

VISUAL DATA MINING

Edward J. Wegman

Center for Computational Statistics, George Mason University

10th of December 1999

Massive data sets pose particular challenges for statisticians, especially when the data is simultaneously high dimensional. These lectures will open with a discussion of the computational complexity and data set size. We will outline the limits of feasibility from both a computational and visualization perspective. A second theme will be to discuss three techniques which allow for the visual analysis of large-scale, high dimensional data: 1) parallel coordinate displays, 2) d-dimensional grand tour, and 3) saturation brushing. These tools when used in concert allow a number of standard statistical tasks to be accomplished including density estimation, rapid data editing, inverse regression, tree-structured decision rules, dimension reduction, clustering and classification. We will conclude the lectures with the analysis of several real data sets illustrating how to use our methods to achieve the data analysis.