NISS Short Course:
A Survey of Modern Data Science

Course Type: Half Day

Course Title: NISS Shortcourse: A Survey of Modern Data Science

Course Description:

Outline: Modern data science is driven by applications, and these often entail Big Data and machine learning perspectives. This short course reviews key ideas and methods in nonparametric regression (starting with cross-validation and light bootstrap asymptotics, then moving on to the additive model, the generalized additive model, and neural networks. It also covers variable selection, with the Lasso and the Median Model, and describes the $p >> n$ problem in the context of contributions by Candes and Tao, Donoho and Tanner, and Wainwright. The course next treats classification, with emphasis upon Random Forests, boosting, and ensemble strategies such as bagging, stacking and boosting.

Objectives: The course intends to convey the intuition and heuristics that underlay the evolution of data mining, machine learning, and data science from the 1990s to the present day. The target audience is MS-level practitioners who have some comfort with regression analysis.

Instructor Qualifications: David Banks is a professor at Duke University who has taught this material in a graduate course on machine learning on multiple occasions. In 2017, he taught this short course at the Kansas State University's Agricultural Statistics conference.

Relevance to Conference Goals: This short course aligns with the CSP's Theme 3: Data Science and Big Data. It will introduce people to a toolkit of methodologies, with instruction and guidance on when and why to use these tools, and what issues may arise. Attendees will learn statistical methods that should help them to advance in their analytical careers.

Software: No specific software will be taught. Most of the methods discussed have implementations in R, Matlab, and (sometimes) SAS.

Instructor: David Banks, Department of Statistical Science, Duke University.