

How R Markdown + Git + GitHub changed my (teaching?) life

R Summit & Workshop 2015

Dr. Jennifer (Jenny) Bryan

Dept. of Statistics & Michael Smith Laboratories, UBC

jenny@stat.ubc.ca

<http://stat545-ubc.github.io>

<http://www.stat.ubc.ca/~jenny/>

 [@JennyBryan](https://twitter.com/JennyBryan)

 [@STAT545](https://twitter.com/STAT545)

 [@jennybc](https://github.com/jennybc)

GitHub repo for this talk
gives a bunch of links + PDF of slides

https://github.com/jennybc/2015-06-28_r-summit-talk

Data wrangling, exploration, and analysis with R

UBC STAT 545A and 547M

Learn how to

- explore, groom, visualize, and analyze data
- make all of that reproducible, reusable, and shareable
- using R

<http://stat545-ubc.github.io>

<https://github.com/STAT545-UBC/STAT545-UBC.github.io>

MOST YEARS OF R
USE BEFORE
DEVELOPING AND
DISTRIBUTING A
PACKAGE



What's so great about
(R) Markdown + Git(Hub)?

**weak links in the chain:
process, packaging and
presentation**



R + markdown

Do your work

Get a presentable, web-friendly version for free

Present-ability is BAKED IN

... not a separate process you never get around to

stuff you
need to
write



stuff people
like to
read

stuff you
need to
write



stuff people
like to
read



stuff you
need to
write



stuff people
like to
read

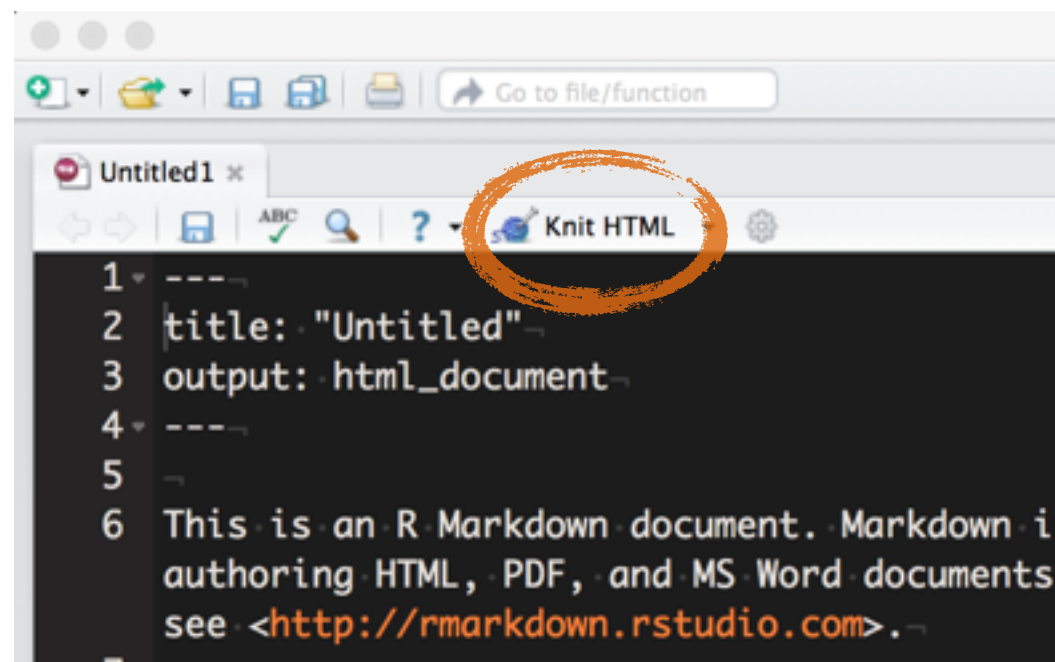
foo.R
foo.Rmd



foo.md
foo.html

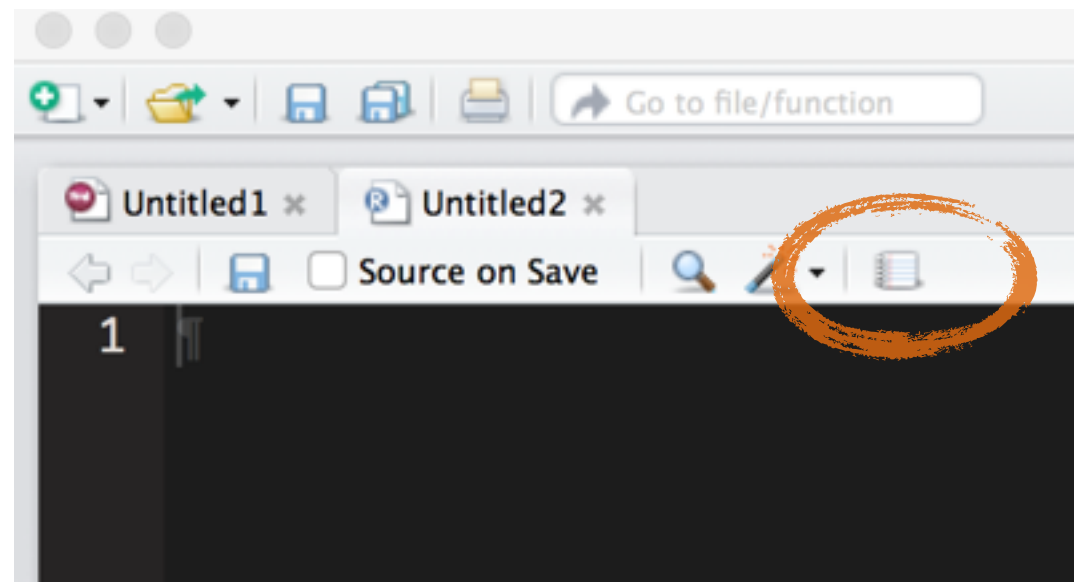


```
library(rmarkdown)  
render("foo.Rmd")
```





```
library(rmarkdown)  
render("foo.R")
```



