

New tools and workflows for data analysis

Dr. Jennifer (Jenny) Bryan

Dept. of Statistics & Michael Smith Laboratories, UBC

Workshop on Visualization for Big Data @ Fields Institute

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)

<https://github.com/jennybc>



in a wide array of academic fields,

the **ability to effectively process data**

is superseding other more classical modes of research

I'm Scott Olechowski, Co-Founder of Plex, and This Is How I Work

lifehacker.com/im-scott-olechowski-co-founder-of-plex-and-this-is-ho-1685226003

Reader

DISCOVER MORE

LOG IN

SIGN UP

TRENDING ON KINJA

1.  Scout On Jameis Winston: "How Could You Let That Guy In The Building?"
on Deadspin
2. Everything You Need to Have Seen From Last Night's Oscars
on Defamer
3. Patricia Arquette: Time for Gays, People of Color to Stand Up for Women
on Jezebel

lifehacker

+ FOLLOW

I'm Scott Olechowski, Co-Founder of Plex, and This Is How I Work



Andy Orin

Filed to: HOW I WORK 2/11/15 2:00pm

34,435 ⌂ 5 ★ ▾



spirit of my talk!

Fact: I don't work for these companies. I don't represent them. I am not an author of these packages.



R Markdown v2



links, files, etc. available here

The screenshot shows a GitHub repository page. The URL in the address bar is https://github.com/jennybc/2015-02-23_bryan-fields-talk. The repository name is **jennybc / 2015-02-23_bryan-fields-talk**. The repository has 3 commits, 1 branch, 0 releases, and 1 contributor. The master branch is selected. The repository description is: "Talk at Workshop on Visualization for Big Data: Strategies and Principles, Fields Institute". A link to the workshop website is provided: <http://www.fields.utoronto.ca/programs/scientific/14-15/bigdata/visualization/>. The repository contains files: .gitignore, 2015-02-23_bryan-fields-talk.Rproj, README.md, and README.md. The latest commit was made a minute ago by jennybc. The repository sidebar includes links for Code, Issues (0), Pull Requests (0), Wiki, Pulse, Graphs, and Settings. A clone URL is also provided: https://github.com/jennybc/2015-02-23_bryan-fields-talk.

