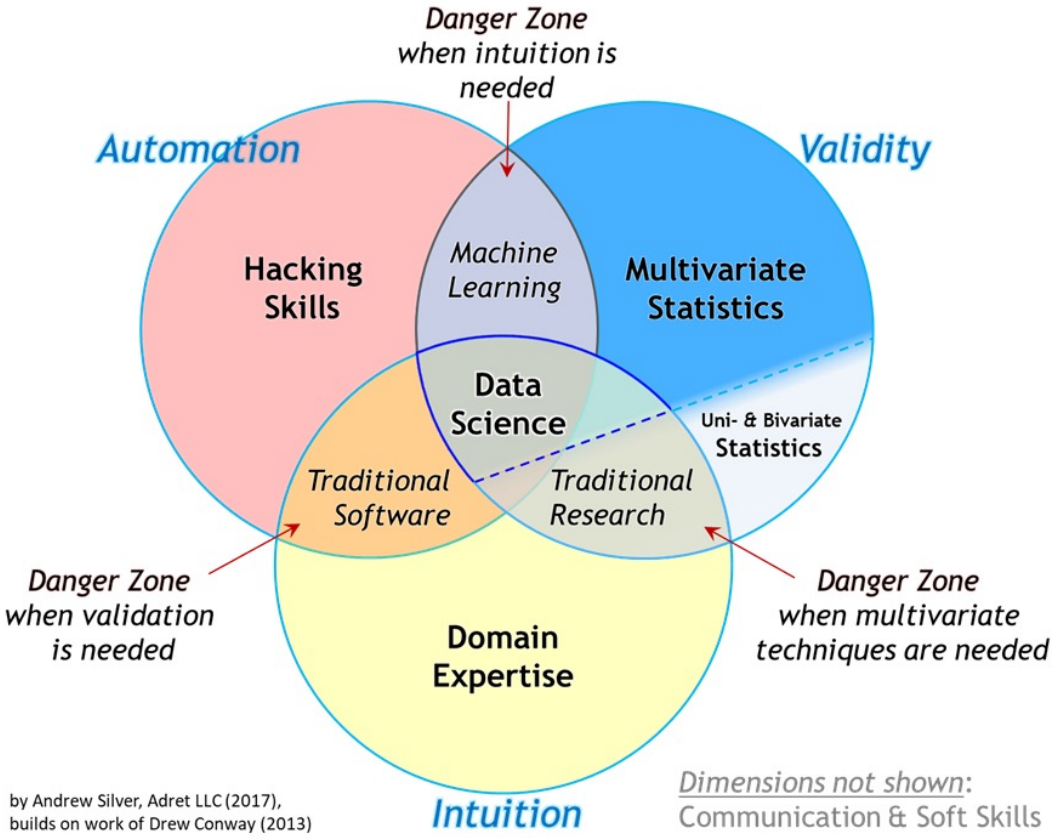
STAT 4060J: Computational Methods for Statistics and Data Science

Meet the teaching team

- Instructor: Ailin Zhang (ailin.zhang@sjtu.edu.cn)
- Office Hour: Wed 2-4 pm
- TA: Jiayuan Rao (jy_rao@sjtu.edu.cn)
  - Senior, ECE
  - Office Hour: Thur 4-6 pm
  - Current Research Direction: Computer Vision, Multi-model Learning
  - Hobby: Tennis, Music, Basketball
  - Contact me via any way is fine ;-)
- we want to help you succeed, please speak up for help!

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, SQL, Linux/unix, TensorFlow, SAS

Topics to be covered

- R and Python Basics
- Least squares regression, sweep operator, QR decomposition
- Eigen computation, Principal Component Analysis
- Logistic regression, Newton-Raphson.
- Feed-forward neural network, back-propagation
- Adaboost, coordinate descent.
- Ridge regression, spline.
- Lasso, stagewise regression, solution path.
- Factor analysis, EM.
- Random number generators, linear congruential, rejection, polar.

Course Logistics: Textbook

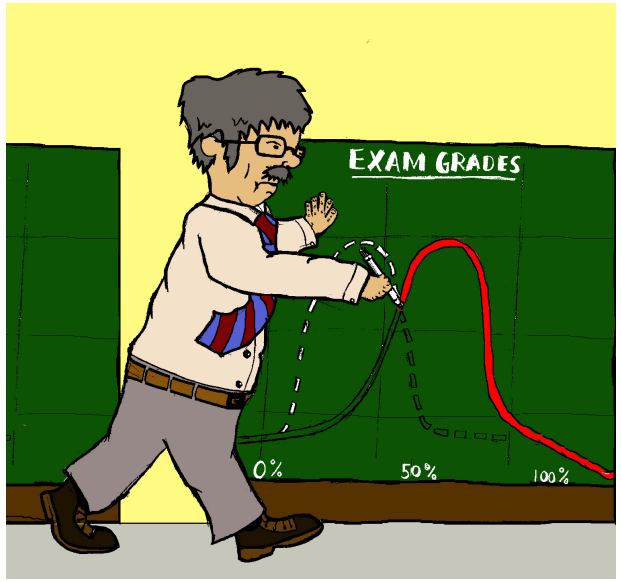
The course is **not** built on top of a single textbook. Therefore I would strongly recommend you participate in the lecture. However, the following references are useful:

- 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

Grading Policy

- Guarantee >30% of A
- **10%** Attendance and Participation.
    - To get the full credit, you need to make > 90% of the attendance check. The rest 10% is your flex-time. Feel free to use it if you have something emergent.
    - If you only make > 80% of the attendance check, you will get 6 pt.
    - Otherwise, you get 0 pt.
- **40%** Homework
- **20%** Midterm (Could be switched to quizzes)
- **30%** Final Project (15% for presentation, and 15% for the written report)
- **1%*** Extra Credit

No final exam, midterm subject to change!



Final Project

- Analyze some data using the methods we have talked about in class.
- You will write up your analysis in a written report and present your work
- Find your own dataset
  - something of genuine interest to you and where you have some knowledge about the topic.
  - Regression style data, i.e. a response variable, and for each observation, a bunch explanatory variables.
  - Predict the response variable
  - Do not use existing library! (e.g. Scikit-learn)

Meet you

Please turn on your camera and 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