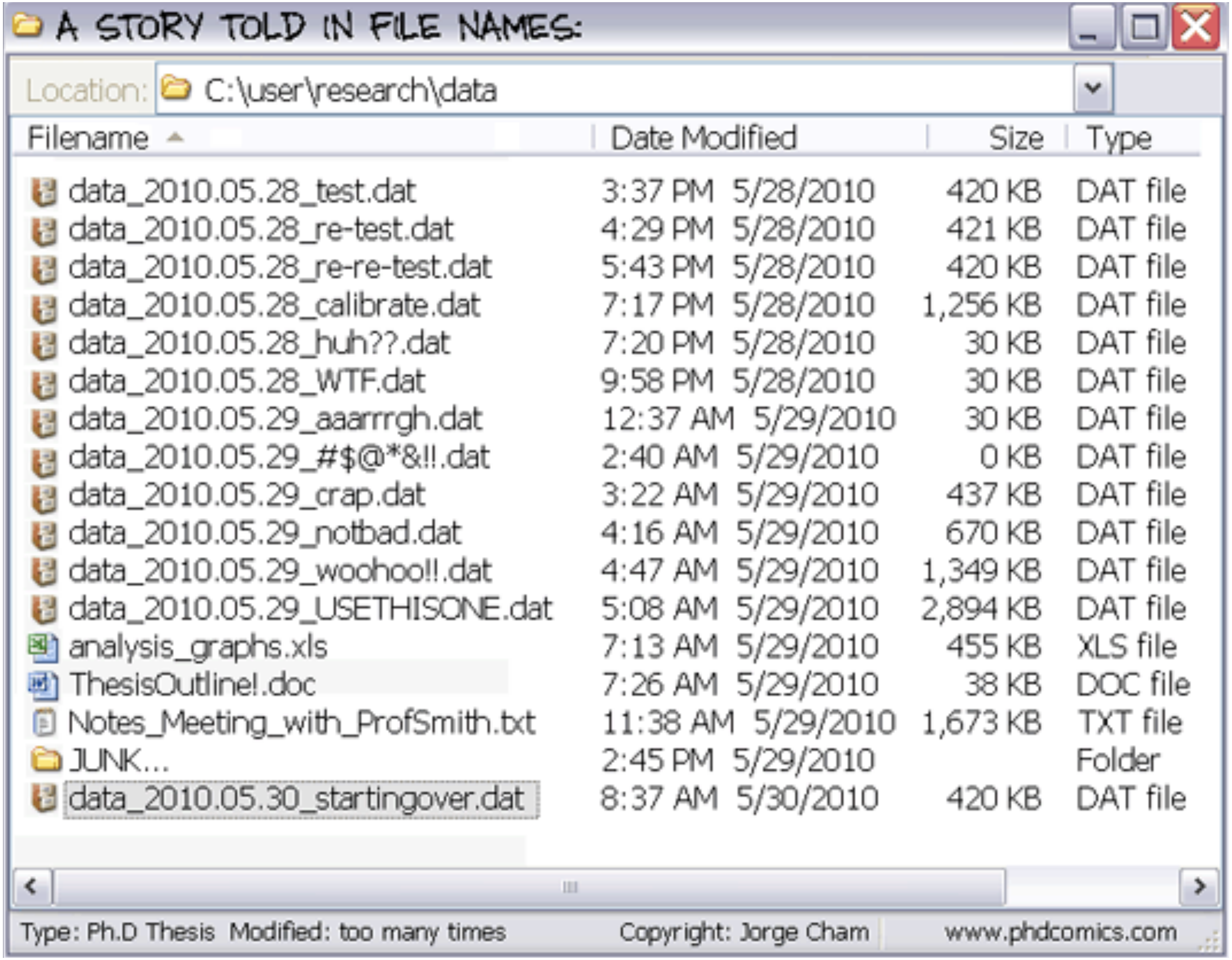
Git & github

A (very) short intro
by: @spsaaibi

References:
https://github.com/hadley

Overview

- Focus mostly on the why, not the how
- Learning git & github is definitely frustrating at first
- Two big pay offs: **safety** & **community**



