The bigmemoRy package: handling large data sets in R using RAM and shared memory

John Emerson
⟨john.emerson@yale.edu⟩

Michael Kane

Multi-gigabyte data sets challenge and frustrate R users even on well-equipped hardware. C programming provides memory efficiency and speed improvements, but is cumbersome for interactive data analysis and lacks R's flexibility and power. The new package **bigmemoRy** bridges this gap, implementing massive matrices in memory (managed in R but implemented in C) and supporting their basic manipulation and exploration. It is ideal for problems involving the analysis in R of manageable subsets of the data, or when an analysis is conducted mostly in C.

In a Unix environment, the data structure may be allocated to shared memory, allowing separate R processes on the same computer to share access to a single copy of the data set; mutual exclusions (mutexes) are provided to avoid conflicts. This opens the door for more powerful parallel analyses and data mining of massive data sets.