https://github.com/jennybc/2015-02-23_bryan-fields-talk

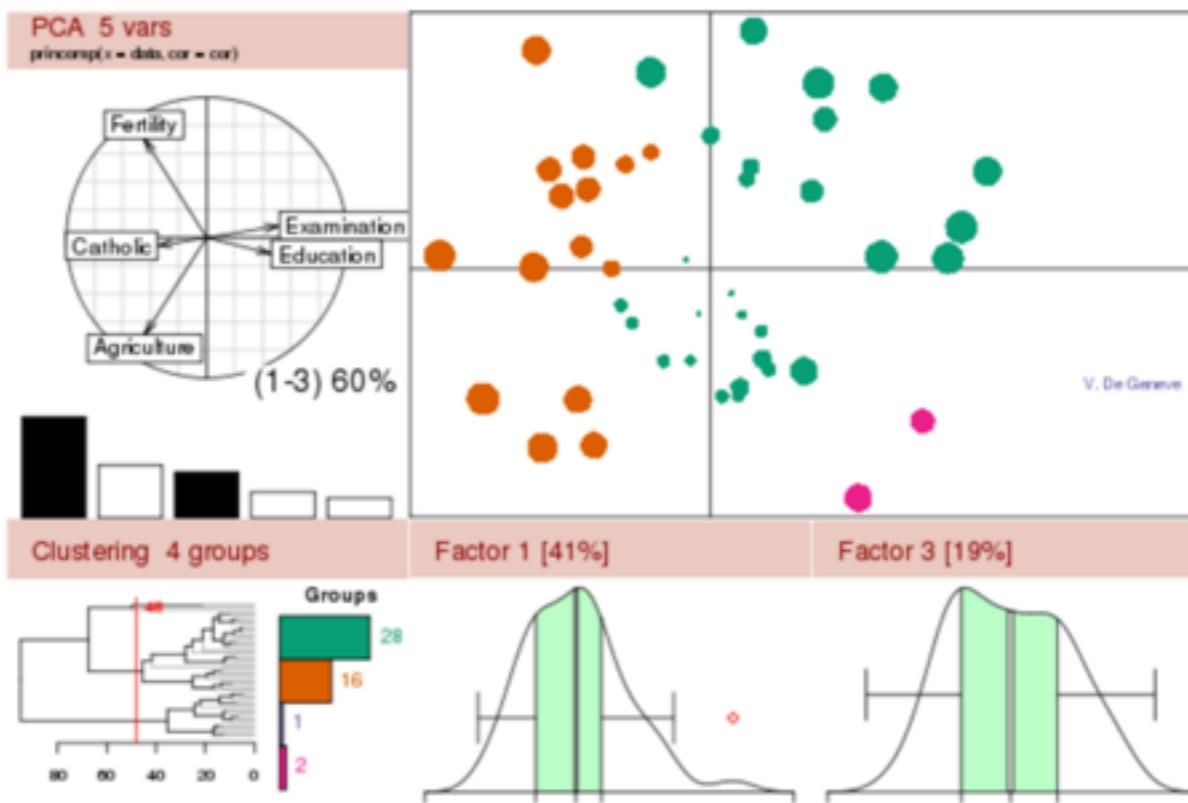


[About R](#)
[What is R?](#)
[Contributors](#)
[Screenshots](#)
[What's new?](#)

[Download, Packages](#)
[CRAN](#)

[R Project](#)
[Foundation](#)
[Members & Donors](#)
[Mailing Lists](#)
[Bug Tracking](#)
[Developer Page](#)
[Conferences](#)
[Search](#)

[Documentation](#)
[Manuals](#)
[FAQs](#)
[The R Journal](#)
[Wiki](#)
[Books](#)
[Certification](#)
[Other](#)



Getting Started:

- R is a free software environment for statistical computing and graphics. It compiles and runs on a wide variety of UNIX platforms, Windows and MacOS. To [download R](#), please choose your preferred [CRAN mirror](#).
- If you have questions about R like how to download and install the software, or what the license terms are, please read our [answers to frequently asked questions](#) before you send an email.

Douglas Bates
John Chambers
Peter Dalgaard
Seth Falcon
Robert Gentleman
Kurt Hornik
Ross Ihaka

Michael Lawrence
Friedrich Leisch
Uwe Ligges
Thomas Lumley
Martin Maechler
Martin Morgan
Duncan Murdoch

Paul Murrell
Martyn Plummer
Brian Ripley
Deepayan Sarkar
Duncan Temple Lang
Luke Tierney
Simon Urbanek



NSERC

www.nserc-crsng.gc.ca

| [Contact Us](#) | [Help](#) |

[eConsole](#) >

Main Menu

[Logout](#)

Proactive Disclosure

[Proactive Disclosure](#)

eConsole

Version 5.52

Welcome Jennifer Bryan

Users of the eSubmission system will no longer be required to periodically change their passwords. However, for their own protection, users are encouraged to change their passwords regularly. Keep your password safe and confidential; do not divulge it to anyone. NSERC will not be held liable for any loss of your data should you neglect to protect your password.

Account Management

[Change Password](#)
[Maintain User Profile](#)

Forms Management

[Forms - Researcher](#)
[Forms - Student](#)
[Forms - Reviewer](#)
[Forms - Partners](#)
[Forms - Department Head](#)

Amazon.com: Online Shopping for Electronics, Apparel, Computers, Books, DVDs & more

Amazon.ca | Reader |

Shopping from Canada? Magazinez-vous depuis le Canada? Visit Visitez amazon.ca >

JENNIFER's Amazon.com Today's Deals Gift Cards Sell Help Tax Central Sponsored by

Shop by Department ▾ Search All Go Hello, JENNIFER Your Account ▾ Your Prime ▾ Cart ▾