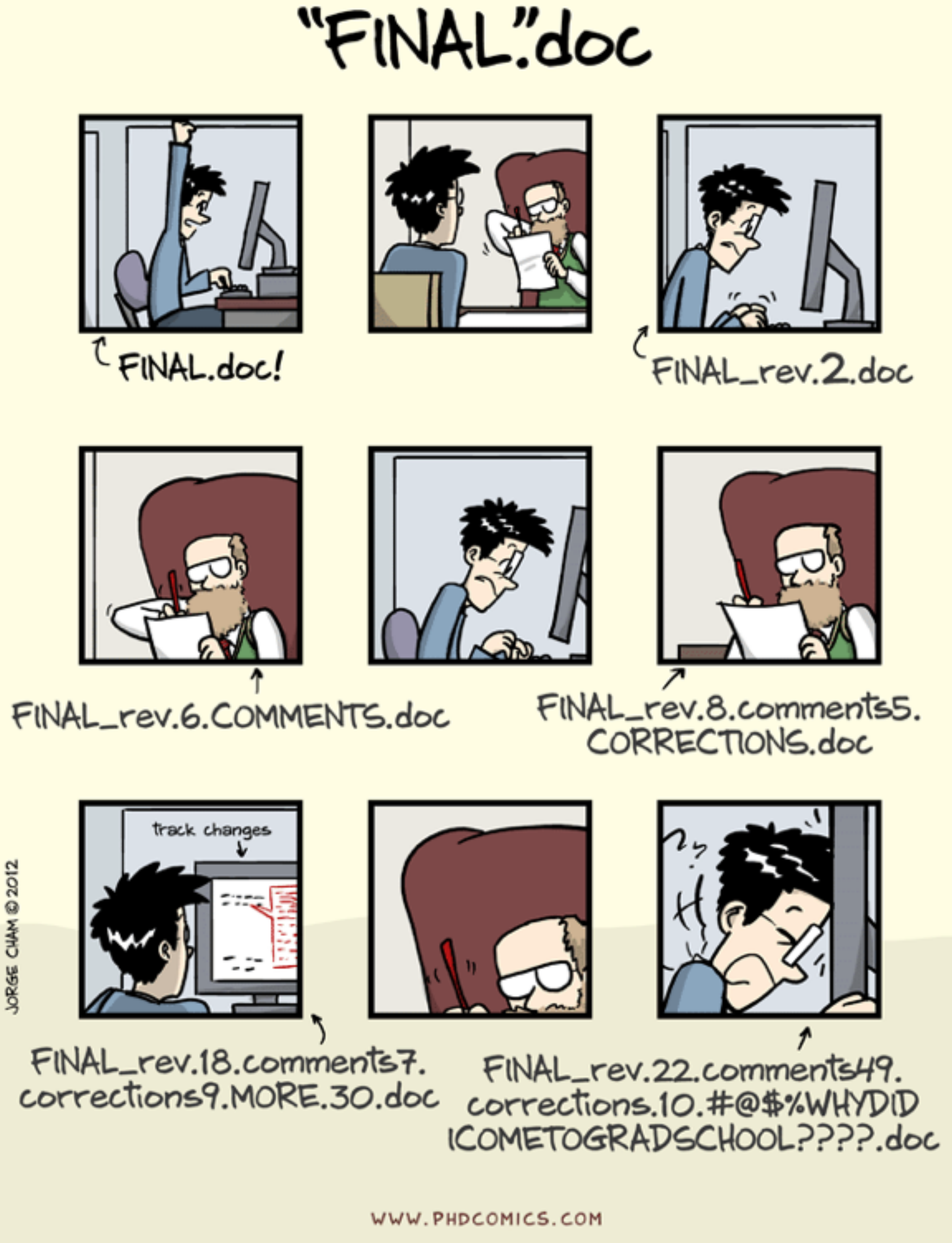
A STORY TOLD IN FILE NAMES:

Location: C:\user\research\data

Filename	Date Modified	Size	Type
data_2010.05.28_test.dat	3:37 PM 5/28/2010	420 KB	DAT file
data_2010.05.28_re-test.dat	4:29 PM 5/28/2010	421 KB	DAT file
data_2010.05.28_re-re-test.dat	5:43 PM 5/28/2010	420 KB	DAT file
data_2010.05.28_calibrate.dat	7:17 PM 5/28/2010	1,256 KB	DAT file
data_2010.05.28_huh??.dat	7:20 PM 5/28/2010	30 KB	DAT file
data_2010.05.28_WTF.dat	9:58 PM 5/28/2010	30 KB	DAT file
data_2010.05.29_aaarrrgh.dat	12:37 AM 5/29/2010	30 KB	DAT file
data_2010.05.29_#$@*&!!.dat	2:40 AM 5/29/2010	0 KB	DAT file
data_2010.05.29_crap.dat	3:22 AM 5/29/2010	437 KB	DAT file
data_2010.05.29_notbad.dat	4:16 AM 5/29/2010	670 KB	DAT file
data_2010.05.29_woohoo!!.dat	4:47 AM 5/29/2010	1,349 KB	DAT file
data_2010.05.29_USETHISONE.dat	5:08 AM 5/29/2010	2,894 KB	DAT file
