Course Proposal

STA141A Fundamentals of Statistical Computing, Visualization, Exploratory Data Analysis and Simulation

Units: 4

Format: 3 hours of lecture per week.

Discussion: 1 hour of TA-led discussion.

There will be significant contact and instruction via an online forum (e.g., Piazza).

Prerequisites: STA10 or STA13 or STA32 or STA100

Catalog Description: Introduction to computing for data analysis and visualization, and simulation, using a high-level language (e.g., R). Computational reasoning, computationally intensive statistical methods, reading tabular and non-standard data.

Topics: Fundamental syntax, semantics, computational model and data structures of the programming language (e.g., R); essentials of data manipulation; writing & debugging functions; data visualization - concepts and functionality; exploratory data analysis; reading data - tabular and complex; statistical modeling software concepts;

Course Content:

- Introduction to R (or Python).
- Visualization: concepts (e.g., perception, composition) and tools (packages & functions).
- Exploratory Data Analysis (EDA), outliers, missing data.
- Identifying and explaining insights and analyses with visualization (telling stories with and about data).
- Fundamental Data Structures (vectors, lists, data frames, factors, matrices).
- Manipulation of data (subsetting, group-by operations, iteration, vectorization).
- Reading data - tabular and complex formats.
- Basic text processing for non-standard data.
- Regular expressions and text pattern matching.
- R's computational model (function calls, argument matching, pass-by-value, call frames).
- Formula language: modeling, lattice.
- Computational reasoning for data problems (algorithms and decomposing tasks).
- Writing and designing functions.
- Debugging tools and strategies.

- Coding style.
- Simulation: reinforcing statistical concepts from other courses via simulation, including random number generation.
- Resampling, bootstrapping, cross-validation.
- Optional: UNIX shell basics (interactive use).
- For Graduate Students: (accessible to UGs, but not required)
    - Important workflow tools: Version control (git); reproducability & dynamic documents (e.g. knitr, iPython)
    - Basics of creating R packages
    - Advanced graphics (e.g., grid, ggplot2).
    - Basics of the UNIX shell.

Statistical Methods covered may include

- k-nearest neighbors,
- classification trees,
- numerical optimization,
- mixture models
- ensembles (e.g. Random Forests).

Grading: A series of 6 - 8 computer assignments. The instructor may choose to use participation (in class, discussion section and online forum). Similarly, the instructor may include a mid-term and/or final exam. Given the practical computer-based nature of this course, an instructor may deem written exams inappropriate.

Course Goals: Students become proficient in data manipulation and exploratory data analysis, and finding and conveying features of interest. They learn to map mathematical descriptions of statistical procedures to code, decompose a problem into sub-tasks, and to create reusable functions. They develop ability to transform complex data as text into data structures amenable to analysis. They learn how and why to simulate random processes, and are introduced to statistical methods they do not see in other courses.

Illustrative Reading:

- R in a Nutshell, Adler.
- The Art of R Programming, Matloff.
- R Graphics, Murrell.
- R Graphics Cookbook, Chang.
- ggplot2: Elegant Graphics for Data Analysis, Wickham

Course Overlaps: This course overlaps significantly with the existing STA141 course which this course will replace. STA242 is a more advanced statistical computing course that covers more material. This also will be replaced with a graduate version of this STA141A course. ECS145 involves R programming. However, the focus of that course is very different, focusing on more fundamental computer science tasks and also comparing high-level scripting languages. R is used in many courses across campus. This course teaches the

fundamentals of R and in more depth that is intentionally not done in these other courses. Furthermore, the combination of topics covered in this course (computational fundamentals, exploratory data analysis and visualization, and simulation) is unique to this course.