foo.Rmd → foo.html

```
---
title: "Untitled"
output: html_document
---
```

foo.Rmd → foo.md → foo.html

```
---
title: "Untitled"
output:
  html_document:
    keep_md: yes
---
```

foo.Rmd → foo.md

```
---
output:
  md_document:
    variant: markdown_github
---
```

foo.R → foo.html

```
#' —  
#′ title: "Untitled"  
#′ output: html_document  
#′ ---
```

foo.R → foo.md → foo.html

```
#′ ---  
#′ title: "Untitled"  
#′ output:  
#′   html_document:  
#′     keep_md: yes  
#′ ---
```

foo.R → foo.md

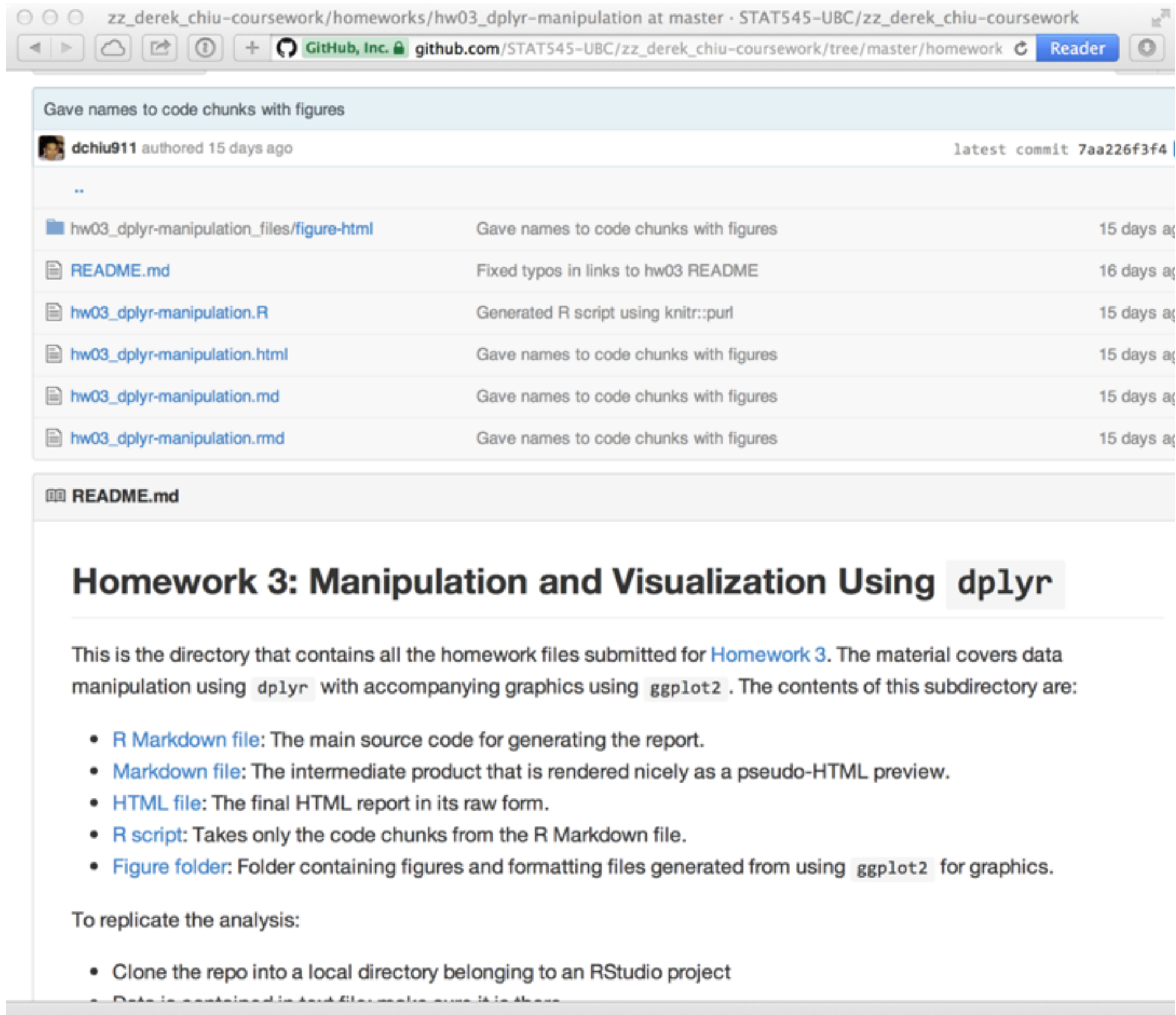
```
#′ ---  
#′ output:  
#′   md_document:  
#′     variant: markdown_github  
#′ ---
```

