The Challenges of Heterogeneous Data

Susan Holmes, Stanford University

Finding the right distance or dissimilarity often solves difficult statistical problems. This talk will provide a survey of mining heterogeneous biological data including networks, trees, images and heteroscedastic variables using weighted dissimilarities and locally defined distances.

Carefully tailored “distances” can incorporate prior information on data structure such as hierarchical dependencies between rows of a data matrix or the graph of correlations between the column-variables. Links to differential geometry are useful in incorporating localized information for these complex data structures. Distances are central to the statistical endeavor and enable generalizations of the notions of variance decomposition, nearest neighbor classification and clustering as I will show through several applications.