Advanced R

David Walling

Data Group

Texas Advanced Computing Center

walling@tacc.utexas.edu

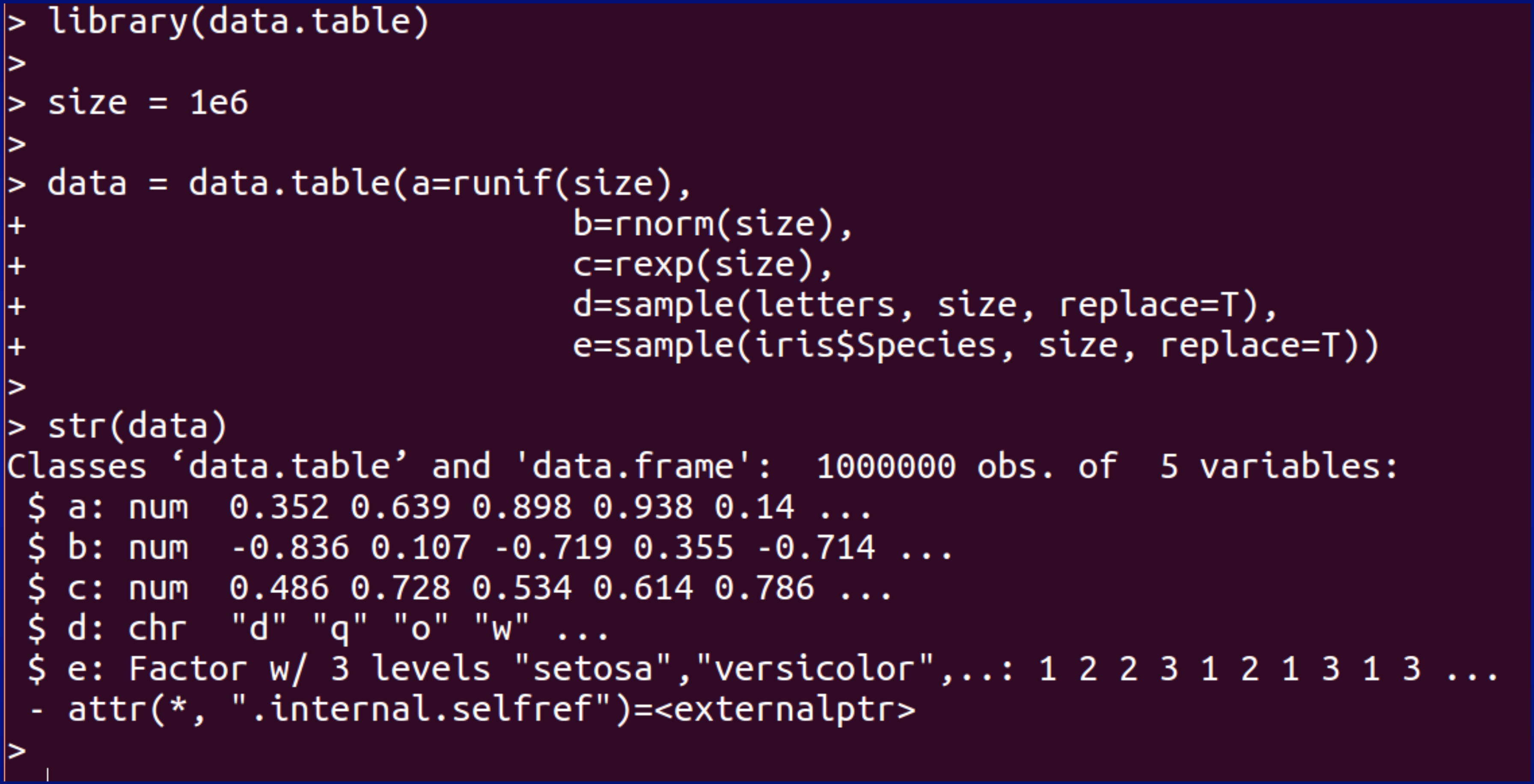
Topics

- data.table
- devtools & package installs
- Rcpp

data.table

- data.frame written in C++
- Very fast fread and fwrite
- A superset of data.frame functionality
- Automatic column names ‘attachment’
- Different syntax: DT[i, j, by]
    - i = filter/select
    - j = do something (aggregations)
    - by = group

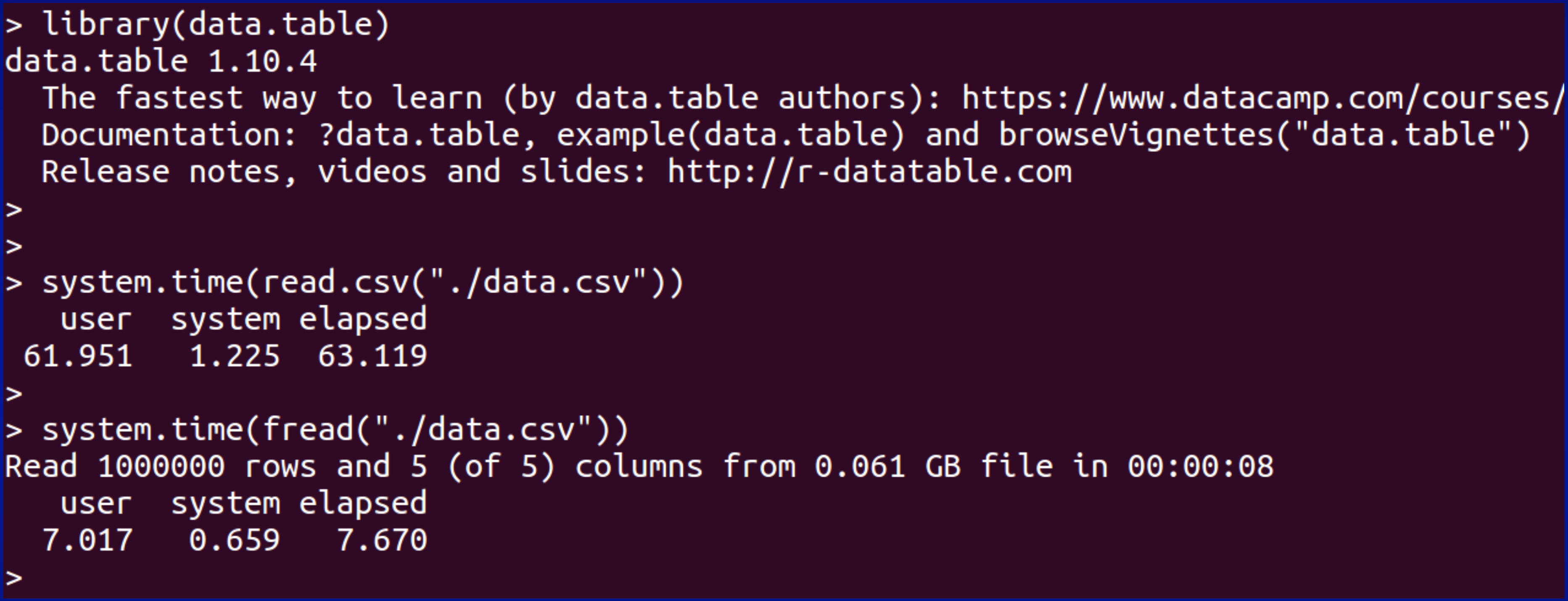
data.table Example



```
> library(data.table)
>
> size = 1e6
>
> data = data.table(a=runif(size),
+                           b=rnorm(size),
+                           c=rexp(size),
+                           d=sample(letters, size, replace=T),
+                           e=sample(iris$Species, size, replace=T))
>
> str(data)
Classes ‘data.table’ and 'data.frame':  1000000 obs. of  5 variables:
 $ a: num  0.352 0.639 0.898 0.938 0.14 ...
 $ b: num  -0.836 0.107 -0.719 0.355 -0.714 ...
 $ c: num  0.486 0.728 0.534 0.614 0.786 ...
 $ d: chr  "d" "q" "o" "w" ...
 $ e: Factor w/ 3 levels "setosa","versicolor",..: 1 2 2 3 1 2 1 3 1 3 ...
 - attr(*, ".internal.selfref")=<externalptr>
>
```