Markdown, and even Rmarkdown,
are very privileged on GitHub

Essentially rendered as HTML



When I mark homework ... this is what I see.



The screenshot shows a web browser window displaying a GitHub repository. The address bar shows the URL: `github.com/STAT545-UBC/zz_derek_chiu-coursework/tree/master/homework`. The repository name is `zz_derek_chiu-coursework/homeworks/hw03_dplyr-manipulation` at master. The commit history shows a commit by `dchiu911` 15 days ago with the message "Gave names to code chunks with figures". The commit hash is `7aa226f3f4`. The file list includes:

- `hw03_dplyr-manipulation_files/figure-html` (15 days ago)
- `README.md` (16 days ago)
- `hw03_dplyr-manipulation.R` (15 days ago)
- `hw03_dplyr-manipulation.html` (15 days ago)
- `hw03_dplyr-manipulation.md` (15 days ago)
- `hw03_dplyr-manipulation.rmd` (15 days ago)

The `README.md` file is selected, showing the following content:

Homework 3: Manipulation and Visualization Using `dplyr`

This is the directory that contains all the homework files submitted for [Homework 3](#). The material covers data manipulation using `dplyr` with accompanying graphics using `ggplot2`. The contents of this subdirectory are:

- [R Markdown file](#): The main source code for generating the report.
- [Markdown file](#): The intermediate product that is rendered nicely as a pseudo-HTML preview.
- [HTML file](#): The final HTML report in its raw form.
- [R script](#): Takes only the code chunks from the R Markdown file.
- [Figure folder](#): Folder containing figures and formatting files generated from using `ggplot2` for graphics.

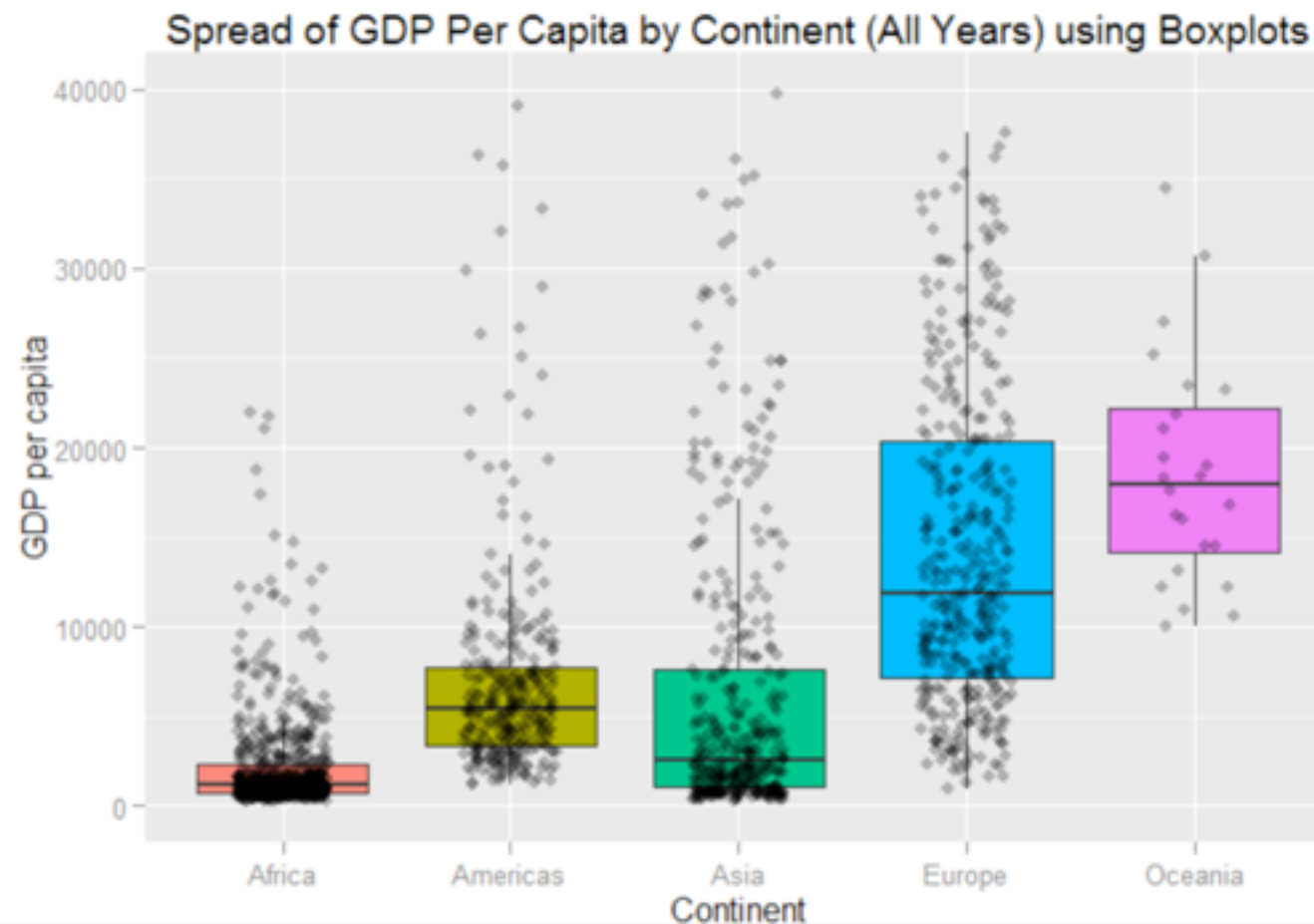
To replicate the analysis:

