BigR data

Hadley Wickham

Keywords: big data, ggplot2, plyr, visualisation, transformation

R has a notorious reputation for not being able to deal with "big" data (and **ggplot2** and **plyr** are frequent culprits). Fortunately, this isn't an underlying problem with *R*, and it's something that we can fix with good programming practices and intelligent use of compiled code. In this talk, I'll introduce two new packages, **bigvis** and **dplyr**, that aim to make it easier (and faster) to work with much larger datasets.

Bigvis makes it possible to visualise 10-100 million observations in just a few seconds. It is built around a pipeline of group, summarise, smooth and visualise, and makes minimal sacrifices of flexibility to achieve fast performance. As well as discussing the visualisation challenges when you have 10s of millions of observations, I'll also discuss the performance challenges, and how *C++* and **Rcpp** make it pleasurable to integrate compiled code into R.

Dplyr is an iteration of **plyr** that focusses on the tools people use most frequently (ddply, dlply and ldply), speed and on flexible data stores, so that you can use the same code regardless of whether you data is in a data frame, data table, or data base. I'll talk a little about how easy it is to compile simple *R* expressions into *SQL*, and on integrating *R* into a workflow when your complete dataset can't fit into memory, or even on the hard drive of a single machine.