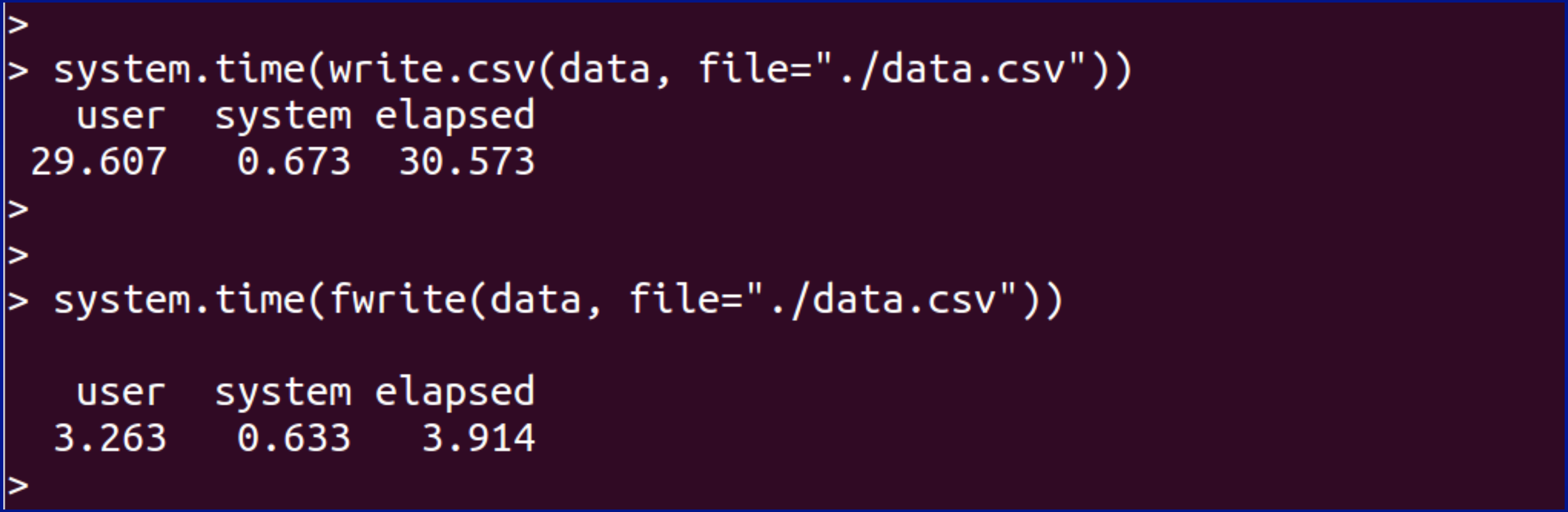
dataTable.R

data.table fread



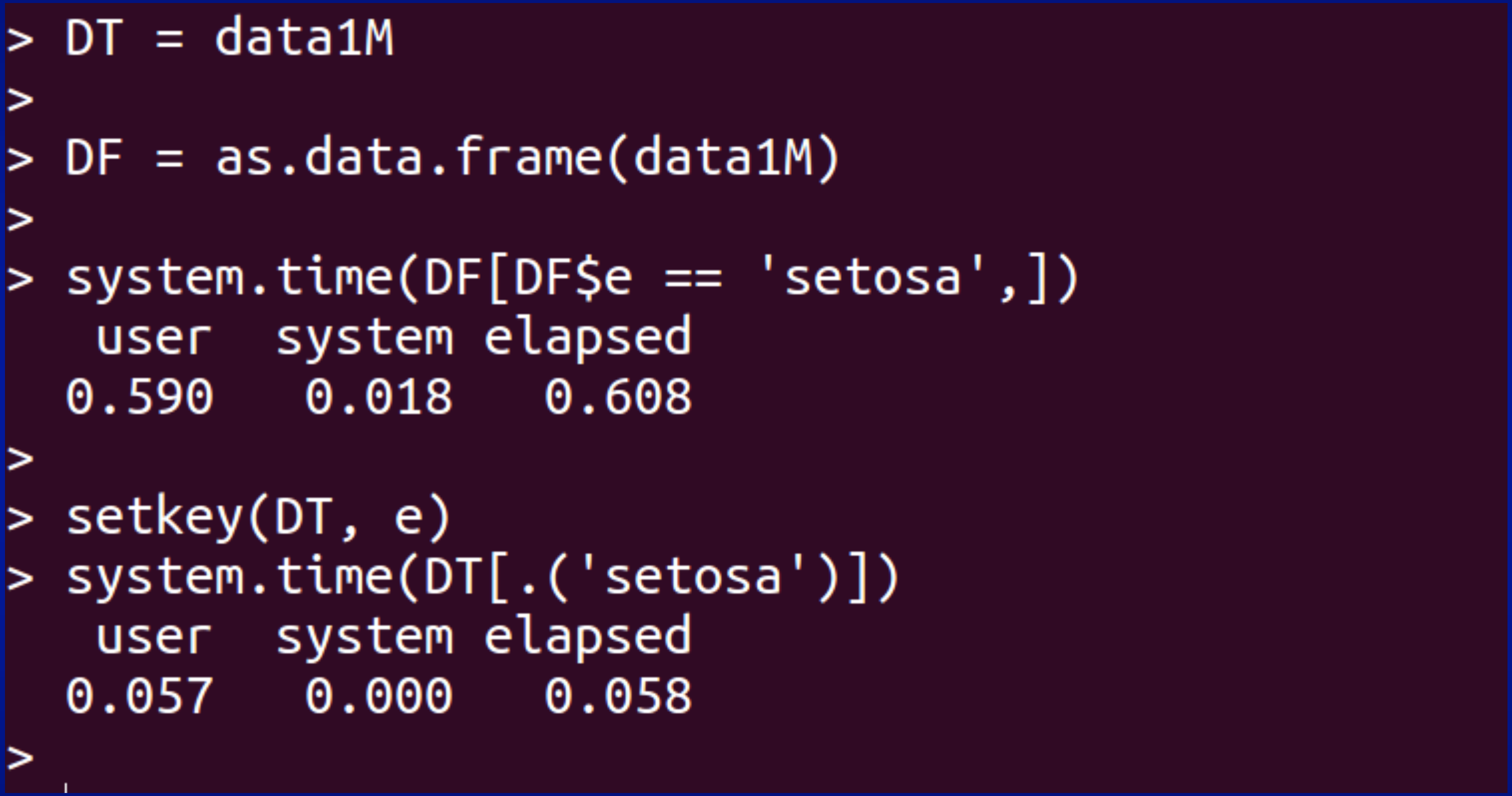
```
> library(data.table)
data.table 1.10.4
  The fastest way to learn (by data.table authors): https://www.datacamp.com/courses/
  Documentation: ?data.table, example(data.table) and browseVignettes("data.table")
  Release notes, videos and slides: http://r-datatable.com
>
>
> system.time(read.csv("./data.csv"))
   user  system elapsed