- Clone the repo into a local directory belonging to an RStudio project
- Data is contained in text files, make sure it is there

When I mark homework ... this is what I see.

In this section, we will use two similar graphs to visualize spread: the **boxplot** and the **violin plot**. Note that the original unmanipulated data frame `gtbl` will be used here.

```
ggplot(gtbl, aes(continent, gdpPercap))+  
  geom_boxplot(aes(fill = continent), outlier.shape = NA)+  
  geom_jitter(alpha = 0.3, position = position_jitter(width = 0.2))+  
  xlab("Continent")+  
  ylab("GDP per capita")+  
  ylim(c(0,40000))+  
  theme(legend.position = "none")+  
  ggtitle("Spread of GDP Per Capita by Continent (All Years) using Boxplots")
```



When I mark homework ... this is what I see.

STAT 545A

- [Homework 1](#): Edit README.md and experiment with Markdown
- [Homework 2](#): Exploring the Gapminder Dataset
- [Homework 3](#): Manipulation and Visualization Using `dplyr`
- [Homework 4](#): Writing and Testing Functions
- [Homework 5](#): Factor Control and File I/O
- [Homework 6](#): *Optional*: Transition Activities

STAT 547M

- [Homework 7](#): Data Wrangling Grand Finale
- [Homework 8](#): Data Cleaning
- [Homework 9](#): Automating Data Analysis Pipelines
- [Homework 10](#): Building an R Package
- [Homework 11](#): Building a Shiny App
- [Homework 12](#): Getting Data off the Web

What's so great about
(R) Markdown + Git(Hub)?

What's so great about Git(Hub)?

search

linky-ness

meta-stuff (comments, issues, ...)

Problem:

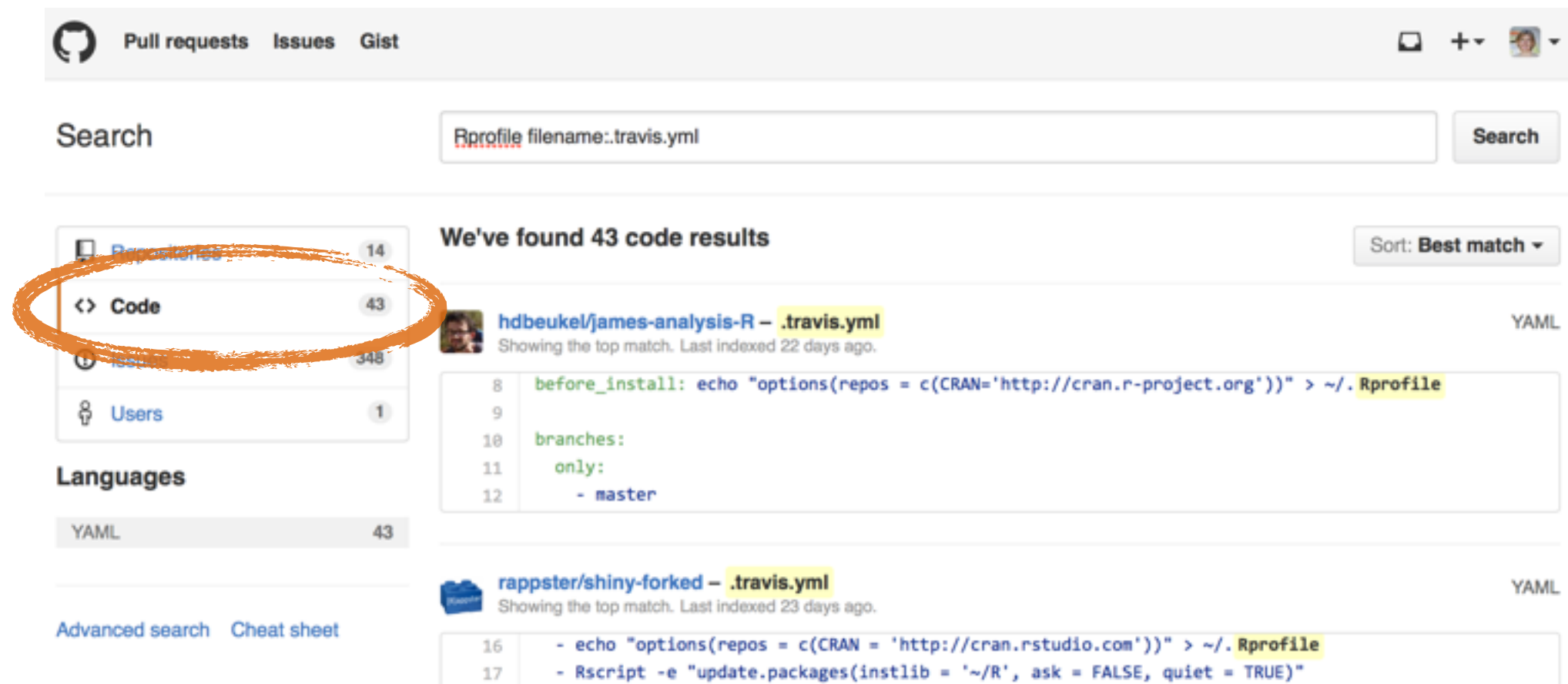
NOTE from R CMD check when testing a package on Travis ... but I *am* specifying my CRAN repo?

```
* checking package dependencies ... NOTE  
No repository set, so cyclic dependency check skipped
```