New For You See more

real
world
data



statistical
theory

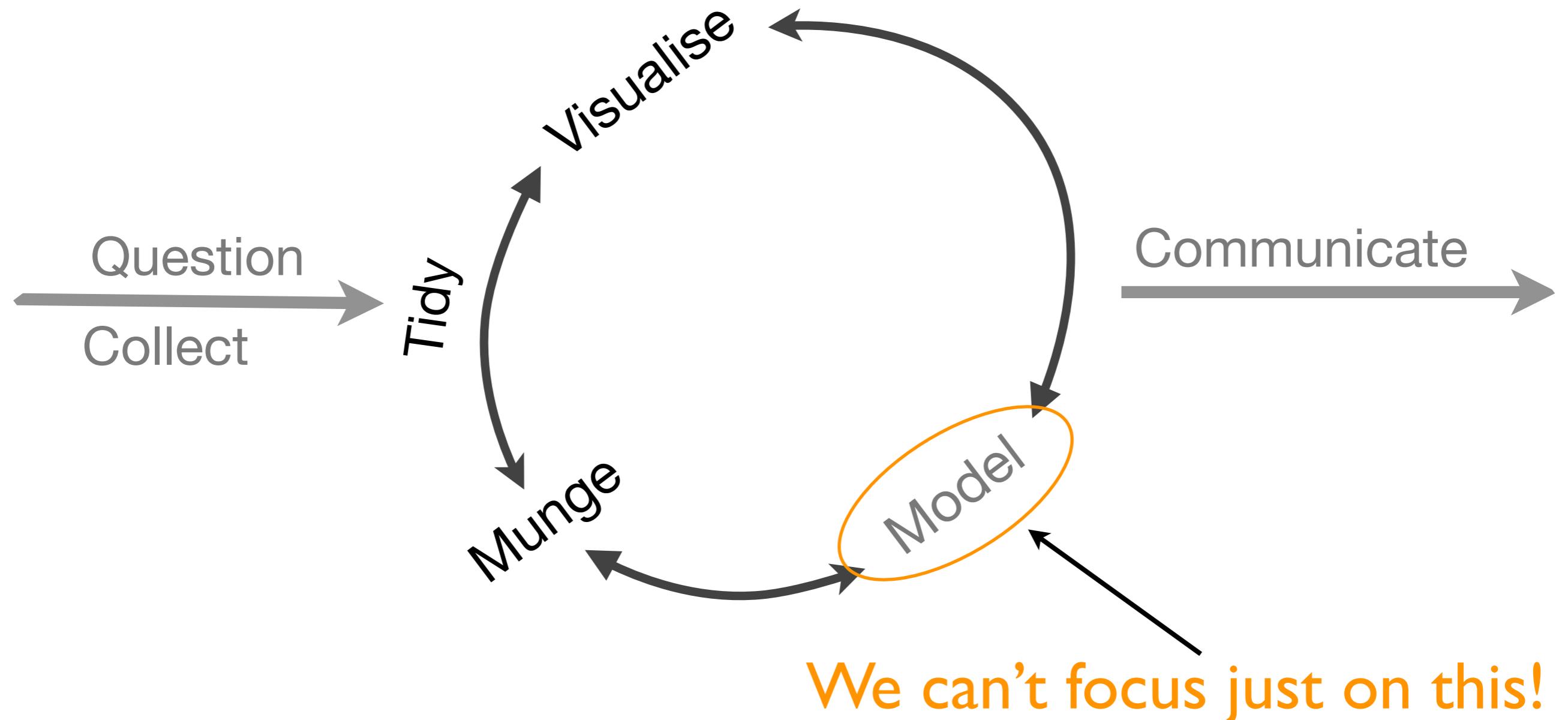
Data wrangling, exploration, and analysis with R

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>



For Big-Data Scientists, ‘Janitor Work’ Is Key Hurdle to Insights

By STEVE LOHR AUG. 17, 2014



Monica Rogati, Jawbone's vice president for data science, with Brian Wilt, a senior data scientist.
Peter DaSilva for The New York Times

Data scientists spend 50 - 80% of their time mired in this more mundane labor of collecting and preparing unruly digital data, before it can be explored for useful nuggets.

[nytimes 2014-08-18](#)

Data carpentry

WRITTEN BY DAVID MIMNO

The New York Times has an article titled [For Big-Data Scientists, ‘Janitor Work’ Is Key Hurdle to Insights](#). Mostly I really like it. The fact that raw data is rarely usable for analysis without significant work is a point I try hard to make with my students. I told them “do not underestimate the difficulty of data preparation”. When they turned in their projects, many of them reported that they had underestimated the difficulty of data preparation. Recognizing this as a hard problem is great.

What I’m less thrilled about is calling this “janitor work”. For one thing, it’s not particularly respectful of custodians, whose work I really appreciate. But it also mischaracterizes what this type of work is about. I’d like to propose a different analogy that I think fits a lot better: *data carpentry*.

Note: [data carpentry](#) seems to already be a thing

Why is woodworking a better analogy? The article uses a few other terms, like data wrangling (data as unruly beasts to be tamed?) and munging (what is that, anyway?), neither of which mean much to me. I also like *data curation* but that’s also

“data science is ‘just’ statistics”

“data wrangling is not statistics”

if you value self-consistency, you can hold at most
one of these opinions

View all in Data Analysis

Data Science

A Specialization on Coursera: Your Pathway to Expertise

Final Capstone Project created with:



DATA SCIENCE

AN 11-WEEK TECHNOLOGY COURSE SUPPORTED BY



The advertisement features a black and white photograph of a person wearing glasses and a striped shirt, standing in front of a chalkboard with various graphs and data plots. On the left, there's a circular logo with a bar chart and a city skyline silhouette. The text "NYC DATA SCIENCE ACADEMY" is prominently displayed in large, bold, white letters. Below it, a blue banner contains the text "12-WEEK DATA SCIENCE BOOTCAMP". At the bottom, the text "NEXT BATCH BEGINNING JUNE 1st" and "Background in STEM" are separated by a vertical line.

