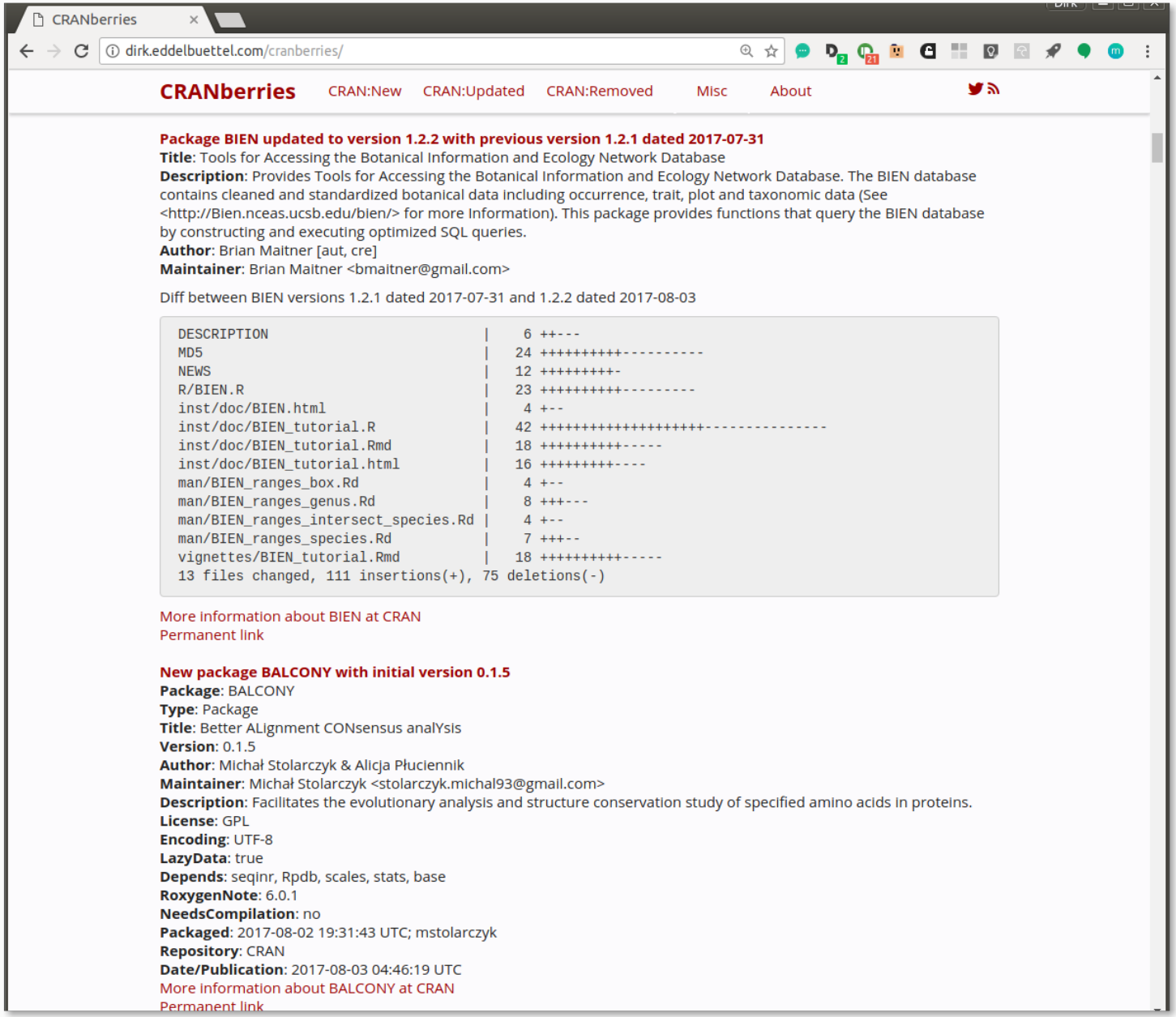
R for System Adminstration: An Example

Lightning Talk @ *Open Source Open Mic*

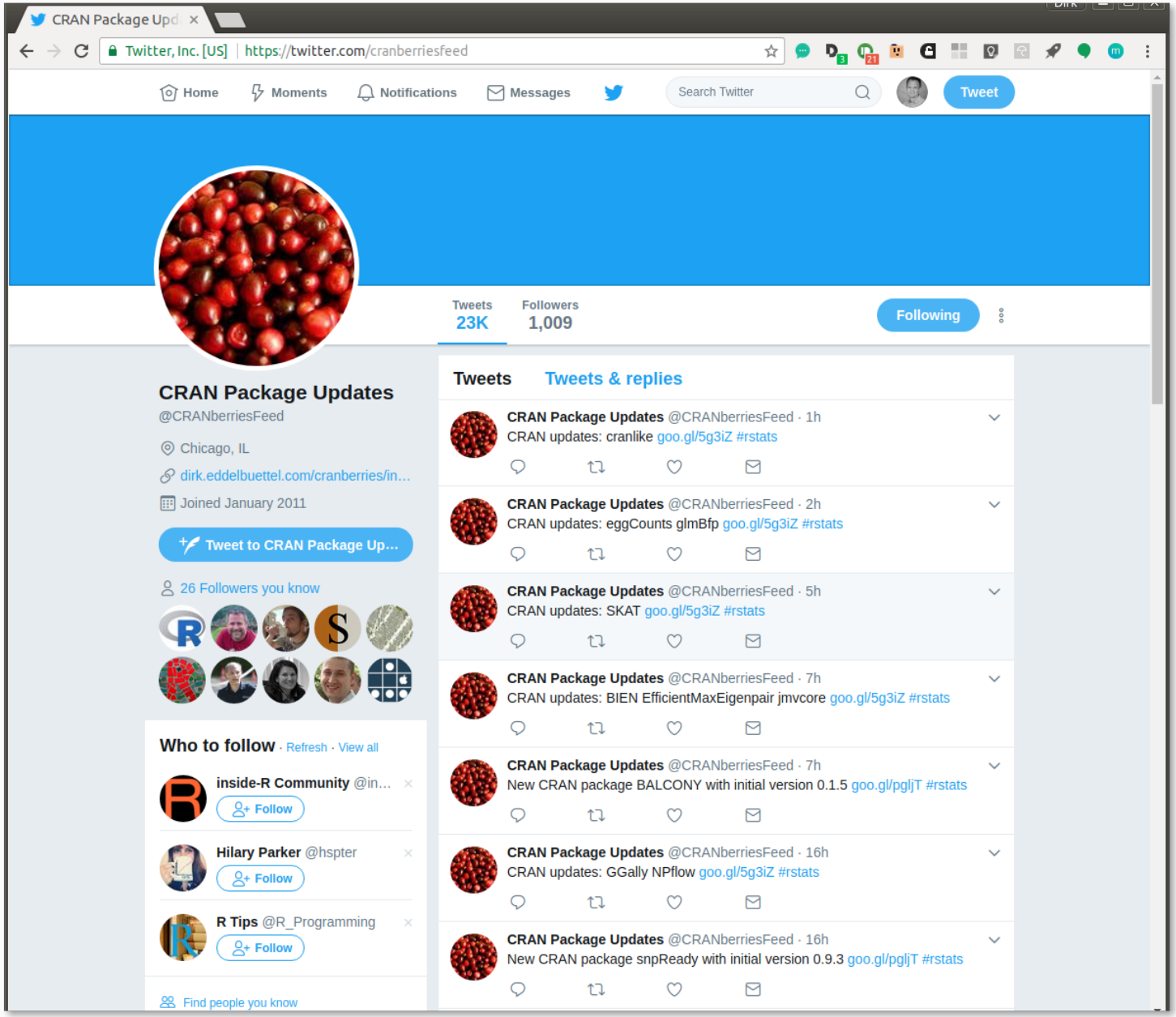
Dirk Eddelbuettel

August 3, 2017

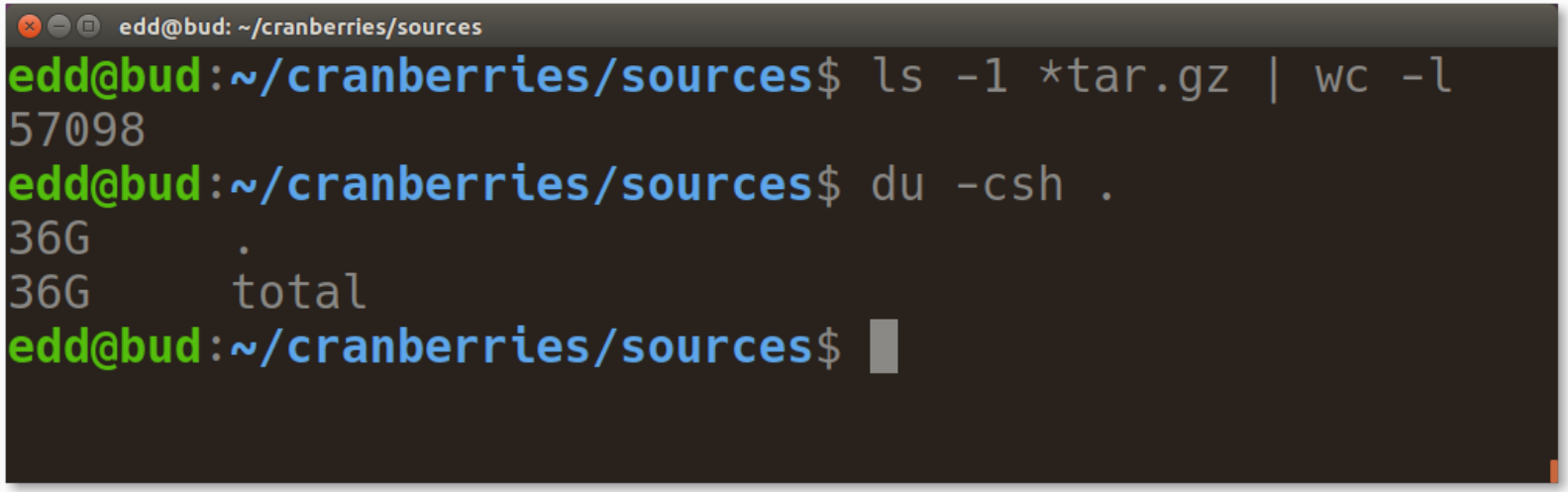
Motivation I



Motivation II

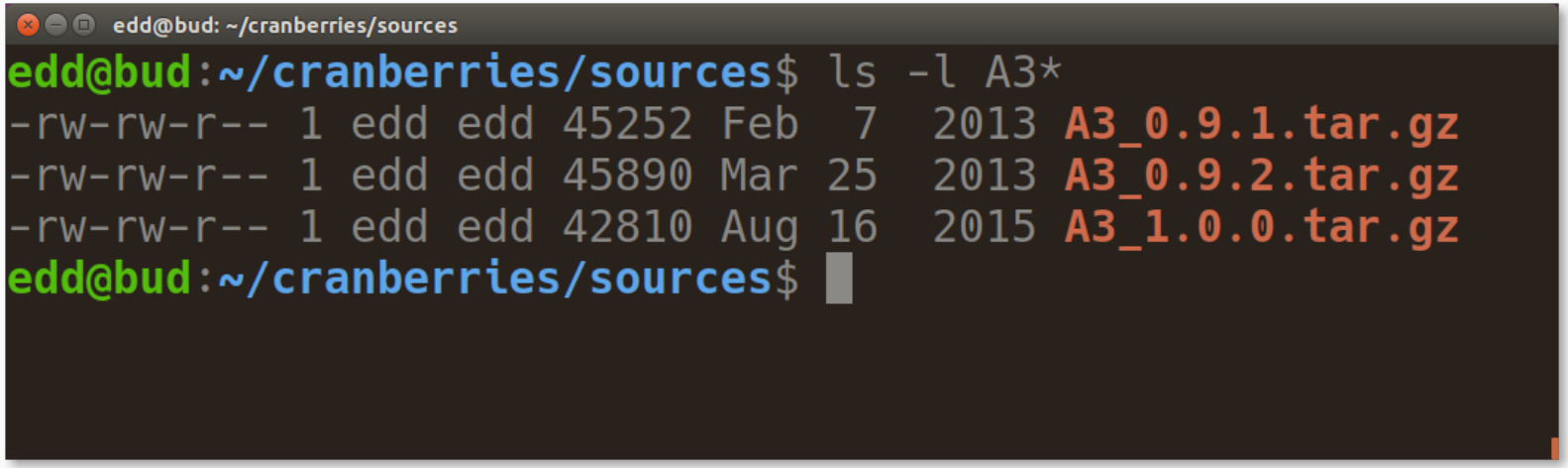


“Cost”



Over 57k files across 11k distinct packages

“Problem”



For each of the 11k package ‘sets’, find the not-newest ones.

(One of many possible) Solution(s)

data.table

- Highly-efficient data.frame extension
- Fast (by reference) aggregation/add/delete/join/group/...
- Operates via SQL-alike `dt[i, j, by]`
- More at r-datatable.com

Step 1

```
## load package, silently
suppressMessages(library("data.table"))

