STAT 4060J: Computational Methods for Statistics and Data Science

Meet the teaching team

- Instructor: Ailin Zhang (ailin.zhang@sjtu.edu.cn)
- Office Hour: By Appointment (very likely available on Wed 2-4 pm)

This semester, I am teaching STAT 4060 and STAT 4510.

We will provide online lectures (Feishu meeting and recordings will be provided), but the lecture notes will be presented on board.

For those who are apply for grad school/ programs, if you need a recommendation letter from me, please check to be qualified for one the following conditions :

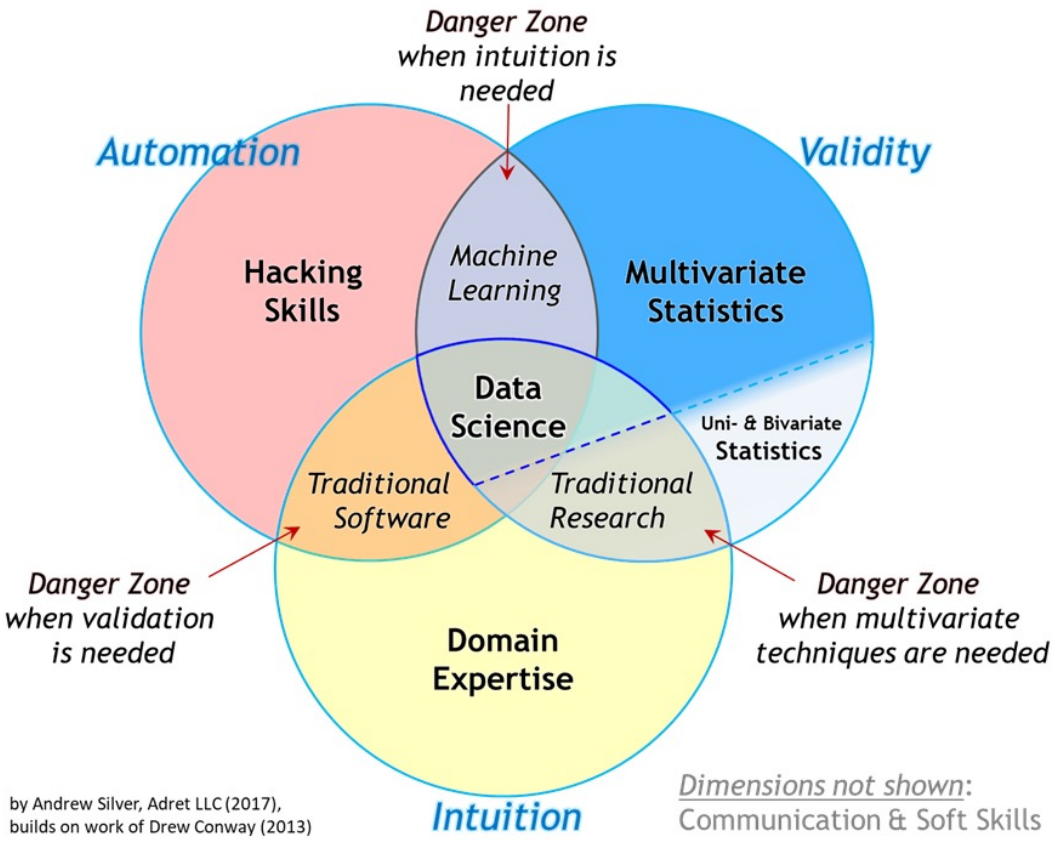
1. You have taken two courses from me: the lowest scores are A and A- (You need to have at least one A and one A- or higher).
2. You have taken one course from me with a grade of A (or higher) and you are taking another course right now.
3. Meet with me in person (If I don't know you well, I might not be able to contribute)

Meet the teaching team

- **Name**: Boyuan Zhang
- **Major/Year**: ECE Senior
- **Email**: zhangboyuan-sgr@sjtu.edu.cn
- **Study of Interest**: Big Data Technology, Data Analytics, Data Science, Statistics

- **Related Course Studied**:
  - STAT1000J
  - STAT4060J
  - STAT4130J
  - ECE4710J/STAT4710J

What is Computational Statistics?



- Bond between statistics and computer science.
- Statistical methods that are enabled by using computational methods.
- Focus on the computational side of the commonly used modern statistical and machine learning methods.
- The main theme is linear regression and its rich variations that span much of statistics and machine learning.
- Write R and Python code to implement these methods enable us to gain first-hand experiences with these methods.
- We will also learn Rcpp, R parallel, TensorFlow, etc..