We cannot expect anyone to know anything we didn't teach them ourselves.

Sarah Bryce

To a very great degree, daily work by other people sounds easy -- certainly easier than what we have to do.

Gretchen Rubin

STAT 545 now

~~permission~~ requirement to invest time in setting up tools and to develop proficiency

“simple” descriptive stats
exploration through visualization

tame data from “the wild”

alpha to omega: raw data to a web page/app

readiness for open science and automation

how to organize your work?

how to make work more pleasant for you?

how to make it navigable by others?

how to reduce tedium and manual processes?

how to reduce friction for collaboration?

how to reduce friction for communication?

specific tools and habits can build a lot of this into
the normal coding and analysis process

weak links in the chain: process, packaging and presentation



RStudio is an integrated development environment (IDE) for R

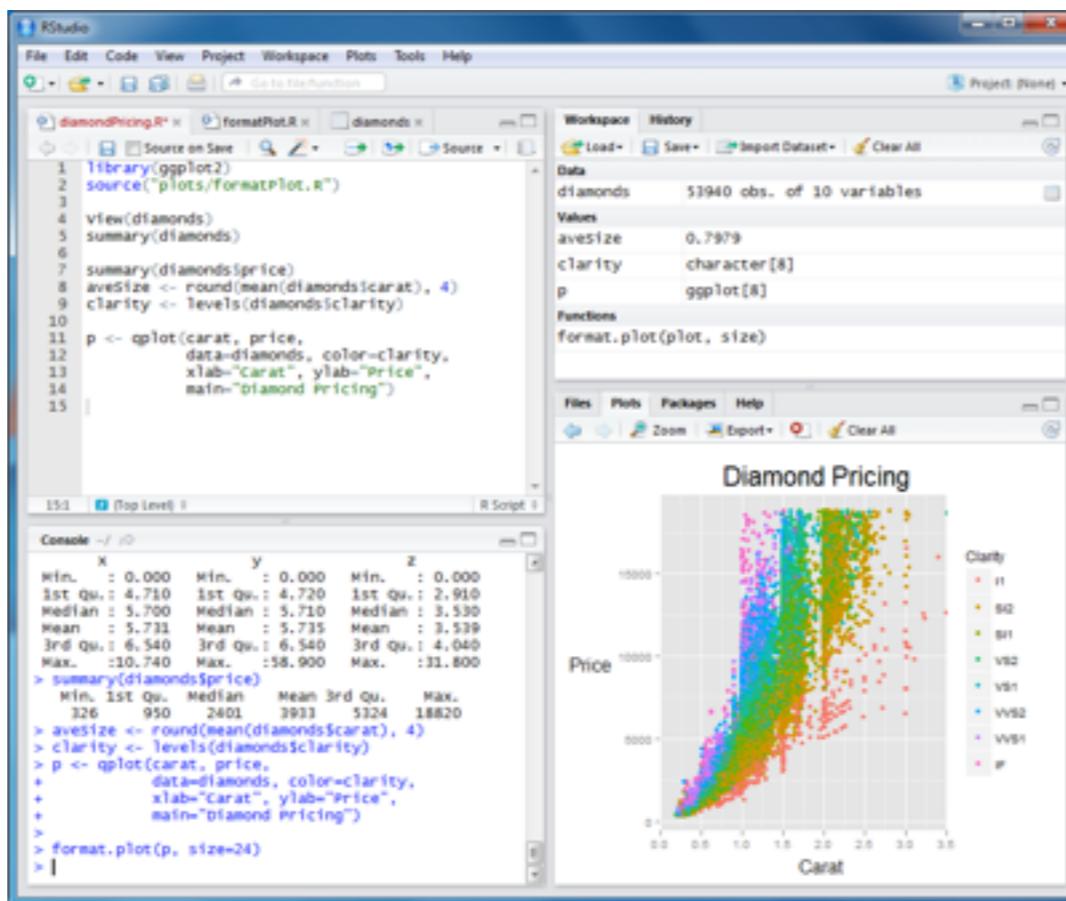
The screenshot displays the RStudio interface with the following components:

- Script Editor (Left):** Shows an R script named "diamondPricing.R" containing code to load ggplot2, source a plotting function, view and summary the diamonds dataset, calculate average size, and create a qplot for diamond pricing.
- Console (Bottom Left):** Displays the results of running the R script, including the summary statistics for the diamonds dataset and the generated plot command.
- Workspace Browser (Top Right):** Shows the "diamonds" dataset with 53940 observations and 10 variables, and a variable "p" which is a ggplot object.
- Plots (Bottom Right):** A scatter plot titled "Diamond Pricing" showing Price vs. Carat. The plot uses color to represent diamond clarity levels, with a legend on the right mapping colors to clarity grades: I1 (red), SI2 (orange), SI1 (green), VS2 (light green), VS1 (blue), VVS2 (cyan), VVS1 (purple), and IF (pink).