analysis_graphs.xls	7:13 AM 5/29/2010	455 KB	XLS file
ThesisOutline!.doc	7:26 AM 5/29/2010	38 KB	DOC file
Notes_Meeting_with_ProfSmith.txt	11:38 AM 5/29/2010	1,673 KB	TXT file
JUNK...	2:45 PM 5/29/2010		Folder
data_2010.05.30_startingover.dat	8:37 AM 5/30/2010	420 KB	DAT file

Type: Ph.D Thesis Modified: too many times Copyright: Jorge Cham www.phdcomics.com

http://www.phdcomics.com/comics/archive.php?comicid=1323


"FINAL".doc
FINAL.doc!
FINAL_rev.2.doc
FINAL_rev.6.COMMENTS.doc
FINAL_rev.8.comments5.CORRECTIONS.doc
track changes
FINAL_rev.18.comments7.corrections9.MORE.30.doc
FINAL_rev.22.comments49.corrections.10.#@$%WHYDIDICOMETOGRADSCHOOL????.doc
JORGE CHAM © 2012
WWW.PHDCOMICS.COM

Changes

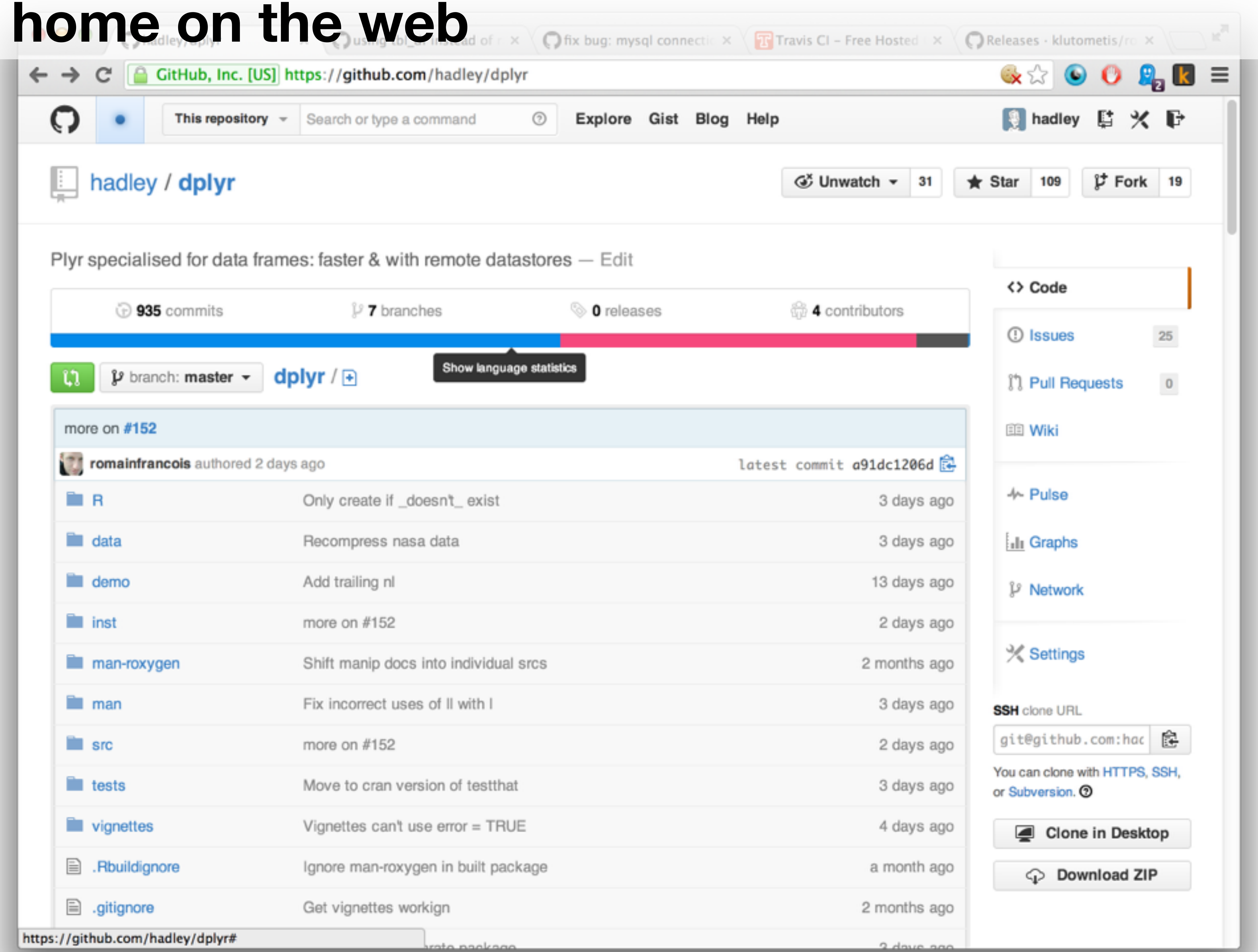
- **A unique identifier**
- What changed?
- **When did it change?**
- Who changed it?
- **Why did it change?**