Topics to be covered

- Matrix Algebra
- R Basics and R library
- Least squares regression, sweep operator, QR decomposition
- Eigen computation, Principal Component Analysis
- Logistic regression, Newton-Raphson.
- Adaboost, coordinate descent.
- Ridge regression, spline regression.
- Lasso, stagewise regression, solution path.
- Feed-forward neural network, back-propagation
- TensorFlow, Pytorch
- Deep learning (Resnet, RNN, Transformer etc..)

Course Logistics: Textbook

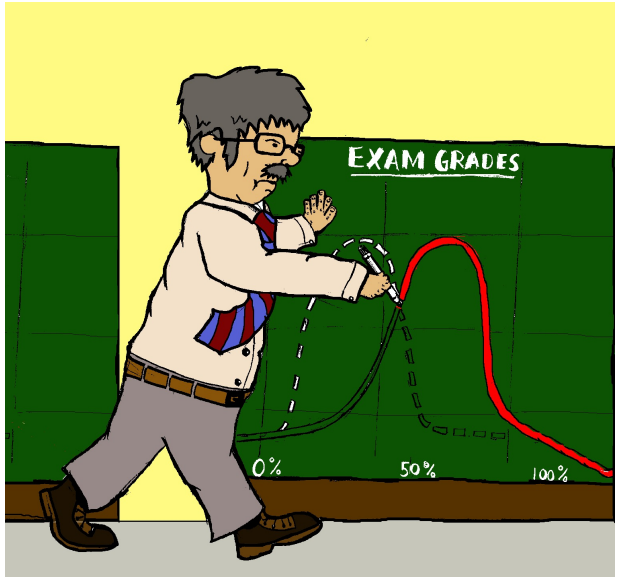
The course is **not** built on top of a single textbook. Therefore I would strongly recommend you participate in the lecture. However, the following references are useful:

- R Programming for Data Science (2020) by Roger Peng.
  - https://bookdown.org/rdpeng/rprogdatascience
- Introduction to Data Science (2020) by Rafael Irizarry.
  - https://rafalab.github.io/dsbook
- Machine Learning: a Probabilistic Perspective (2012) by Kevin Murphy.
  - https://probml.github.io/pml-book/book0.html

Grading Policy

- Guarantee >30% of A
- **35%** Homework and Participation.
  - Regular assignments will check your understanding on both theories and programming, and you will have 1-2 weeks to finish each assignment.
  - At the same time, we will have some short exercises which are closely related to our lecture, which will due by the end of each lecture day (you have less than 24 hours to submit it, no late submission will be accepted)
- **40%** Midterm
- **25%** Final Project (15% for presentation, and 10% for the written report)
- **1%*** Extra Credit

No final exam



Final Project

- Individual project
- You will be asked to implement an algorithm/ part of an algorithm from scratch
- You will write up your analysis in a written report and present your work
- Do not use existing library! (e.g. Scikit-learn)

Meet you

Please give a brief self-introduction

- Name, year, major
- Proficiency: R, Python, C++
    - It is ok to say I have no background at all
- What do you think is the most important topic in computational methods for statistics and data science?

R, RStudio, RMarkdown

When we say R", we are referring to three interrelated things:

- A language
- A community
- An implementation or environment

R is a statistical programming language

- R is specifically design to load, manipulate, and analyze tabular data (versus Python, Java, C++)
- We can use R to easily code up new algorithms, methods (versus Stata, SAS)
- We interact with R via scripts containing textual input (versus Minitab, Excel)

Key concepts:

- Store data in variables, usually vectors, matrices, and data frames.
- Manipulate data using functions, iteration, and high level declarations.
- Process data using scripts and RMarkdown documents.

R: The community

- The Comprehensive R Archive Network (CRAN) is a collection of user submitted packages.
- R is supported via: textbooks, official mailing lists, StackOverflow, R Bloggers, etc
- R is being adopted by Fortune 500 companies, government, start ups, applied academic disciplines, many others.

R: The environment

- The official R implementation consists of an command line interface for entering R commands, a batch file processor for handling scripts, and a basic graphical user interface for handling plots.
- We will be using RStudio which adds:
  - Projects to handle multiple R files, data files.
  - More complete file editor with syntax completion
  - Help system and graph tab
  - Integration with external software development tools
  - RMarkdown to PDF support
  - Desktop and server instances

RMarkdown

- RMarkdown is a plain text file that contains structured text and R snippets.
- It can be processed into a PDF or HTML file.
- Some great features:
  - Put the description and the implementation in one place.
  - Inline R code allows printing out values – no more copy and paste errors.
  - Includes a math language for writing up analytical results.

Take home task:

- Install R and RStudio
- Install Anaconda