Statistical Computing and Robust Statistical Methods for 'Big' Data

Karen Kafadar

Commonwealth Professor
Department of Statistics, The University of Virginia

Today's massive datasets require statistical methods for efficient computing and informative displays even more now than when the terms "Statistical Computing" and "Statistical Graphics" evolved as disciplines 50 years ago. Because the central goals of data analysis are insight and inference, and because rarely should all data be displayed, statistical methods continue to inform our algorithms, analyses, and displays. Further, `big data' invariably contain exotic values, outliers, or mixtures of distributions, and hence require robust techniques. Finally, more data may have more information, especially when they are not representative of their target populations. This talk will emphasize the role of statistics in uncovering sampling biases, classification, robust estimation of model parameters, and graphical displays, as well as offer some thoughts on when statistical models can be usefully replaced by `black-box' algorithms.

Topics:

1. Two motivating datasets (classification)
2. Are "big" data really informative? (Not always)
3. With so much (?) data, do we still need robust methods and statistical displays? (yes)
4. Why do we need efficient statistical computing?
5. Statistical thinking regarding when to use/avoid `black-box' algorithms