## source directory
dir <- "~/cranberries/sources/"

## read files, optionally recursively (not here)
files <- list.files(dir, pattern="*tar.gz",
                    full.names=TRUE)
```

Output from Step 1

```
R> suppressMessages(library("data.table"))
R> dir <- "~/cranberries/sources"
R> files <- list.files(dir, pattern="*tar.gz", full.names=TRUE)
R> d <- data.table(name=files, file.info(files))
R> d
                                                name   size isdir mode                       mtime
    1:    /home/edd/cranberries/sources/A3_0.9.1.tar.gz  45252 FALSE  664 2013-02-07 04:21:57.729160
    2:    /home/edd/cranberries/sources/A3_0.9.2.tar.gz  45890 FALSE  664 2013-03-25 02:12:03.437618
    3:    /home/edd/cranberries/sources/A3_1.0.0.tar.gz  42810 FALSE  664 2015-08-16 16:13:09.811910
    4:  /home/edd/cranberries/sources/aaMI_1.0-1.tar.gz   3487 FALSE  644 2005-10-17 14:24:18.000000
   ---
57108: /home/edd/cranberries/sources/ztype_0.1.0.tar.gz   3792 FALSE  664 2016-12-22 18:31:08.054666
57109:  /home/edd/cranberries/sources/zyp_0.10-1.tar.gz   7280 FALSE  664 2013-09-19 02:12:34.938652
57110:   /home/edd/cranberries/sources/zyp_0.9-1.tar.gz   5801 FALSE  644 2009-01-09 02:02:28.000000
57111:   /home/edd/cranberries/sources/zyp_0.9-3.tar.gz   6959 FALSE  664 2013-08-23 18:12:29.087391
                           ctime                       atime  uid  gid uname grname
    1: 2017-06-23 07:28:42.504387 2017-06-25 19:36:52.266695 1000 1000   edd    edd
    2: 2017-06-23 07:28:42.504387 2017-06-25 19:36:52.266695 1000 1000   edd    edd
    3: 2017-06-23 07:28:42.660387 2017-06-25 19:36:52.274695 1000 1000   edd    edd
    4: 2017-06-23 07:54:58.099709 2017-06-25 19:57:58.533000 1000 1000   edd    edd
   ---
57108: 2017-06-23 08:46:25.754497 2017-06-25 20:33:16.887890 1000 1000   edd    edd
57109: 2017-06-23 08:46:25.990497 2017-06-25 20:33:16.895890 1000 1000   edd    edd
57110: 2017-06-23 08:46:25.990497 2017-06-25 20:33:16.895890 1000 1000   edd    edd
57111: 2017-06-23 08:46:25.990497 2017-06-25 20:33:16.895890 1000 1000   edd    edd
R>
```