 61.951   1.225  63.119
>
> system.time(fread("./data.csv"))
Read 1000000 rows and 5 (of 5) columns from 0.061 GB file in 00:00:08
   user  system elapsed
  7.017   0.659   7.670
>
```

data.table fwrite



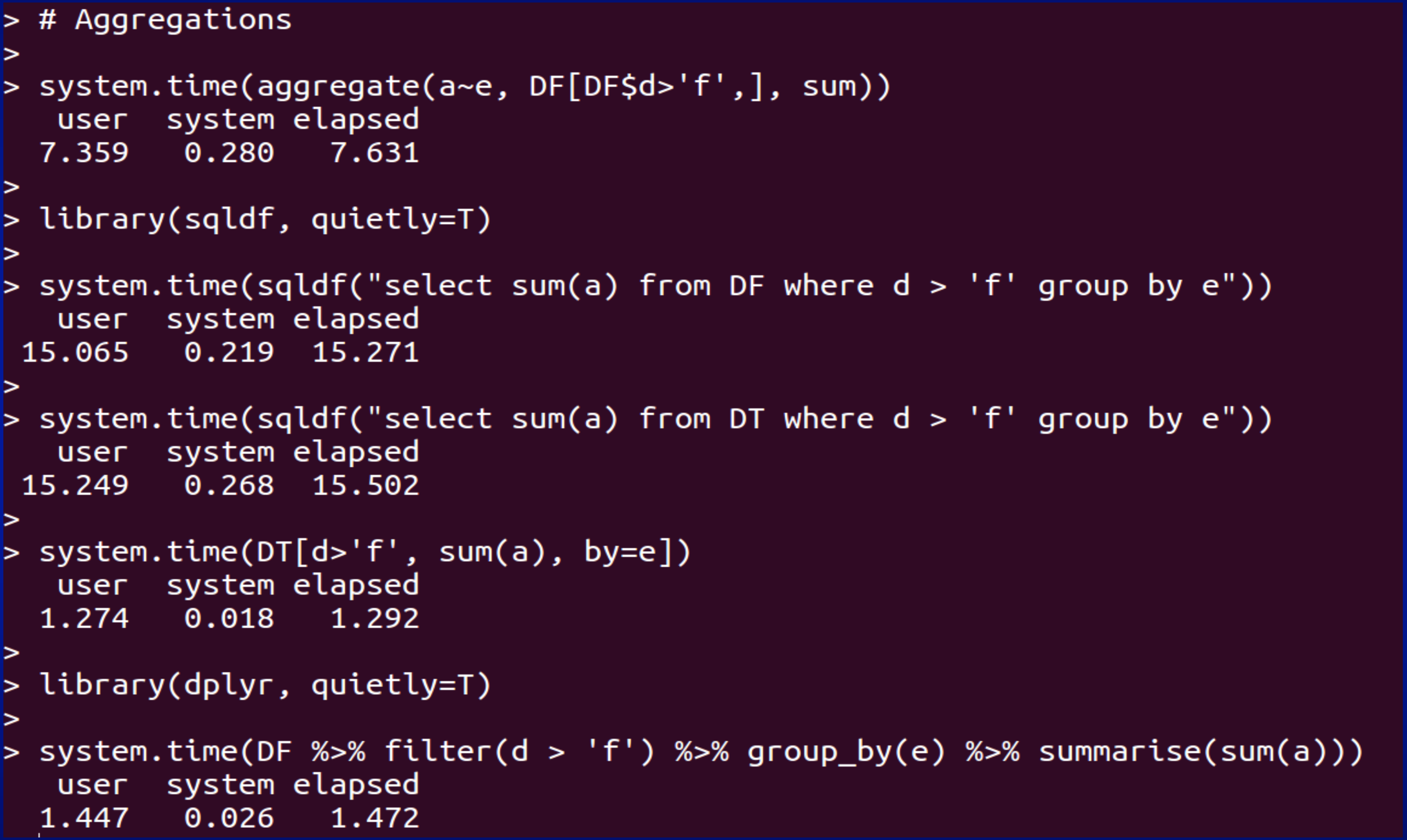
```
>
> system.time(write.csv(data, file="./data.csv"))
   user  system elapsed
 29.607   0.673  30.573
>
>
> system.time(fwrite(data, file="./data.csv"))

   user  system elapsed
  3.263   0.633   3.914
>
```

