

News from data.table 1.4 and 1.5

Matthew Dowle

LondonR, July 2010

3 new vignettes

- Quick 10 minute introduction
- FAQs
- Reproducible timings :

```
      base data.table times faster
==      5.556      0.012      462
tapply 11.636      0.832      13
```

Radix sorting

- Added to data.table by Tom Short
- `?base::sort.list(x,method="radix")`
- Very fast
- http://en.wikipedia.org/wiki/Radix_sort
- Specifically for integers **only**
- That's *why* key columns must be integer/factor
- Might add `decimal()` in future

IDate

- Added to data.table by Tom Short
- Like Date but integer storage
- We need integer for radix

Example :

```
DT[, fitdistr(d1-d2,"normal"), by=month(d1)]
```

Fast grouping

- Allocates memory for largest group
- Reuses that same memory for all groups
- Allocates result data.table up front
- Implemented in C

Ad-hoc vs key'd by

- Key'd by :

`key(DT) = "colA" # same as setkey(DT,colA)`

`DT[, sum(colB), by=colA]`

- Ad-hoc by :

`key(DT) = NULL`

`DT[, sum(colB), by=colA]`

- Difference is memcopy internally

Inherits from data.frame

- `class(data.table) = c("data.table", "data.frame")`
- `is.data.frame()` is now TRUE
- Now works easily with packages that **only** accept `data.frame`, such as `ggplot` and `lattice`.
- Technically tricky, otherwise would have done this years ago. Uses `base::topenv()`.

Small syntax changes

- Dropped DT() alias
- j is now list() of expressions
- by is now list() of expressions
- .SD available to j
- Reuse of expressions :
 `e = quote(list(fun1(colA),fun2(colB)))`
 `DT[, mean(colC), by=eval(e)]`

Unit tests continue to grow

- `test.data.table()`
- Tests have grown to 172
- Simple self contained test method
- Called from `.Rd` example, so 'R CMD check' runs `test.data.table()`
- Runs every night on R-Forge and CRAN