R ≠ RStudio

RStudio mediates your interaction with R; it would replace Emacs + ESS or Tinn-R, but not R itself

Rstudio is a product of -- actually, more a driver of -- the emergence of R Markdown, knitr, R + Git(Hub)



markdown

What is Markdown?

- Markdown is a lightweight markup language for creating HTML (or XHTML) documents.
- Markup languages are designed produce documents from human readable text (and annotations).
- Some of you may be familiar with *LaTeX*. This is another (less human friendly) markup language for creating pdf documents.
- Why I love Markdown:
 - Easy to learn and use.
 - Focus on **content**, rather than **coding** and debugging **errors**.
 - Once you have the basics down, you can get fancy and add HTML, JavaScript & CSS.

<http://cpsievert.github.io/slides/markdown/#/5>

Markdown



HTML

foo.md



foo.html

**easy to write
(and read!)**

**easy to publish
easy to read in
browser**

Markdown



HTML

```
Title (header 1, actually)  
=====
```

This is a Markdown document.

```
## Medium header (header 2, actually)
```

It's easy to do *italics* or make things bold.

> All models are wrong, but some are useful. An approximate answer to the right problem is worth a good deal more than an exact answer to an approximate problem. Absolute certainty is a privilege of uneducated minds-and fanatics. It is, for scientific folk, an you do every day matter once in a while. We can anything we didn't teach Enthusiasm is a form of

Code block below. Just we'll get to R Markdown

```
~~~  
x <- 3 * 4  
~~~
```

I can haz equations. Inline equations, such as ... the average is computed as $\frac{1}{n} \sum_{i=1}^n x_i$. Or display equations like this:

```
$$  
\begin{equation*}  
|x| =  
\begin{cases} x & \text{if } x \geq 0, \\ -x & \text{if } x \leq 0. \end{cases} \quad \quad  
\end{equation*}  
$$
```



```
<!DOCTYPE html>  
<html>  
<head>  
<meta http-equiv="Content-Type" content="text/html;  
charset=utf-8"/>
```

```
<title>Title (header 1, actually)</title>
```

```
<!-- MathJax scripts -->  
<script type="text/javascript" src="https://  
c328740.ssl.cf1.rackcdn.com/mathjax/2.0-latest/  
MathJax.js?config=TeX-AMS-MML_HTMLorMML">  
</script>
```

Fess up: How many of you still hand-code HTML?

```
<h1>Title (header 1, actually)</h1>
```

```
<p>This is a Markdown document.</p>
```

```
<h2>Medium header (header 2, actually)</h2>
```

```
<p>It's easy to do <em>italics</em> or  
<strong>make things bold</strong>.</p>
```

```
<blockquote>
```

```
<p>All models are wrong, but some are...
```

```
<p>Code block below. Just affects formatting here  
but we'll get to R Markdown for the real fun  
soon!</p>
```

```
<pre><code>x < 3 * 4  
</code></pre>
```

Markdown



HTML

Title (header 1, actually)

This is a Markdown document.

Medium header (header 2, actually)

It's easy to do *italics* or make things bold.

> All models are wrong, but some are useful. An approximate answer to the right problem is worth a good deal more than an exact answer to an approximate problem. Absolute certainty is a privilege of uneducated minds-and fanatics. It is, for scientific folk, an unattainable ideal. What you do every day matters more than what you do once in a while. We cannot expect anyone to know anything we didn't teach them ourselves.

Enthusiasm is a form of social courage.

Code block below. Just affects formatting here but we'll get to R Markdown for the real fun soon!

```
```  
x <- 3 * 4
```
```

I can haz equations. Inline equations, such as ... the average is computed as $\frac{1}{n} \sum_{i=1}^n x_i$. Or display equations like this:

```
$$  
\begin{equation*}  
|x| =  
\begin{cases} x & \text{if } x \geq 0, \\ -x & \text{if } x \leq 0. \end{cases} \\\\  
\end{equation*}  
$$
```



Title (header 1, actually)

This is a Markdown document.

Medium header (header 2, actually)

It's easy to do *italics* or **make things bold**.

All models are wrong, but some are useful. An approximate answer to the right problem is worth a good deal more than an exact answer to an approximate problem. Absolute certainty is a privilege of uneducated minds-and fanatics. It is, for scientific folk, an unattainable ideal. What you do every day matters more than what you do once in a while. We cannot expect anyone to know anything we didn't teach them ourselves. Enthusiasm is a form of social courage.