Hmm ... maybe the CRAN environment variable isn't consulted, maybe I need to set the CRAN repo set in `.Rprofile` ... how do I *that* on Travis?

Solution:

Find a ton of other `.travis.yml` files that mention `.Rprofile`!



GitHub interface showing search results for `Rprofile filename:.travis.yml`.

Search bar: `Rprofile filename:.travis.yml` Search

Filters:

- Repositories: 14
- Code: 43** (circled in orange)
- Issues: 348
- Users: 1

Languages: YAML 43

Advanced search Cheat sheet

We've found 43 code results Sort: Best match

Top result: **hdbeukel/james-analysis-R** — `.travis.yml` (YAML)
Showing the top match. Last indexed 22 days ago.

```
8 before_install: echo "options(repos = c(CRAN='http://cran.r-project.org'))" > ~/.Rprofile
9
10 branches:
11   only:
12     - master
```

Second result: **rappster/shiny-forked** — `.travis.yml` (YAML)
Showing the top match. Last indexed 23 days ago.

```
16 - echo "options(repos = c(CRAN = 'http://cran.rstudio.com'))" > ~/.Rprofile
17 - Rscript -e "update.packages(instlib = '~/R', ask = FALSE, quiet = TRUE)"
```



Jennifer Bryan @JennyBryan · May 29

We have `isTRUE` but not `isFALSE`? 😡 #rstats

RETWEET

1

FAVORITES

9



Replies included discussion of why we have `isTRUE ()` and the existence + history of various “truthy” things

Included links to where all this occurs in the R source and when it was added

```
120 }
121 #endif
122
123 const static char * const truenames[] = {
124     "T",
125     "True",
126     "TRUE",
127     "true",
128     (char *) NULL,
129 };
130
131 const static char * const falsenames[] = {
132     "F",
133     "False",
134     "FALSE",
135     "false",
136     (char *) NULL,
137 };
138
```

<https://github.com/wch/r-source/blob/trunk/src/main/util.c#L445-L452>

```
442
443 /* Function to test whether a string is a true value */
444
445 Rboolean StringTrue(const char *name)
446 {
447     int i;
448     for (i = 0; truenames[i]; i++)
449         if (!strcmp(name, truenames[i]))
450             return TRUE;
451     return FALSE;
452 }
453
454 Rboolean StringFalse(const char *name)
455 {
456     int i;
457     for (i = 0; falsenames[i]; i++)
458         if (!strcmp(name, falsenames[i]))
459             return TRUE;
460     return FALSE;
461 }
```

<https://github.com/wch/r-source/blob/trunk/src/main/util.c#L123-L129>

Updates for new list structure. [Browse files](#)

git-svn-id: <https://svn.r-project.org/R/trunk@1820> 00db46b3-68df-0310-9c12-caf00c1e9a41

ihaka authored on Jul 31, 1998 1 parent 94dadd6 commit 8fee9de69de30815457992145657797bf793b74f

