Building Data Analysis Proofs

Roger Peng

(University of Texas at Austin)

27/05/2026, 4:15 pm

Department of Statistics, Ludwigstr. 33, Seminar Room 144
and online via Zoom (Link)
(Meeting-ID: 631 1190 7291; Password: StatsCol)

Data analyses are often constructed in an imperative manner, where commands representing actions taken on the data are issued sequentially. The publication of these commands, along with the data, is essential to the reproducibility of the analysis by others. However, simply presenting the code and the results of running the code can hide important details about the data analyst's premises, expectations, and assumptions about the data. Understanding this analysis reasoning can be critical to evaluating the quality of an analysis and for suggesting possible improvements. We argue that a formal representation of a data analysis that externalizes its logical construction offers more useful information for statically illustrating an analyst's reasoning. Such a formal representation would allow for the evaluation of some aspects of a data analysis without the need for the data, the visualization of the logical connections leading to a conclusion, and the ability to assess the sensitivity of an analyst's assumptions to unexpected features in the data. In this talk I will describe an implementation of this formal representation and how it might be applied to some common data analysis tasks.

About the Speaker:
Roger D. Peng is a Professor of Statistics and Data Sciences at the University of Texas at Austin. Previously, he was Professor of Biostatistics at the Johns Hopkins Bloomberg School of Public Health and the Co-Director of the Johns Hopkins Data Science Lab. He is the author of the popular book R Programming for Data Science and 10 other books on data science and statistics. Roger is a Fellow of the American Statistical Association and is the recipient of the Mortimer Spiegelman Award from the American Public Health Association, which honors a statistician who has made outstanding contributions to public health. He received a PhD in Statistics from the University of California, Los Angeles. His current research focuses on building analytic design theory for improving the quality of data analyses and on the development of statistical methods for addressing environmental health problems.