(And this obviously gets hard to coordinate for multiple files)

Recipes (*basics*):

```
git pull #get the latest version of course repo
git add . #add my changes to commit
git commit -am ‘msg’ #insightful msg about my changes’
git push # push changes to repo
```

A home on the web



GitHub, Inc. [US] https://github.com/hadley/dplyr

hadley / dplyr

Plyr specialised for data frames: faster & with remote datastores — Edit

935 commits · 7 branches · 0 releases · 4 contributors

more on #152

romainfrancois authored 2 days ago — latest commit a91dc1206d

R	Only create if _doesn't_ exist	3 days ago
data	Recompress nasa data	3 days ago
demo	Add trailing nl	13 days ago
inst	more on #152	2 days ago
man-roxygen	Shift manip docs into individual srcs	2 months ago
man	Fix incorrect uses of ll with l	3 days ago
src	more on #152	2 days ago
tests	Move to cran version of testthat	3 days ago
vignettes	Vignettes can't use error = TRUE	4 days ago
.Rbuildignore	Ignore man-roxygen in built package	a month ago
.gitignore	Get vignettes workign	2 months ago

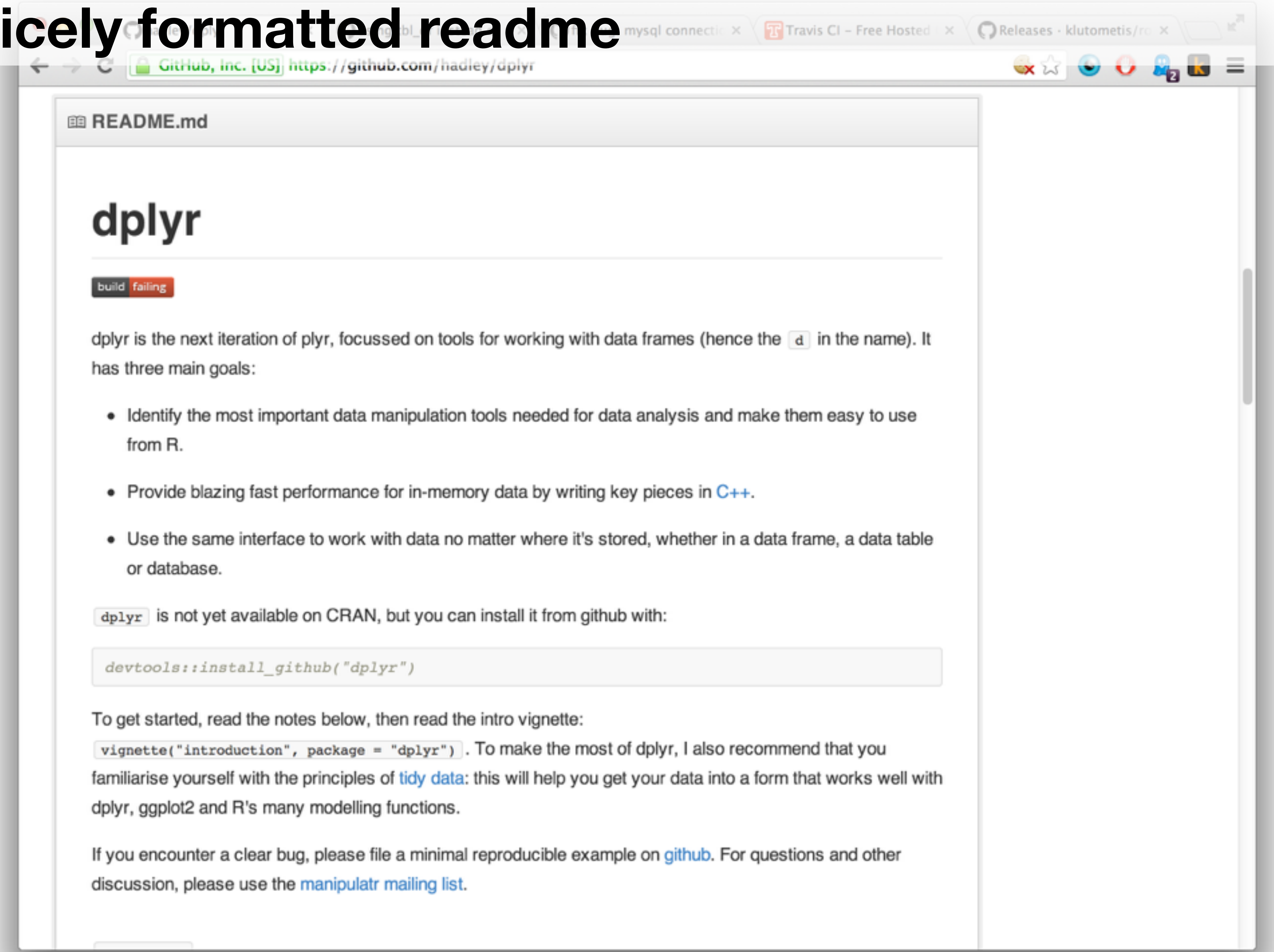
SSH clone URL

git@github.com:hac

You can clone with HTTPS, SSH, or Subversion.

https://github.com/hadley/dplyr#

Nicely formatted readme



dplyr

dplyr is the next iteration of plyr, focussed on tools for working with data frames (hence the `d` in the name). It has three main goals:

- Identify the most important data manipulation tools needed for data analysis and make them easy to use from R.
- Provide blazing fast performance for in-memory data by writing key pieces in C++.
- Use the same interface to work with data no matter where it's stored, whether in a data frame, a data table or database.

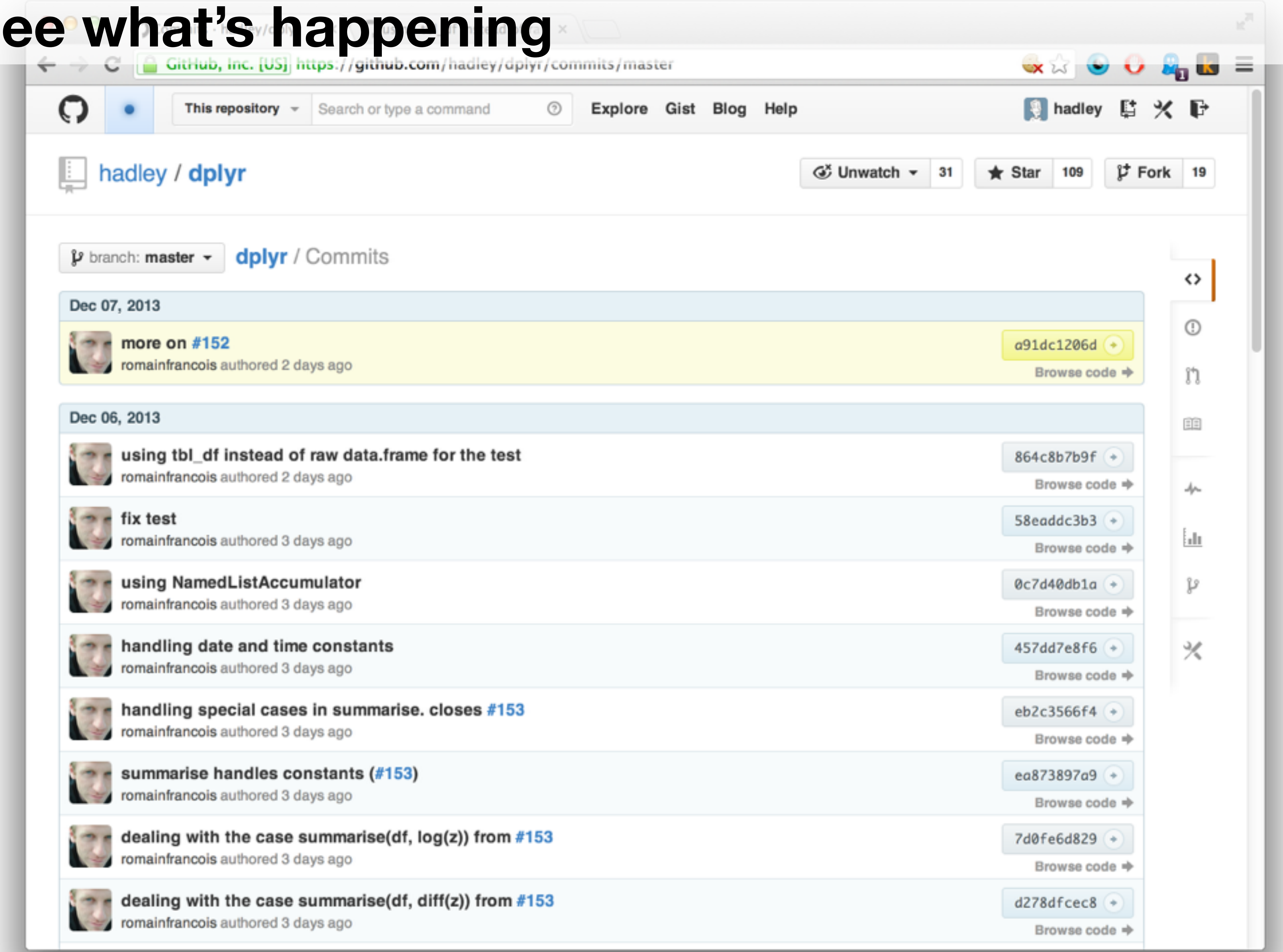
`dplyr` is not yet available on CRAN, but you can install it from github with:

```
devtools::install_github("dplyr")
```

