# Syllabus for ME397M: Applied Engineering Data Analysis, Optimization, and Visualization (or Too Big to Excel)

**Spring 2018**

**Joshua D. Rhodes, PhD**

**Unique Number:** # 17874/18104
**Semester**: Spring 2018
**Day/Time/Location**: Monday and Wednesday / 9:30am to 10:45am / Room: RLM 5.124

**Instructor**: Joshua D. Rhodes, PhD
**Office Location:** FAC 428
**Phone**: (512) 658-2965
**Website**: http://www.webberenergygroup.com
**Email**: joshdr@utexas.edu
**Office Hours:** 2:30-4:00pm, Wednesdays in FAC 428

## Introduction

The modern engineer must also be a data scientist. The future holds more data, not less. In particular engineering datasets are becoming ever larger. The purpose of this course is to teach students how to be comfortable working with data that they might not necessary be able to see in its raw form. Leveraging the database and supercomputing resources of the Texas Advanced Computing Center, the student who completes this course should be able to gather, clean, merge, store, analyze, and visualize various types and sizes of datasets.

Students are highly encouraged to bring their own research projects to use as deliverables for the class. Class projects will be designed to be broad enough to accommodate most engineering research projects.

## Class Outline

1. Data structures
2. What to do when datasets are too large for excel
3. Using TACC's supercomputing resources
4. How to move data to TACC and set up a database
5. Using PostgreSQL to manipulate data in databases
6. Github
7. Data analysis: machine learning
    a. OLS and logit regression
    b. Neural networks
    c. Clustering
    d. Optimization
8. Running analyses using parallel processing

9. Introduction to data visualization
10. Introduction to Geographical Information Systems

## Objectives

Upon successful completion of this course the student should be able to:

1. Obtain/scrape and interact with data files, including those from web-based (and other) APIs
2. Automate the cleaning of data and spot/deal with irregularities/missing entries
3. Format and upload into a PostgreSQL database hosted by the Texas Advanced Computing Center (TACC)
4. Query/analyze/merge/subset that database (remotely) and data subsets (locally)
5. Use remote Linux-based supercomputing facilities (TACC) to run multiple types of parallelized analyses on data from database
6. Work in groups and collaborate on code over Github
7. Use Graphical Information Software (GIS) to perform spatial analysis on data
8. Visualize data in clear, intuitive way

## Coursework

This course is project-based. There will be some suggested reading and texts, but it will be assumed that the person taking this course actually wants to learn the material. There will be a few times where an R function (code) will be submitted for checking of functionality and understanding. There will be a mid-term that will be more conceptually-based and a group project will function as the final exam. The grading breakdown will be 15% code samples, 35% midterm, 45% final, and 5% class participation. Final grades will include + and – distinctions (e.g., a B+ or B- is possible).

## Prerequisites

There are no direct prerequisites for this class, however there will be significant programming involved. Most coding examples will be given in either R or Python. Students will be expected to be self-motivated to learn how to apply the skills and code learned in the class towards their own research projects.

## Observance of University policies:

Standard University policies relating to accommodation for students with disabilities and to scholastic dishonesty will be followed in this course. Information regarding these policies may be found in the General Information Bulletin.

The University of Texas at Austin provides upon request appropriate academic adjustments for qualified students with disabilities. For more information, contact the Office of the Dean of Students at 471-6259, 471-4641 TDD or the College of Engineering Director of Students with Disabilities at 471-4321.