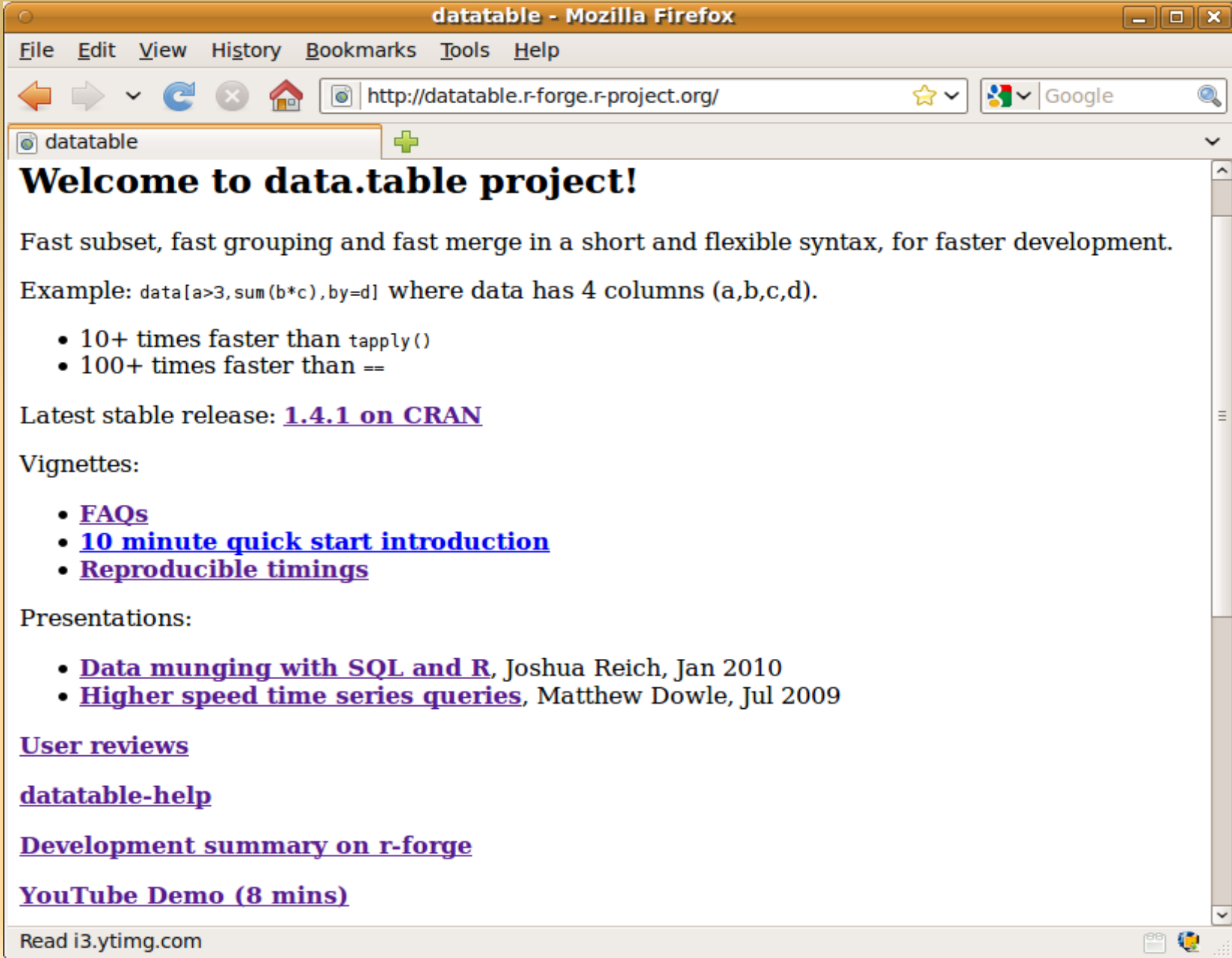
Available on all platforms

CRAN Package Check Results for Package [data.table](#)

Last updated on 2010-07-12 23:53:27.

Flavor	Version	Tinstall	Tcheck	Ttotal	Status	Flags
r-devel-linux-ix86	1.4.1	1.04	56.27	57.31	OK	
r-devel-linux-x86_64-gcc-debian	1.4.1	1.20	68.62	69.82	OK	
r-devel-linux-x86_64-gcc-fedora	1.4.1			76.24	OK	
r-patched-linux-ix86	1.4.1	1.04	63.70	64.74	OK	
r-patched-linux-x86_64	1.4.1	1.25	71.85	73.10	OK	
r-patched-solaris-sparc	1.4.1			558.70	OK	
r-patched-solaris-x86	1.4.1			106.50	OK	
r-release-linux-ix86	1.4.1	1.21	92.55	93.76	OK	
r-release-macosx-ix86	1.4.1	2.00	80.00	82.00	OK	
r-release-windows-ix86	1.4.1	3.00	88.00	91.00	OK	
r-release-windows64-x86_64	1.4.1	4.00	109.00	113.00	OK	
r-oldrel-macosx-ix86	1.4.1	2.00	79.00	81.00	OK	
r-oldrel-windows-ix86	1.4.1	3.00	87.00	90.00	OK	

New homepage



The screenshot shows a Mozilla Firefox browser window with the title "datatable - Mozilla Firefox". The address bar contains "http://datatable.r-forge.r-project.org/". The page content includes a welcome message, a description of the project's speed, an example of data manipulation, and links to FAQs, vignettes, presentations, user reviews, help, development summary, and a YouTube demo.

Welcome to data.table project!

Fast subset, fast grouping and fast merge in a short and flexible syntax, for faster development.

Example: `data[a>3, sum(b*c), by=d]` where data has 4 columns (a,b,c,d).

- 10+ times faster than `tapply()`
- 100+ times faster than `==`

Latest stable release: [1.4.1 on CRAN](#)

Vignettes:

- [FAQs](#)
- [10 minute quick start introduction](#)
- [Reproducible timings](#)

Presentations:

- [Data munging with SQL and R](#), Joshua Reich, Jan 2010
- [Higher speed time series queries](#), Matthew Dowle, Jul 2009

[User reviews](#)

[datatable-help](#)

[Development summary on r-forge](#)

[YouTube Demo \(8 mins\)](#)