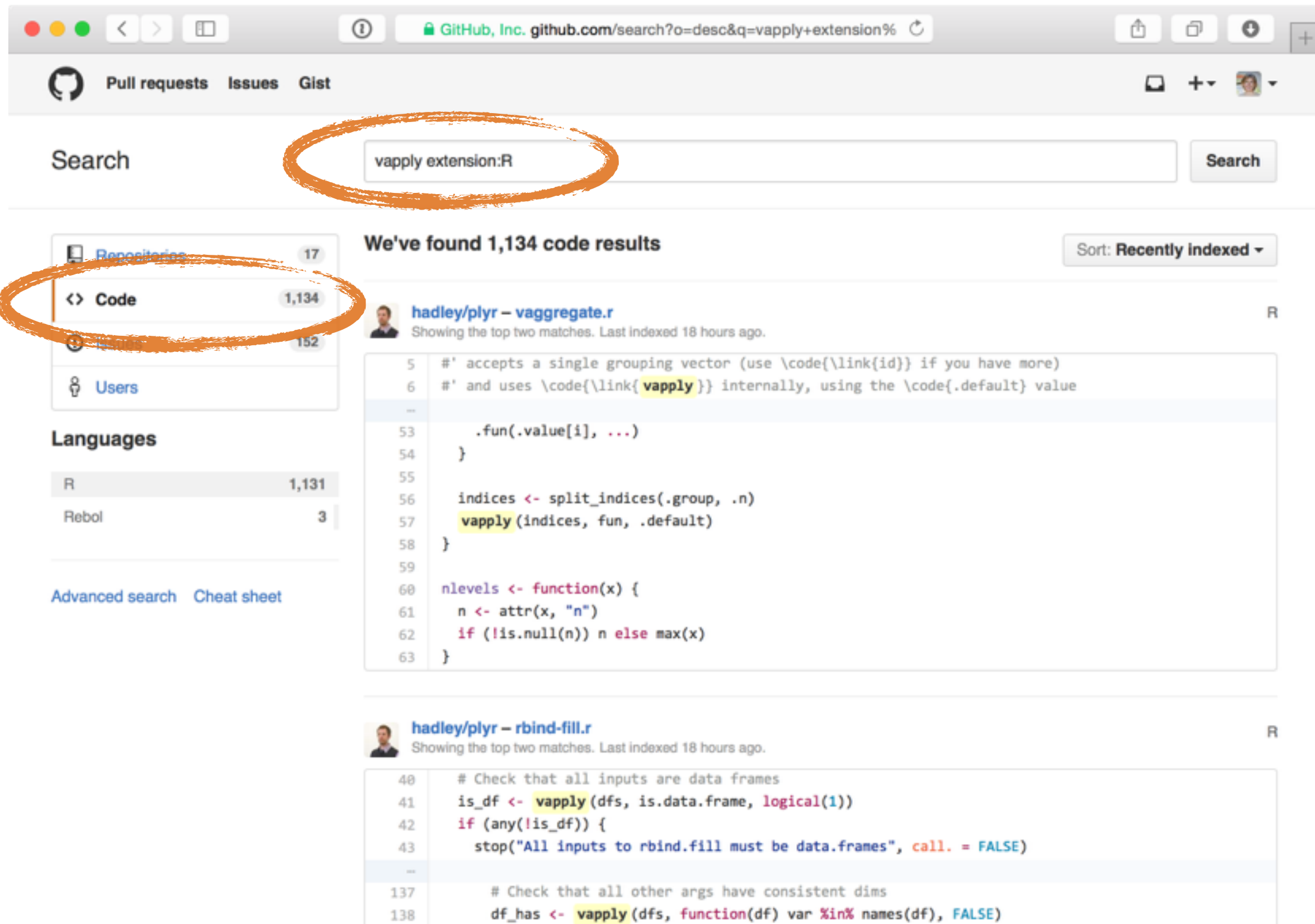
Showing 11 changed files with 5,842 additions and 5,074 deletions. [Unified](#) [Split](#)

```
63 +
64 +static char *truenames[] = {
65 +   "T",
66 +   "True",
67 +   "TRUE",
68 +   "true",
69 +   (char *) 0,
70 +};
71 +
72 +static char *falsenames[] = {
73 +   "F",
74 +   "False",
75 +   "FALSE",
76 +   "false",
77 +   (char *) 0,
78 +};
79
80 int asInteger(SEXP x)
81 {
82     if (isVector(x)) {
83         if (LENGTH(x) < 1)
84             return NA_INTEGER;
85         switch (TYPEOF(x)) {
86         case LGLSXP:
87             return (LOGICAL(x)[0] == NA_LOGICAL) ? NA_INTEGER : ((LOGICAL(x)[0]) != 0);
88         case INTSXP:
89             return (INTEGER(x)[0]);
90         case REALSXP:
91             return (REAL(x)[0]);
92         case CPLXSXP:
93             return (INTEGER(x)[0]);
94         case STRSXP:
95             return (INTEGER(x)[0]);
96         case SYMSXP:
97             return (INTEGER(x)[0]);
98         case ENVSXP:
99             return (INTEGER(x)[0]);
100        case BUILTINSXP:
101            return (INTEGER(x)[0]);
102        }
103    }
104    return NA_INTEGER;
```

<https://github.com/wch/r-source/commit/8fee9de69de30815457992145657797bf793b74f#diff-b98c76a66da58d25f4436ff77e8a2ac2R64>

Github's ease of search, navigation, & linking enriched this conversation

Are the official examples for a function kind of thin?
Search Github to see the function “in the wild”.