Step 2

```
d[, baseNM := basename(name)]
d[, nameNE := gsub(".tar.gz$", "", baseNM)]
d[, pkg := gsub("(.*)_.*$", "\\1", nameNE)]
d[, ver := gsub(".*_(.*)$", "\\1", nameNE)]
d
```

Output from Step 2

```
R> d[, baseNM := basename(name)]
R> d[, nameNE := gsub(".tar.gz$", "", baseNM)]
R> d[, pkg := gsub("(.*)_.*$", "\\1", nameNE)]
R> d[, ver := gsub(".*_(.*)$", "\\1", nameNE)]
R> d
                                                  name   size isdir mode                       mtime
    1:    /home/edd/cranberries/sources/A3_0.9.1.tar.gz  45252 FALSE  664 2013-02-07 04:21:57.729160
    2:    /home/edd/cranberries/sources/A3_0.9.2.tar.gz  45890 FALSE  664 2013-03-25 02:12:03.437618
   ---
57110:   /home/edd/cranberries/sources/zyp_0.9-1.tar.gz   5801 FALSE  644 2009-01-09 02:02:28.000000
57111:   /home/edd/cranberries/sources/zyp_0.9-3.tar.gz   6959 FALSE  664 2013-08-23 18:12:29.087391
                          ctime                       atime  uid  gid uname grname             baseNM
    1: 2017-06-23 07:28:42.504387 2017-06-25 19:36:52.266695 1000 1000   edd    edd    A3_0.9.1.tar.gz
    2: 2017-06-23 07:28:42.504387 2017-06-25 19:36:52.266695 1000 1000   edd    edd    A3_0.9.2.tar.gz
   ---
57110: 2017-06-23 08:46:25.990497 2017-06-25 20:33:16.895890 1000 1000   edd    edd   zyp_0.9-1.tar.gz
57111: 2017-06-23 08:46:25.990497 2017-06-25 20:33:16.895890 1000 1000   edd    edd   zyp_0.9-3.tar.gz
            nameNE    pkg    ver
    1:    A3_0.9.1     A3  0.9.1
    2:    A3_0.9.2     A3  0.9.2
   ---
57110:   zyp_0.9-1    zyp  0.9-1
57111:   zyp_0.9-3    zyp  0.9-3
R>
```

Step 3

```
## set keys on package and mod.time
setkeyv(d, c("pkg", "mtime"))

## compute boolean on 'is it latesg
d[ , newest := (ver == last(ver)), by=pkg]

## view 'name' of those that are not latest, cap'ed at 5
d[ newest==FALSE, name][1:5]
```

Output from Step 3

```
R> setkeyv(d, c("pkg", "mtime"))
R>
R> d[ , newest := (ver == last(ver)), by=pkg]
R>
R> d[ newest==FALSE, name][1:5]
[1] "/home/edd/cranberries/sources/A3_0.9.1.tar.gz"
[2] "/home/edd/cranberries/sources/A3_0.9.2.tar.gz"
[3] "/home/edd/cranberries/sources/ABCanalysis_1.0.tar.gz"
[4] "/home/edd/cranberries/sources/ABCanalysis_1.0.1.tar.gz"
[5] "/home/edd/cranberries/sources/ABCanalysis_1.0.2.tar.gz"
R>
```

and those are our files to `unlink()` or move or ...

Summary

R

- lets us access standard POSIX functions
- in a vectorised manner

Operating ‘sys-admin’ style on files and folders

- is just Programming with Data
- for which R is ideally suited

Contact

Email

dirk@eddelbuettel.com

Website

http://dirk.eddelbuettel.com

Twitter

@eddelbuettel

GitHub

eddelbuettel