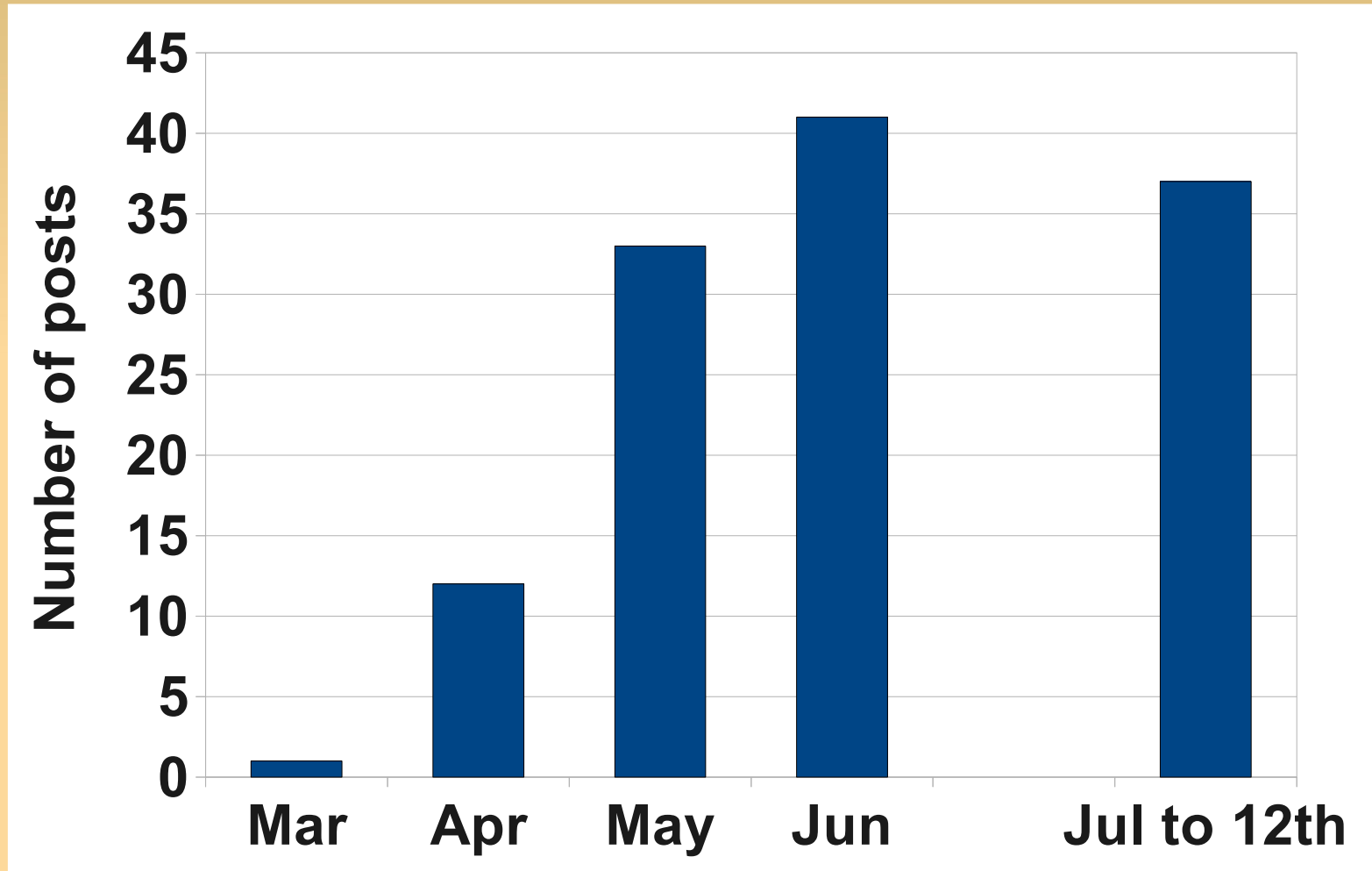
Read i3.ytimg.com

YouTube video

Last year's demo here at LondonR (700 views) :



datatable-help



Quotes from users

- "I don't know where I would be w/o data.table"
- "Fast splitting/sorting operations in frames"
- "The fast way to do SQL like operations in R"
- "I can't believe my eyes, I wish I had looked earlier"

Other relevant packages

- plyr by Hadley Wickham
 - sqldf by Gabor Grothendieck
 - doBy by Søren Højsgaard
 - indexing by Jeff Ryan
 - mmap by Jeff Ryan
 - ff by Jens Oehlschlägel and team
 - refdata (package ref) by Jens Oehlschlägel
 - remix by David Hajage
 - reshape by Hadley Wickham
 - bigtabulate by Michael Kane and John Emerson
- + more ? ... please let me know.