data.table subsetting



```
> DT = data1M
>
> DF = as.data.frame(data1M)
>
> system.time(DF[DF$e == 'setosa',])
   user  system elapsed
  0.590   0.018   0.608
>
> setkey(DT, e)
> system.time(DT[.('setosa')])
   user  system elapsed
  0.057   0.000   0.058
>
```

data.table + timings



```
> # Aggregations
>
> system.time(aggregate(a~e, DF[DF$d>'f',], sum))
   user  system elapsed
  7.359   0.280   7.631
>
> library(sqldf, quietly=T)
>
> system.time(sqldf("select sum(a) from DF where d > 'f' group by e"))
   user  system elapsed
 15.065   0.219  15.271
>
> system.time(sqldf("select sum(a) from DT where d > 'f' group by e"))
   user  system elapsed
 15.249   0.268  15.502
>
> system.time(DT[d>'f', sum(a), by=e])
   user  system elapsed
  1.274   0.018   1.292
>
> library(dplyr, quietly=T)
>
> system.time(DF %>% filter(d > 'f') %>% group_by(e) %>% summarise(sum(a)))
   user  system elapsed
  1.447   0.026   1.472
```

data.table Exercise

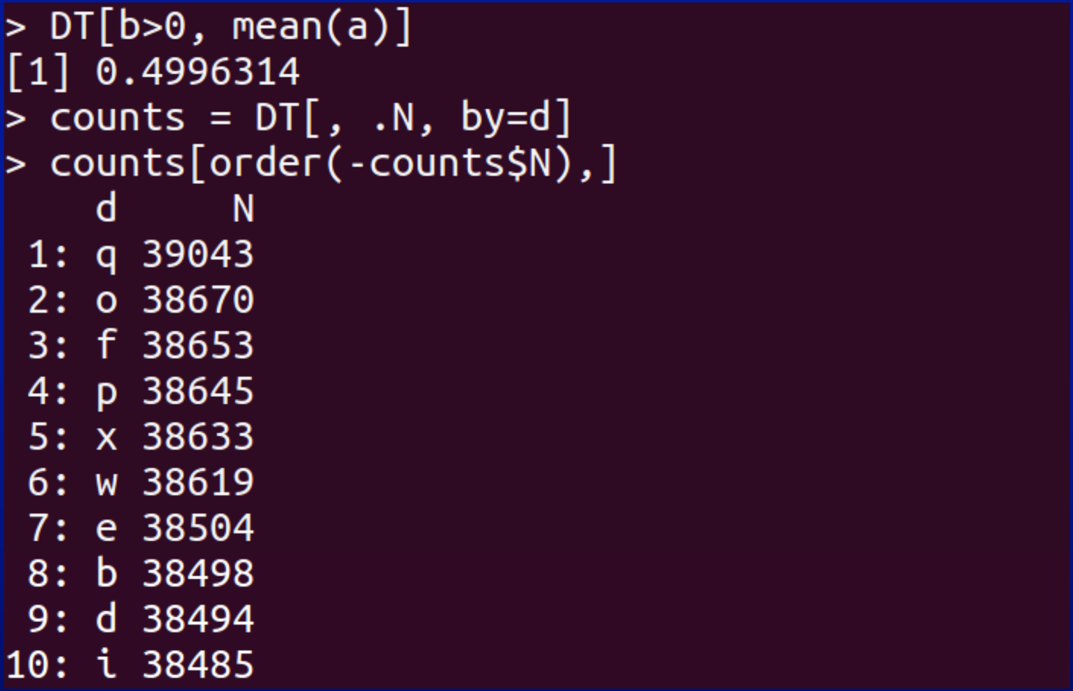
- Setup
  - –Launch an idev job on either ‘normal’ or ‘hadoop’ queue
  - –Start R session
  - –set.seed(1)
  - –Create our test data at size 1e6 (dataTable.R)

- Questions
  - –What is the average value of column a, for all rows where column b > 0?

  - –Which letter appears most frequently in column d? (Hint: .N gives counts for the ‘what’ part of data.table syntax)

data.table Exercise

- Questions
  - –What is the average value of column a, for all rows where column b > 0?
  - –Which letter appears most frequently in column d? (Hint: .N gives counts for the ‘what’ part of data.table syntax)



```
> DT[b>0, mean(a)]
[1] 0.4996314
> counts = DT[, .N, by=d]
> counts[order(-counts$N),]
    d     N
 1: q 39043
 2: o 38670
 3: f 38653
 4: p 38645
 5: x 38633
 6: w 38619
 7: e 38504
 8: b 38498
 9: d 38494
10: i 38485
```

