Data Integration: Data-Driven Discovery from Diverse Data Sources

Genevera I. Allen, Rice University

Data integration, or the strategic analysis of multiple sources of data simultaneously, can often lead to discoveries that may be hidden in individual analyses of a single data source. In this talk, we present several new techniques for data integration of mixed, multi-view data where multiple sets of features, possibly each of a different domain, are measured for the same set of samples. This type of data is common in heathcare, biomedicine, national security, multi-senor recordings, multi-modal imaging, and online advertising, among others. In this talk, we specifically highlight how mixed graphical models and new feature selection techniques for mixed, mutli-view data allow us to explore relationships amongst features from different domains. Next, we present new frameworks for integrated principal components analysis and integrated generalized convex clustering that leverage diverse data sources to discover joint patterns amongst the samples. We apply these techniques to integrative genomic studies in cancer and neurodegenerative diseases to make scientific discoveries that would not be possible from analysis of a single data set.
Keywords: data integration; multi-view data; multi-modal data; mixed data; unsupervised learning