To get started, read the notes below, then read the intro vignette: `vignette("introduction", package = "dplyr")`. To make the most of dplyr, I also recommend that you familiarise yourself with the principles of tidy data: this will help you get your data into a form that works well with dplyr, ggplot2 and R's many modelling functions.

If you encounter a clear bug, please file a minimal reproducible example on github. For questions and other discussion, please use the manipulatr mailing list.

See what’s happening



hadley / dplyr

dplyr / Commits

Dec 07, 2013

more on #152
romainfrancois authored 2 days ago — a91dc1206d

Dec 06, 2013

using tbl_df instead of raw data.frame for the test
romainfrancois authored 2 days ago — 864c8b7b9f

fix test
romainfrancois authored 3 days ago — 58eaddc3b3

using NamedListAccumulator
romainfrancois authored 3 days ago — 0c7d40db1a

handling date and time constants
romainfrancois authored 3 days ago — 457dd7e8f6

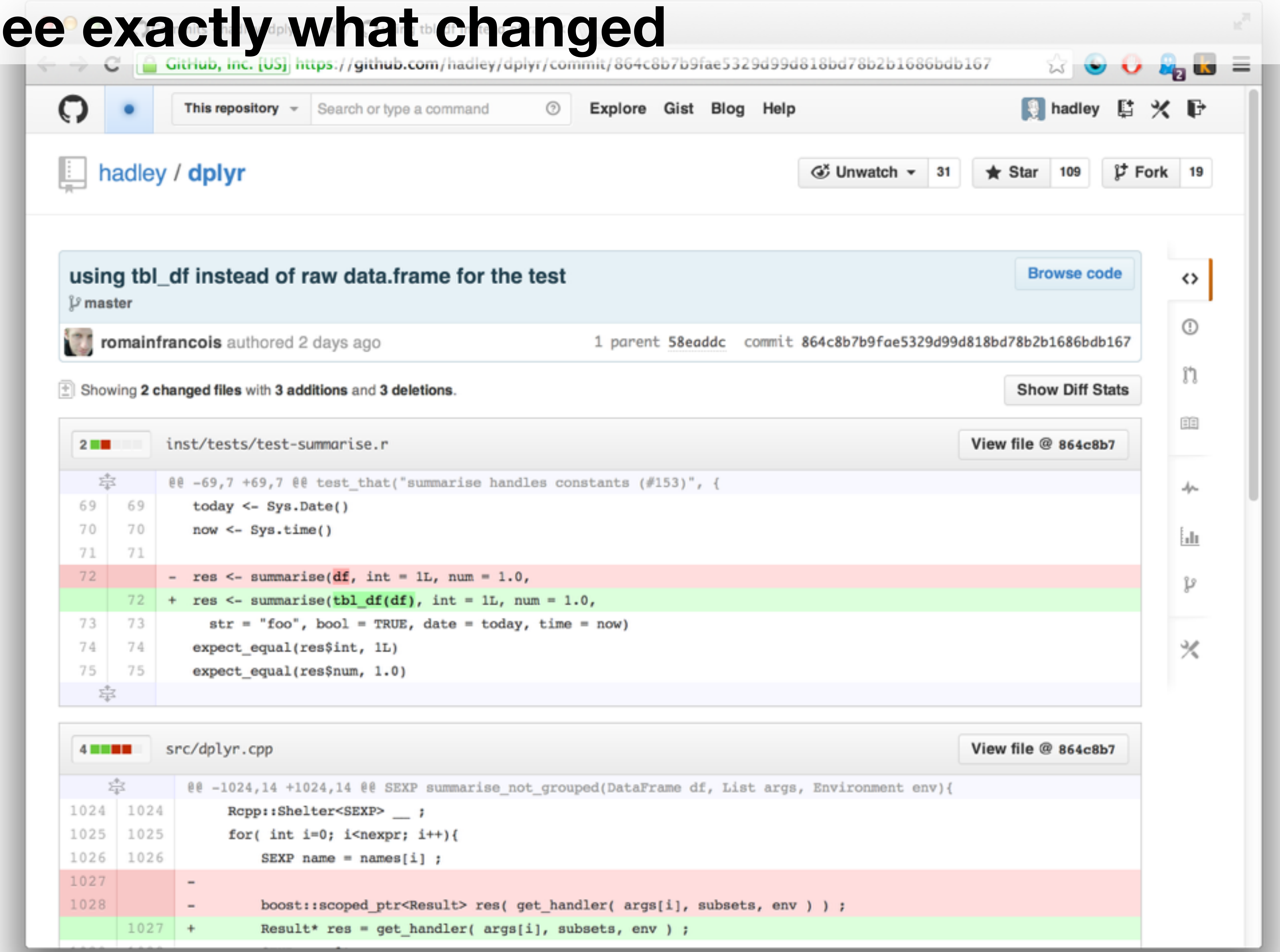
handling special cases in summarise. closes #153
romainfrancois authored 3 days ago — eb2c3566f4

summarise handles constants (#153)
romainfrancois authored 3 days ago — ea873897a9

dealing with the case summarise(df, log(z)) from #153
romainfrancois authored 3 days ago — 7d0fe6d829

dealing with the case summarise(df, diff(z)) from #153
romainfrancois authored 3 days ago — d278dfcec8

See exactly what changed



hadley / dplyr

using tbl_df instead of raw data.frame for the test

master

romainfrancois authored 2 days ago 1 parent 58eaddc commit 864c8b7b9fae5329d99d818bd78b2b1686bdb167

Showing **2 changed files** with **3 additions** and **3 deletions**.