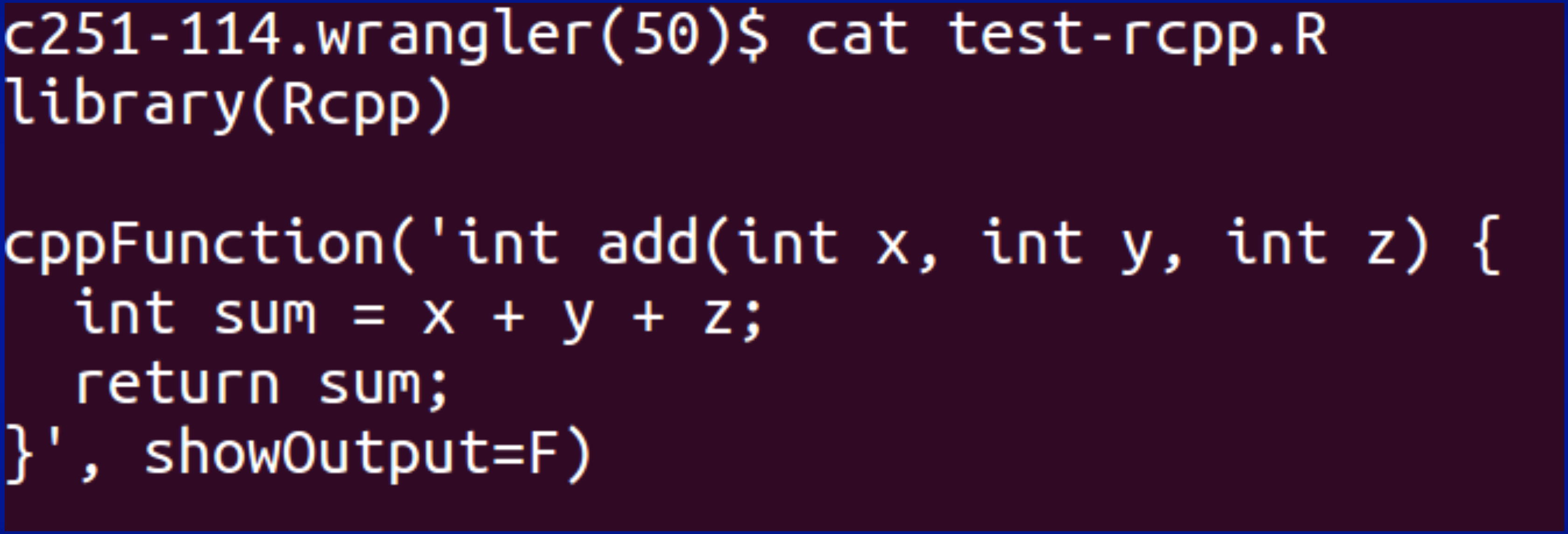
data.table In Depth

- Matt Dowle
  - https://rawgit.com/wiki/Rdatatable/data.table/vignettes/datatable-intro.html
- datacamp.com
  - https://www.datacamp.com/community/tutorials/data-table-r-tutorial

Rcpp

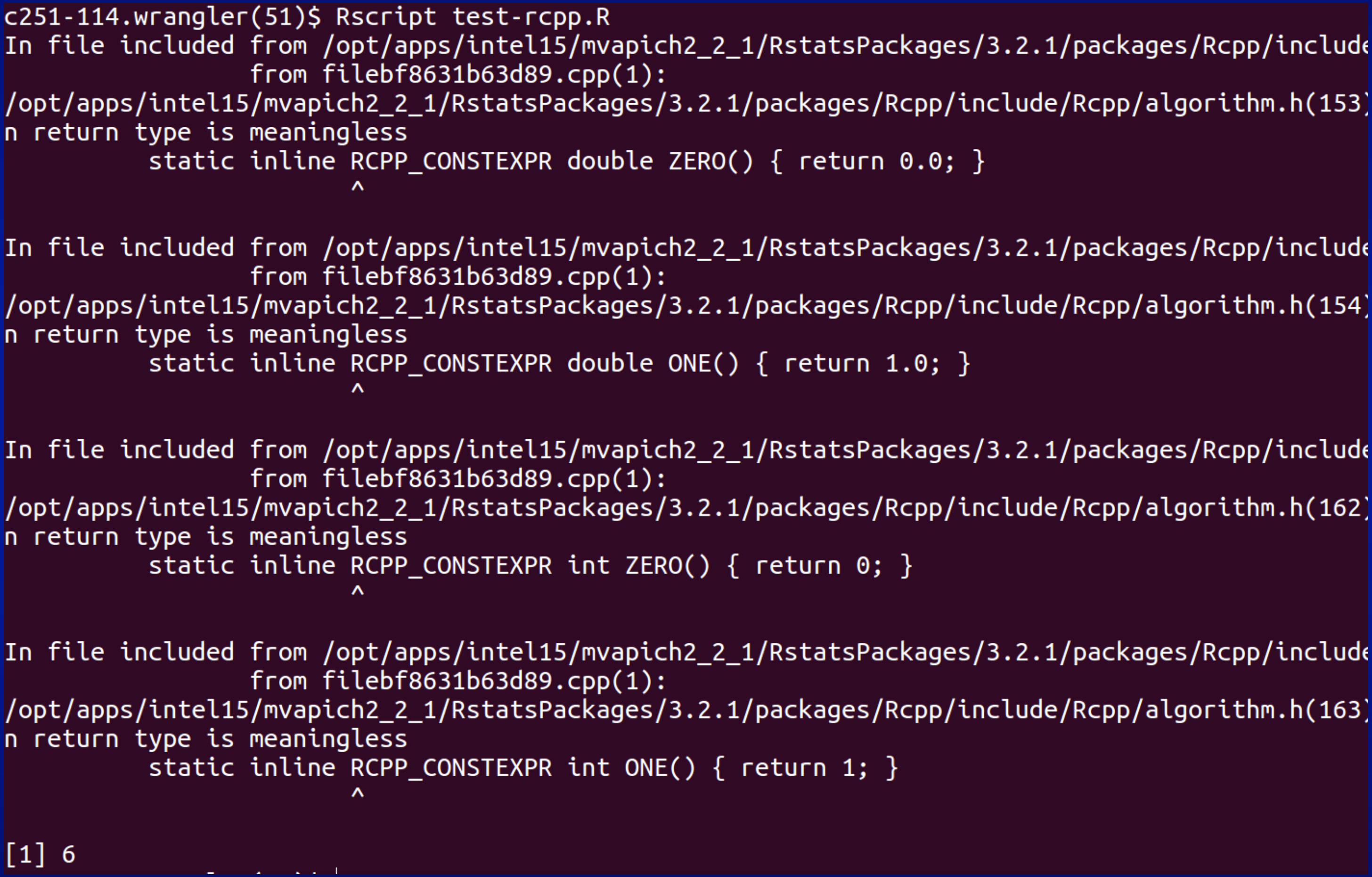
- Core of R is mostly C
- Some things are still slow
- Can often re-write bottlenecks directly in C++ for dramatic speed ups
- for loops and recursive functions are primary candidates

Rcpp Example



```
c251-114.wrangler(50)$ cat test-rcpp.R
library(Rcpp)

cppFunction('int add(int x, int y, int z) {
  int sum = x + y + z;
  return sum;
}', showOutput=F)
```