GitHub, Inc. github.com/search?o=desc&q=vapply+extension%

Pull requests Issues Gist

Search Search

We've found 1,134 code results Sort: Recently indexed

Code 1,134

Languages

Language	Count
R	1,131
Rebol	3

Advanced search Cheat sheet

hadley/plyr – vaggregate.r R

Showing the top two matches. Last indexed 18 hours ago.

```
5 #' accepts a single grouping vector (use \code{\link{id}} if you have more)
6 #' and uses \code{\link{vapply}} internally, using the \code{.default} value
...
53 .fun(.value[i], ...)
54 }
55
56 indices <- split_indices(.group, .n)
57 vapply(indices, fun, .default)
58 }
59
60 nlevels <- function(x) {
61   n <- attr(x, "n")
62   if (!is.null(n)) n else max(x)
63 }
```

hadley/plyr – rbind.fill.r R

Showing the top two matches. Last indexed 18 hours ago.

```
40 # Check that all inputs are data frames
41 is_df <- vapply(dfs, is.data.frame, logical(1))
42 if (any(!is_df)) {
43   stop("All inputs to rbind.fill must be data.frames", call. = FALSE)
...
137 # Check that all other args have consistent dims
138 df_has <- vapply(dfs, function(df) var %in% names(df), FALSE)
```

Are the official examples for a function kind of thin?
Search Github to see the function “in the wild”.

GitHub, Inc. github.com/search?q=vapply+user%3Acran&type=Code

Pull requests Issues Gist

Search Search

We've found 894 code results Sort: Best match

Code 894

Languages

R	866
RMarkdown	4
Text	3
Rebol	2
HTML	2
Org	1
Markdown	1
INI	1

cran/tumblr - oauth.encode.R R
Showing the top match. Last indexed 25 days ago.

```
1  oauth.encode <-  
2  function(x) vapply(x, oauth.encode1, character(1))
```

cran/tumblr - compact.R R
Showing the top match. Last indexed 25 days ago.

```
1  compact <-  
2  function(x) {  
3      null <- vapply(x, is.null, logical(1))  
4      x[!null]  
5  }
```

cran/sdprisk - almost.equal.R R
Showing the top match. Last indexed 27 days ago.

```
1  almost.equal <- function(x, value) {  
2      vapply(Map(all.equal, x, rep.int(value, length(x))), isTRUE, logical(1))  
3  }
```

[Advanced search](#) [Cheat sheet](#)



STAT 545 @ University of British Columbia

Course in data wrangling, exploration, and analysis with R

Vancouver, BC

Filters ▾

Find a repository...

+ New repository

STAT545-UBC.github.io

HTML ★ 47 📄 39

Main repository for STAT 545 @ University of British Columbia, a course in data wrangling, exploration, and analysis with R.

Updated 4 days ago

InstructorsOnly PRIVATE

R ★ 0 📄 0

Private repository for STAT 545 instructors

Updated on Jan 25

Discussion

★ 3 📄 1

Public discussion

Updated on Sep 17, 2014

zz_derek_chiu-coursework PRIVATE

HTML ★ 0 📄 1

coursework created for zz_derek_chiu

Updated on Mar 25

← Github Organization

← Content, website

← Shh... secret

← Discussion forum

← 1 repo per student

Issues as discussion forum

Issues · STAT545-UBC/Discussion

github.com/STAT545-UBC/Discussion/issues

STAT545-UBC / Discussion

Unwatch 23 Star 1 Fork 1

Issues Pull requests Labels Milestones Filters is:issue is:open New Issue

19 Open 64 Closed

Author	Labels	Milestones	Assignee	Sort
<input type="checkbox"/> Future plans for your STAT 545 / 547 coursework repository				3
#83 opened on Dec 18, 2014 by jennybc				
<input type="checkbox"/> Worked example: when two data.frames are almost, but not quite, the same				1
#82 opened on Dec 16, 2014 by jennybc				
<input type="checkbox"/> Using `rplots` in RMarkdown without including API key inside RMarkdown?				5
#81 opened on Dec 5, 2014 by spencerfrei				
<input type="checkbox"/> Remixing data -- `dplyr` bug				0
#80 opened on Dec 3, 2014 by aammd				
<input type="checkbox"/> Small detail: Indented Rstudio code not displaying nicely with github's tab setting of 8 spaces	Mac OS Windows			0
#78 opened on Nov 26, 2014 by jooolia				
<input type="checkbox"/> Deploying our shiny apps to the UBC stats server				2
#76 opened on Nov 25, 2014 by daattali				

Issues to submit, peer review, and mark homework

The screenshot shows a web browser window displaying the GitHub interface for the repository 'STAT545-UBC / zz_derek_chiu-coursework'. The browser's address bar shows the URL: `github.com/STAT545-UBC/zz_derek_chiu-coursework/issues?q=is%3Aissue+is%3Aclosed`. The repository is marked as 'PRIVATE'. The user 'jennybc' is logged in. The 'Issues' tab is selected, and the search filter 'is:issue is:closed' is applied. The issue list shows 0 Open and 33 Closed issues. The visible issues are:

Issue Title	Author	Labels	Comments
Mark homework 12 of Derek-Chiu	dchiu911		2
Peer review of derek_chiu's hw11 by beryl_zhuang	jennybc	hw11, peer-review	2
Peer review of derek_chiu's hw11 by abrar_wafa	jennybc	hw11, peer-review	2
Mark homework 11 of Derek-Chiu	dchiu911		2
Peer review of derek_chiu's hw10 by omar_alomeir	jennybc	hw10, peer-review	3

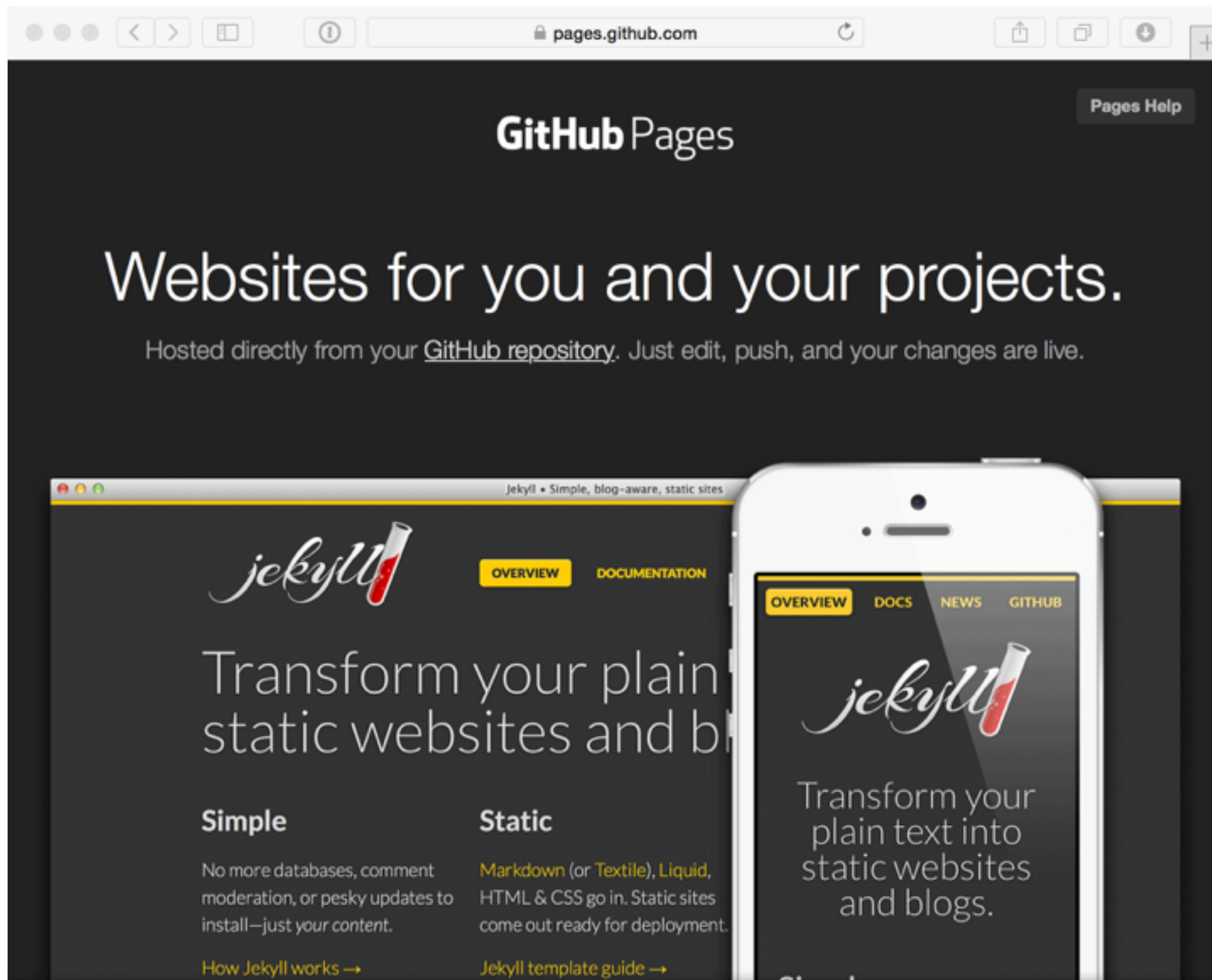
The unreasonable effectiveness of GitHub browsability



Commit .md files!
They are not like .o files or executables.
Get over it.

When (intermediate) Markdown is no longer enough ...

`<my_thing>.github.io`



reproducible
presentable

search

linky-ness

meta-stuff, e.g. issues

R Markdown v2

+



=



collaboration: “link, don’t attach”
machine- and human-readable
meta self-documentation