2 inst/tests/test-summarise.r

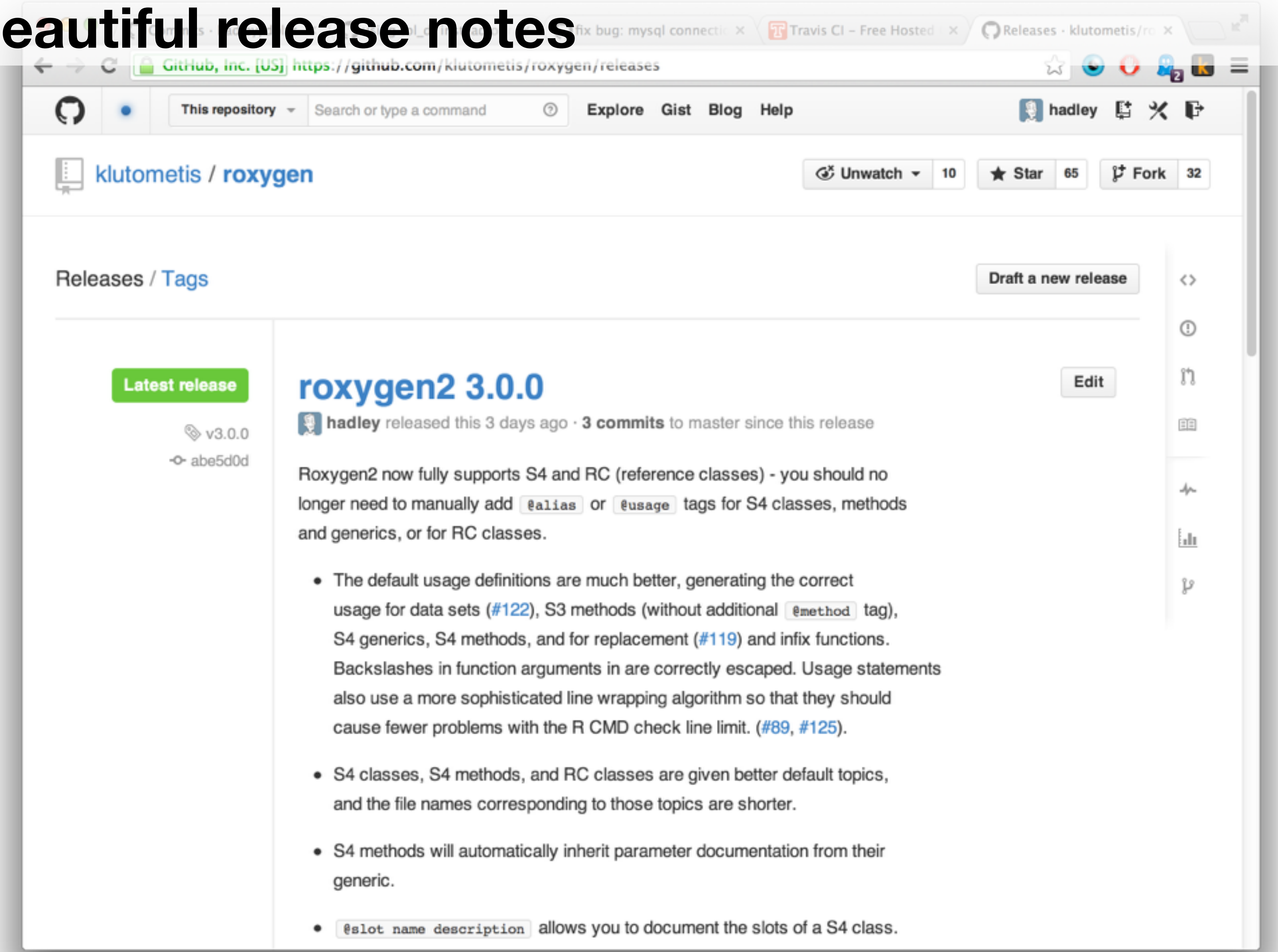
```
@@ -69,7 +69,7 @@ test_that("summarise handles constants (#153)", {
   today <- Sys.Date()
   now <- Sys.time()

-  res <- summarise(df, int = 1L, num = 1.0,
+  res <- summarise(tbl_df(df), int = 1L, num = 1.0,
     str = "foo", bool = TRUE, date = today, time = now)
   expect_equal(res$int, 1L)
   expect_equal(res$num, 1.0)
```

4 src/dplyr.cpp

```
@@ -1024,14 +1024,14 @@ SEXP summarise_not_grouped(DataFrame df, List args, Environment env){
     Rcpp::Shelter<SEXP> __ ;
     for( int i=0; i<nexpr; i++){
         SEXP name = names[i] ;
-
-        boost::scoped_ptr<Result> res( get_handler( args[i], subsets, env ) ) ;
+        Result* res = get_handler( args[i], subsets, env ) ;
```

Beautiful release notes



roxygen2 3.0.0

hadley released this 3 days ago · **3 commits** to master since this release

Roxygen2 now fully supports S4 and RC (reference classes) - you should no longer need to manually add `@alias` or `@usage` tags for S4 classes, methods and generics, or for RC classes.

- The default usage definitions are much better, generating the correct usage for data sets (#122), S3 methods (without additional `@method` tag), S4 generics, S4 methods, and for replacement (#119) and infix functions. Backslashes in function arguments in are correctly escaped. Usage statements also use a more sophisticated line wrapping algorithm so that they should cause fewer problems with the R CMD check line limit. (#89, #125).
- S4 classes, S4 methods, and RC classes are given better default topics, and the file names corresponding to those topics are shorter.
- S4 methods will automatically inherit parameter documentation from their generic.
- `@slot name description` allows you to document the slots of a S4 class.

Learning Git & Github

- Is frustrating at first & there’s a lot of new vocab to learn.
- Pays off by increasing the pace of your coding because you don't have to worry as much about mistakes.
- Pays off by connecting you with a community of like minded people who can help make your code better.