Rcpp Example



```
c251-114.wrangler(51)$ Rscript test-rcpp.R
In file included from /opt/apps/intel15/mvapich2_2_1/RstatsPackages/3.2.1/packages/Rcpp/include
                 from filebf8631b63d89.cpp(1):
/opt/apps/intel15/mvapich2_2_1/RstatsPackages/3.2.1/packages/Rcpp/include/Rcpp/algorithm.h(153
n return type is meaningless
          static inline RCPP_CONSTEXPR double ZERO() { return 0.0; }
                        ^

In file included from /opt/apps/intel15/mvapich2_2_1/RstatsPackages/3.2.1/packages/Rcpp/include
                 from filebf8631b63d89.cpp(1):
/opt/apps/intel15/mvapich2_2_1/RstatsPackages/3.2.1/packages/Rcpp/include/Rcpp/algorithm.h(154
n return type is meaningless
          static inline RCPP_CONSTEXPR double ONE() { return 1.0; }
                        ^

In file included from /opt/apps/intel15/mvapich2_2_1/RstatsPackages/3.2.1/packages/Rcpp/include
                 from filebf8631b63d89.cpp(1):
/opt/apps/intel15/mvapich2_2_1/RstatsPackages/3.2.1/packages/Rcpp/include/Rcpp/algorithm.h(162
n return type is meaningless
          static inline RCPP_CONSTEXPR int ZERO() { return 0; }
                        ^

In file included from /opt/apps/intel15/mvapich2_2_1/RstatsPackages/3.2.1/packages/Rcpp/include
                 from filebf8631b63d89.cpp(1):
/opt/apps/intel15/mvapich2_2_1/RstatsPackages/3.2.1/packages/Rcpp/include/Rcpp/algorithm.h(163
n return type is meaningless
          static inline RCPP_CONSTEXPR int ONE() { return 1; }
                        ^

[1] 6
```

Rcpp In Depth

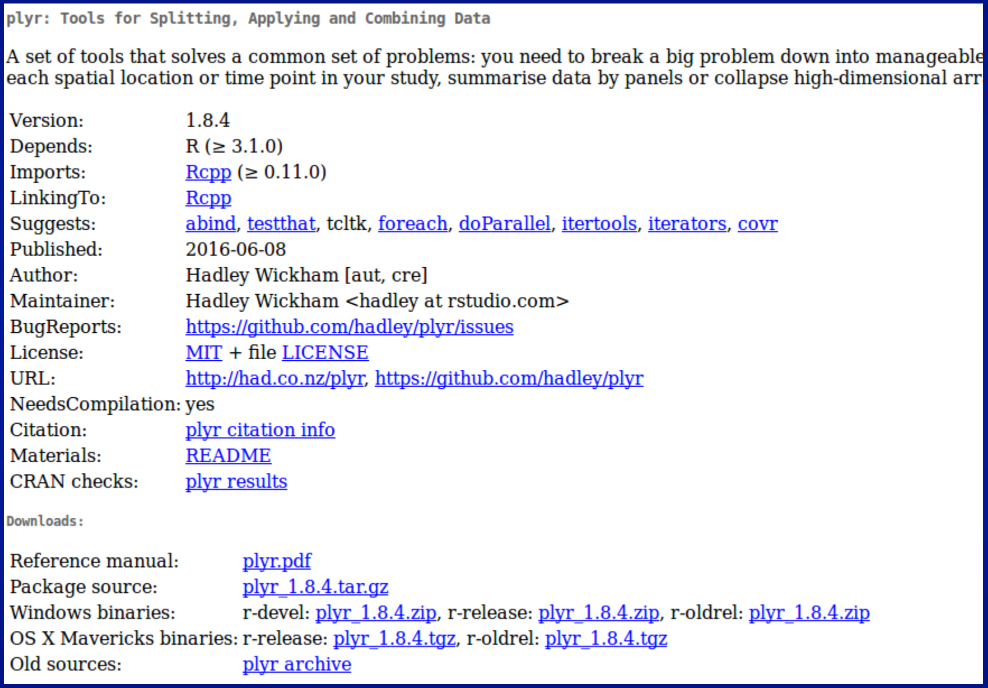
- Dirk Eddelbuettel
  - http://dirk.eddelbuettel.com/papers/rcpp_workshop_introduction_user2012.pdf
- RcppArmadillo
  - http://thecoatlessprofessor.com/programming/r-to-armadillo-using-rcpparmadillo-for-speed-and-portability/
- Hadley Wickham
  - http://adv-r.had.co.nz/Rcpp.html

devtools

- Many packages being distributed in github
  - install_github()
- TACC's R is usually behind latest, might need older versions of given package
  - install_version()

devtools

- Some packages more prone to updating core R dependency
- Our version is always a bit behind
- Option
  - –Build your own R in $WORK
    - Set PATH and LD_LIBRARY_PATH
  - –Install archived version of a package



plyr: Tools for Splitting, Applying and Combining Data

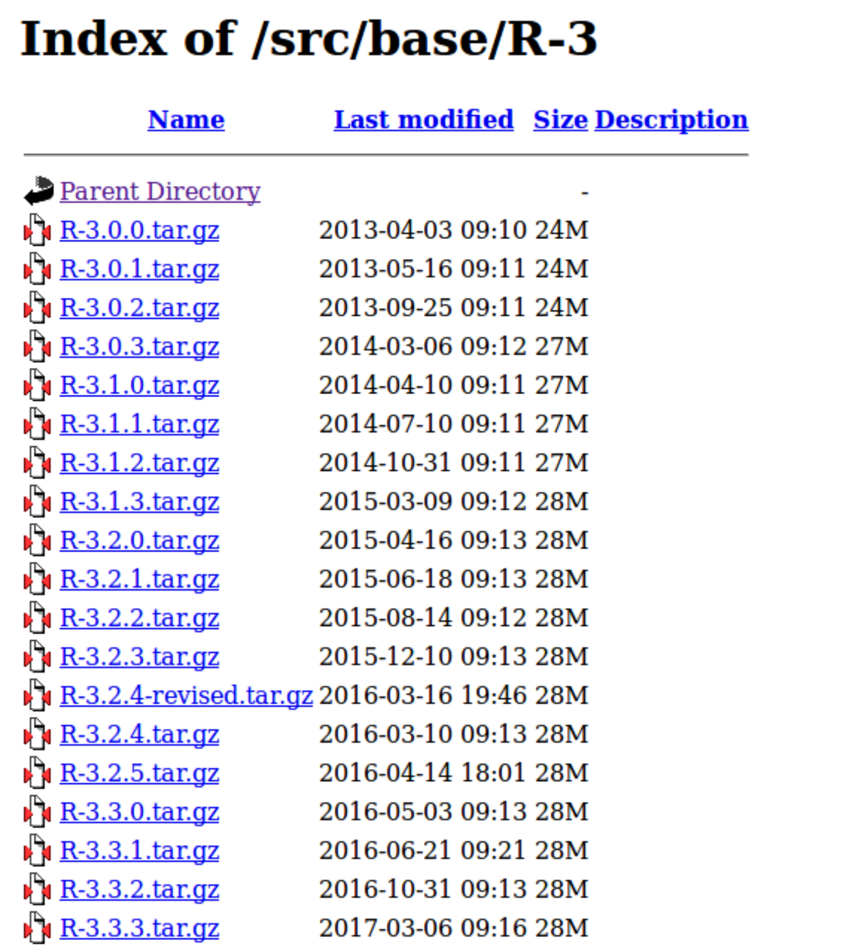
A set of tools that solves a common set of problems: you need to break a big problem down into manageable each spatial location or time point in your study, summarise data by panels or collapse high-dimensional arr

