Data Management and Visualization in R

BSA Computing Workshop

Brian Segal

Department of Biostatistics
University of Michigan

February 12, 2016

Overview

Data Management

Typical workflow

1. Read in
2. Reshape
3. Split, apply, and combine

To do: make diagram of read in → reshape → split, apply, combine, with → visualize at each step

Outline

1. Read in
2. Manipulating datasets

File management

Medium files

Large files

Files too large to fit into memory

Note: see Kerby Shedden's site, BioConductor, etc.

Data processing steps

dplyr

data.table

For maximum speed. Very useful, e.g., if bootstrapping column means or standard deviations.