Code block below. Just affects formatting here but we'll get to R Markdown for the real fun soon!

```
x <- 3 * 4
```

I can haz equations. Inline equations, such as ... the average is computed as $\frac{1}{n} \sum_{i=1}^n x_i$. Or display equations like this:

$$|x| = \begin{cases} x & \text{if } x \geq 0, \\ -x & \text{if } x \leq 0. \end{cases}$$

How is the math getting typeset?

Answer: Mathjax

MathJax is an open source JavaScript display engine for mathematics that works in all browsers.

No more setup for readers. No more browser plugins. No more font installations... It just works.

How painful is that to use?

Not at all. Automagic with knitr and RStudio.

What happens to equations if the reader is not connected to the internet?

The LaTeX is displayed. No great harm.

Code block below. Just affects formatting here but we'll get to R Markdown for the real fun soon!

```
x <- 3 * 4
```

I can haz equations. Inline equations, such as ... the average is computed as $\frac{1}{n} \sum_{i=1}^n x_i$. Or display equations like this:

```
\begin{equation*} |x| = \begin{cases} x & \text{if } x \geq 0, \\ -x & \text{if } x < 0. \end{cases} \end{equation*}
```

If I use Markdown, am I restricted to HTML output?

No.

pandoc = “swiss-army knife” of document conversion
(RStudio will gladly install and invoke for you.)

About pandoc

If you need to convert files from one markup format into another, pandoc is your swiss-army knife. Pandoc can convert documents in [markdown](#), [reStructuredText](#), [textile](#), [HTML](#), [DocBook](#), [LaTeX](#), [MediaWiki markup](#), [OPML](#), or [Haddock markup](#) to

- HTML formats: XHTML, HTML5, and HTML slide shows using [Slidy](#), [reveal.js](#), [Slideous](#), [S5](#), or [DZSlides](#).
- Word processor formats: Microsoft Word [docx](#), OpenOffice/LibreOffice [ODT](#), [OpenDocument XML](#)
- Ebooks: [EPUB](#) version 2 or 3, [FictionBook2](#)
- Documentation formats: [DocBook](#), [GNU TexInfo](#), [Groff man](#) pages, [Haddock markup](#)
- Outline formats: [OPML](#)
- TeX formats: [LaTeX](#), [ConTeXt](#), LaTeX Beamer slides
- [PDF](#) via [LaTeX](#)
- Lightweight markup formats: [Markdown](#), [reStructuredText](#), [AsciiDoc](#), [MediaWiki markup](#), Emacs [Org-Mode](#), [Textile](#)
- Custom formats: custom writers can be written in [lua](#).

If you have an annoying process for authoring for the web

or

If you avoid authoring for the web, because you're not sure how ...

start writing in Markdown.

R markdown

R Markdown

Markdown

R Markdown rocks

This is an R Markdown document.

```
```{r}
x <- rnorm(1000)
head(x)
```
```

See how the R code gets executed and a representation thereof appears in the document? `knitr` gives you control over how to represent all conceivable types of output. In case you care, then average of the `r length(x)` random normal variates we just generated is `r round(mean(x), 3)`. Those numbers are NOT hard-wired but are computed on-the-fly. As is this figure. No more copy-paste ... copy-paste ... oops forgot to copy-paste.

```
```{r}
plot(density(x))
```
```

Note that all the previously demonstrated math typesetting still works. You don't have to choose between having math cred and being web-friendly!

Inline equations, such as ... the average is computed as $\frac{1}{n} \sum_{i=1}^n x_i$. Or display equations like this:

```
$$
\begin{aligned}
|x| = \\
\begin{cases} x & \text{if } x \geq 0, \\ -x & \text{if } x \leq 0. \end{cases}
\end{aligned}
$$
```

R Markdown rocks

This is an R Markdown document.

```
```{r}
x <- rnorm(1000)
head(x)
```
```

```
## [1] -1.3007  0.7715  0.5585 -1.2854  1.1973
2.4157
```
```

See how the R code gets executed and a representation thereof appears in the document? `knitr` gives you control over how to represent all conceivable types of output. In case you care, then average of the 1000 random normal variates we just generated is -0.081. Those numbers are NOT hard-wired but are computed on-the-fly. As is this figure. No more copy-paste ... copy-paste ... oops forgot to copy-paste.

```
```{r}
plot(density(x))
```
```

```
![plot of chunk unnamed-chunk-2](figure/unnamed-
chunk-2.png)
```

```
...
```

# Markdown → HTML

R Markdown rocks

This is an R Markdown document.