Version:	1.8.4
Depends:	R (≥ 3.1.0)
Imports:	Rcpp (≥ 0.11.0)
LinkingTo:	Rcpp
Suggests:	abind, testthat, tcltk, foreach, doParallel, itertools, iterators, covr
Published:	2016-06-08
Author:	Hadley Wickham [aut, cre]
Maintainer:	Hadley Wickham <hadley at rstudio.com>
BugReports:	https://github.com/hadley/plyr/issues
License:	MIT + file LICENSE
URL:	http://had.co.nz/plyr, https://github.com/hadley/plyr
NeedsCompilation:	yes
Citation:	plyr citation info
Materials:	README
CRAN checks:	plyr results

Downloads:

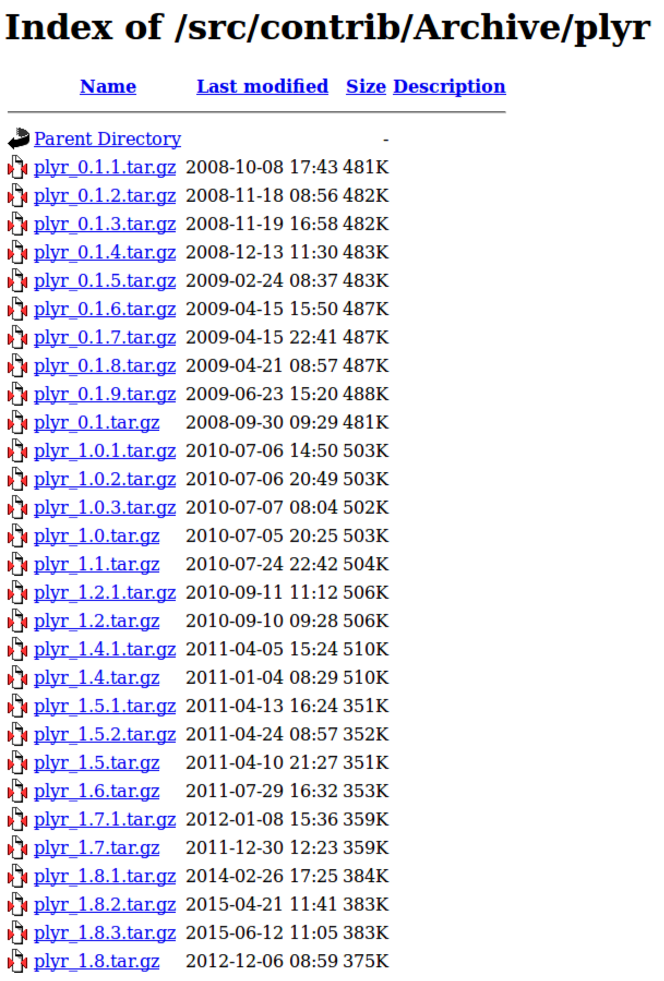
Reference manual:	plyr.pdf
Package source:	plyr_1.8.4.tar.gz
Windows binaries:	r-devel: plyr_1.8.4.zip, r-release: plyr_1.8.4.zip, r-oldrel: plyr_1.8.4.zip
OS X Mavericks binaries:	r-release: plyr_1.8.4.tgz, r-oldrel: plyr_1.8.4.tgz
Old sources:	plyr archive

devtools



Index of /src/base/R-3

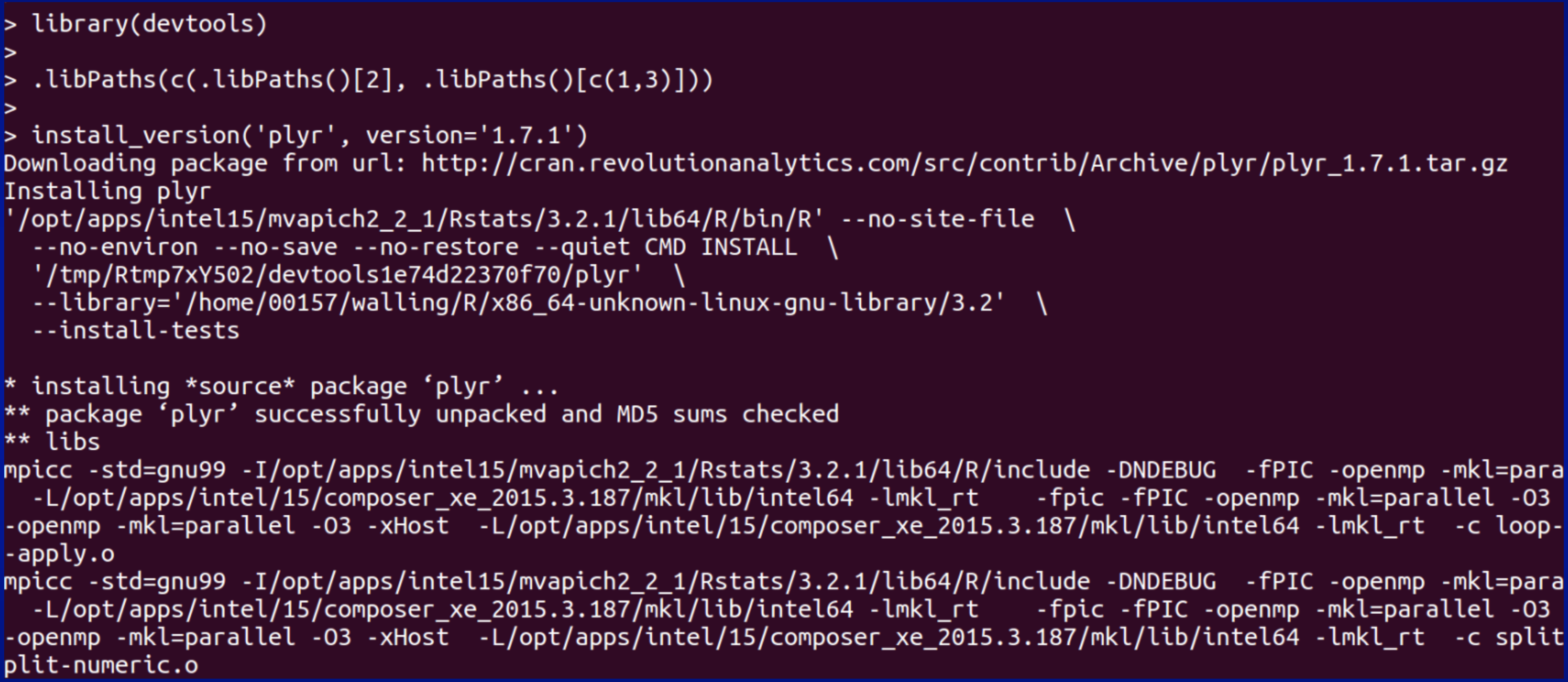
Name	Last modified	Size	Description
Parent Directory		-	
R-3.0.0.tar.gz	2013-04-03 09:10	24M	
R-3.0.1.tar.gz	2013-05-16 09:11	24M	
R-3.0.2.tar.gz	2013-09-25 09:11	24M	
R-3.0.3.tar.gz	2014-03-06 09:12	27M	
R-3.1.0.tar.gz	2014-04-10 09:11	27M	
R-3.1.1.tar.gz	2014-07-10 09:11	27M	
R-3.1.2.tar.gz	2014-10-31 09:11	27M	
R-3.1.3.tar.gz	2015-03-09 09:12	28M	
R-3.2.0.tar.gz	2015-04-16 09:13	28M	
R-3.2.1.tar.gz	2015-06-18 09:13	28M	
R-3.2.2.tar.gz	2015-08-14 09:12	28M	
R-3.2.3.tar.gz	2015-12-10 09:13	28M	
R-3.2.4-revised.tar.gz	2016-03-16 19:46	28M	
R-3.2.4.tar.gz	2016-03-10 09:13	28M	
R-3.2.5.tar.gz	2016-04-14 18:01	28M	
R-3.3.0.tar.gz	2016-05-03 09:13	28M	
R-3.3.1.tar.gz	2016-06-21 09:21	28M	
R-3.3.2.tar.gz	2016-10-31 09:13	28M	
R-3.3.3.tar.gz	2017-03-06 09:16	28M	



