homework, every group should submit one report, generated from R Markdown. In the report, the concrete contributions from each member should be highlighted. If not doing so, points will be deducted. Also, every group has the right (for example, by voting within group) to kick out a member whose contribution is constantly lacking. Late homework will not be accepted unless there is a reasonable justification.

Academic integrity  Collaboration and discussions are allowed and are encouraged in this class, but copying or letting others copy your work amounts to plagiarism. Although I expect high academic awareness in this class, if such plagiarism occurs I will take the following action: a grade 0 will be assigned to all involved in the incident where cheating occurred and I will report the incident to the Academic Senate which will then decide appropriate course of action. Integrity matters, and certainly is NOT A JOKE!