```
```r
x <- rnorm(1000)
head(x)
```
```
## [1] -1.3007  0.7715  0.5585 -1.2854  1.1973
2.4157
````
```

See how the R code gets executed and a representation thereof appears in the document? `knitr` gives you control over how to represent all conceivable types of output. In case you care, then average of the 1000 random normal variates we just generated is -0.081. Those numbers are NOT hard-wired but are computed on-the-fly. As is this figure. No more copy-paste ... copy-paste ... oops forgot to copy-paste.

```
```r
plot(density(x))
````
```

```
![plot of chunk unnamed-chunk-2](figure/unnamed-
chunk-2.png)
```

...

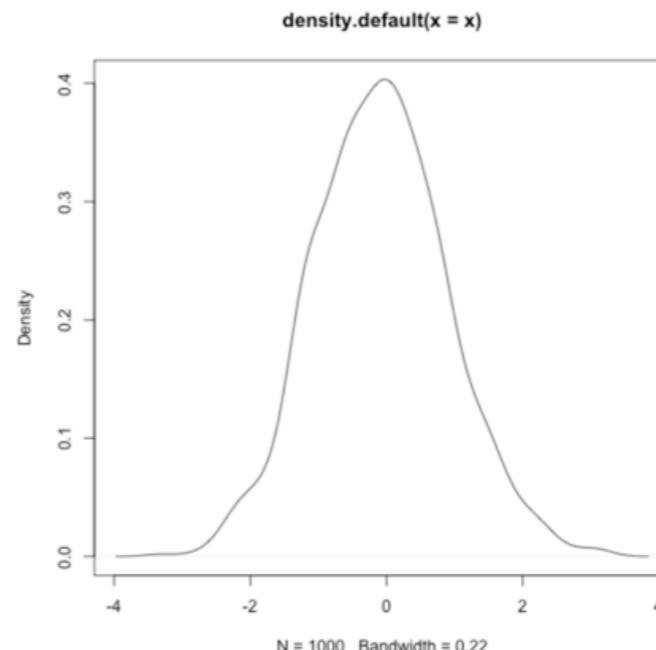
## R Markdown rocks

This is an R Markdown document.

```
x <- rnorm(1000)
head(x)
```

```
[1] -1.3007 0.7715 0.5585 -1.2854 1.1973 2.4157
```

See how the R code gets executed and a representation thereof appears in the document? `knitr` gives you control over how to represent all conceivable types of output. In case you care, then average of the 1000 random normal variates we just generated is -0.081. Those numbers are NOT hard-wired but are computed on-the-fly. As is this figure. No more copy-paste ... copy-paste ... oops forgot to copy-paste.



Note that all the previously demonstrated math typesetting still works. You don't have to choose between having math cred and being web-friendly!

Inline equations, such as ... the average is computed as  $\frac{1}{n} \sum_{i=1}^n x_i$ . Or display equations like this:

$$|x| = \begin{cases} x & \text{if } x \geq 0, \\ -x & \text{if } x \leq 0. \end{cases}$$

R Markdown → Markdown → HTML

**foo.rmd** → **foo.md** → **foo.html**

easy to write  
(and read!)

easy to publish  
easy to read in  
browser

# How do to actually convert Markdown to HTML?

`knitr`, `rmarkdown` add-on packages provide user-friendly functions

RStudio makes them available via button

# R Markdown

→ HTML

R Markdown rocks

=====

This is an R Markdown document.

```
```{r}
x <- rnorm(1000)
head(x)
```
```

See how the R code gets executed and a

represen

`knitr`

conceal

averag

we jus

number

fly. A

paste

```
```{r}
plot(d
```

```

Note t

typese

betwee

Inline

comput

displa

\$\$

\begin{

| x | =

\begin{

- x & \text{if } x \leq 0 \}

\end{cases}

\end{equation\*}

\$\$

# R Markdown rocks

This is an R Markdown document.

```
x <- rnorm(1000)
head(x)
```

```
[1] -1.3007 0.7715 0.5585 -1.2854 1.1973 2.4157
```

See how the R code gets executed and a representation thereof appears in the

## How to achieve at the command line:

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

equations like this:

$$|x| = \begin{cases} x & \text{if } x \geq 0, \\ -x & \text{if } x \leq 0. \end{cases}$$

# R Markdown

→ HTML

R Markdown rocks  
=====

This is an R Markdown document.

```
```{r}  
x <- rnorm(1000)  
head(x)
```

~/tmp/test - RStudio

File | New | Open | Save | Print | Go to file/function

test

foo.Rmd x

1 ---
2 title: "Untitled"
3 author: "Jenny Bryan"
4 date: "22 February, 2015"
5 output: html_document
6 ---
7
8 This is an R Markdown document. Markdown is a simple
formatting syntax for authoring HTML, PDF, and MS
Word documents. For more details see [http://rmarkdown.rstudio.com](#)

1:1 (Top Level) R Markdown

Click here.