Index of /src/contrib/Archive/plyr

Name	Last modified	Size	Description
Parent Directory		-	
plyr_0.1.1.tar.gz	2008-10-08 17:43	481K	
plyr_0.1.2.tar.gz	2008-11-18 08:56	482K	
plyr_0.1.3.tar.gz	2008-11-19 16:58	482K	
plyr_0.1.4.tar.gz	2008-12-13 11:30	483K	
plyr_0.1.5.tar.gz	2009-02-24 08:37	483K	
plyr_0.1.6.tar.gz	2009-04-15 15:50	487K	
plyr_0.1.7.tar.gz	2009-04-15 22:41	487K	
plyr_0.1.8.tar.gz	2009-04-21 08:57	487K	
plyr_0.1.9.tar.gz	2009-06-23 15:20	488K	
plyr_0.1.tar.gz	2008-09-30 09:29	481K	
plyr_1.0.1.tar.gz	2010-07-06 14:50	503K	
plyr_1.0.2.tar.gz	2010-07-06 20:49	503K	
plyr_1.0.3.tar.gz	2010-07-07 08:04	502K	
plyr_1.0.tar.gz	2010-07-05 20:25	503K	
plyr_1.1.tar.gz	2010-07-24 22:42	504K	
plyr_1.2.1.tar.gz	2010-09-11 11:12	506K	
plyr_1.2.tar.gz	2010-09-10 09:28	506K	
plyr_1.4.1.tar.gz	2011-04-05 15:24	510K	
plyr_1.4.tar.gz	2011-01-04 08:29	510K	
plyr_1.5.1.tar.gz	2011-04-13 16:24	351K	
plyr_1.5.2.tar.gz	2011-04-24 08:57	352K	
plyr_1.5.tar.gz	2011-04-10 21:27	351K	
plyr_1.6.tar.gz	2011-07-29 16:32	353K	
plyr_1.7.1.tar.gz	2012-01-08 15:36	359K	
plyr_1.7.tar.gz	2011-12-30 12:23	359K	
plyr_1.8.1.tar.gz	2014-02-26 17:25	384K	
plyr_1.8.2.tar.gz	2015-04-21 11:41	383K	
plyr_1.8.3.tar.gz	2015-06-12 11:05	383K	
plyr_1.8.tar.gz	2012-12-06 08:59	375K	

devtools



```
> library(devtools)
>
> .libPaths(c(.libPaths()[2], .libPaths()[c(1,3)]))
>
> install_version('plyr', version='1.7.1')
Downloading package from url: http://cran.revolutionanalytics.com/src/contrib/Archive/plyr/plyr_1.7.1.tar.gz
Installing plyr
'/opt/apps/intel15/mvapich2_2_1/Rstats/3.2.1/lib64/R/bin/R' --no-site-file  \
  --no-environ --no-save --no-restore --quiet CMD INSTALL  \
  '/tmp/Rtmp7xY502/devtools1e74d22370f70/plyr'  \
  --library='/home/00157/walling/R/x86_64-unknown-linux-gnu-library/3.2'  \
  --install-tests

* installing *source* package ‘plyr’ ...
** package ‘plyr’ successfully unpacked and MD5 sums checked
** libs
mpicc -std=gnu99 -I/opt/apps/intel15/mvapich2_2_1/Rstats/3.2.1/lib64/R/include -DNDEBUG  -fPIC -openmp -mkl=para
  -L/opt/apps/intel/15/composer_xe_2015.3.187/mkl/lib/intel64 -lmkl_rt    -fpic -fPIC -openmp -mkl=parallel -O3
-openmp -mkl=parallel -O3 -xHost  -L/opt/apps/intel/15/composer_xe_2015.3.187/mkl/lib/intel64 -lmkl_rt  -c loop-
-apply.o
mpicc -std=gnu99 -I/opt/apps/intel15/mvapich2_2_1/Rstats/3.2.1/lib64/R/include -DNDEBUG  -fPIC -openmp -mkl=para
  -L/opt/apps/intel/15/composer_xe_2015.3.187/mkl/lib/intel64 -lmkl_rt    -fpic -fPIC -openmp -mkl=parallel -O3
-openmp -mkl=parallel -O3 -xHost  -L/opt/apps/intel/15/composer_xe_2015.3.187/mkl/lib/intel64 -lmkl_rt  -c split
plit-numeric.o
```

David Walling

walling@tacc.utexas.edu