Console ~/tmp/test/
Type 'citation()' on how to cite R or R packages in publications
.
Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

```
\begin{cases} x & \text{if } x \geq 0 \\ -x & \text{if } x \leq 0 \end{cases}
```

R Markdown rocks

This is an R Markdown document.

```
x <- rnorm(1000)  
head(x)
```

~/tmp/test - RStudio

File | New | Open | Save | Print | Go to file/function

Environment History

To Console To Source

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

Files Plots Help Viewer

New Folder Delete Rename

Home > tmp > test

Name	Size	Modified
..		
test.Rproj	257 B	Feb 22, 2015, 9:47 PM
foo.Rmd	723 B	Feb 22, 2015, 9:47 PM

Inline equations, such as ... the average is computed as $\frac{1}{n} \sum_{i=1}^n x_i$. Or display equations like this:

$$|x| = \begin{cases} x & \text{if } x \geq 0, \\ -x & \text{if } x \leq 0. \end{cases}$$

Do I have to do everything in R markdown? What about plain R scripts?

Use `rmarkdown::render()` or Rstudio's Compile Notebook button to get a satisfying stand-alone webpage based on an R script.

simple R script: toyline.R

```
1 a <- 2
2 b <- 7
3 sigSq <- 0.5
4 n <- 400
5
6 set.seed(1234)
7 x <- runif(n)
8 y <- a + b * x + rnorm(n, sd = sqrt(sigSq))
9
10 (avgX <- mean(x))
11
12 plot(x, y)
13 abline(a, b, col = "blue", lwd = 2)
```

→ **HTML**

toyline.R

jenny — Sep 6, 2013, 3:15 PM

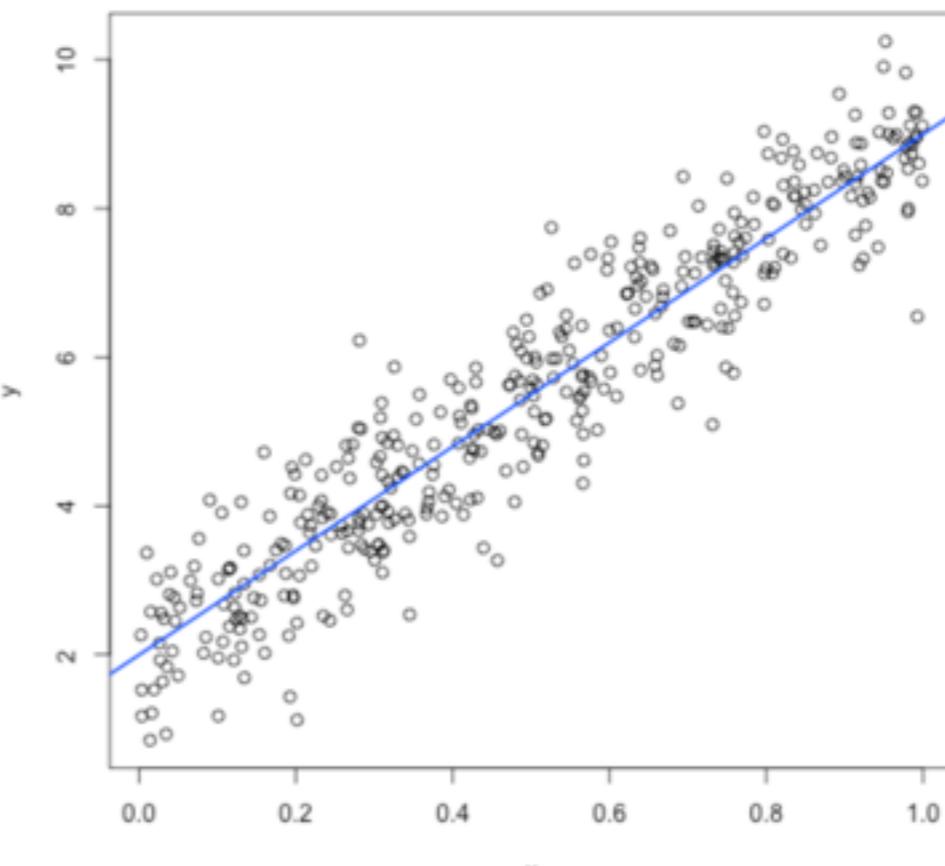
```
a <- 2
b <- 7
sigSq <- 0.5
n <- 400

set.seed(1234)
x <- runif(n)
y <- a + b * x + rnorm(n, sd = sqrt(sigSq))

(avgX <- mean(x))
```

```
[1] 0.4969
```

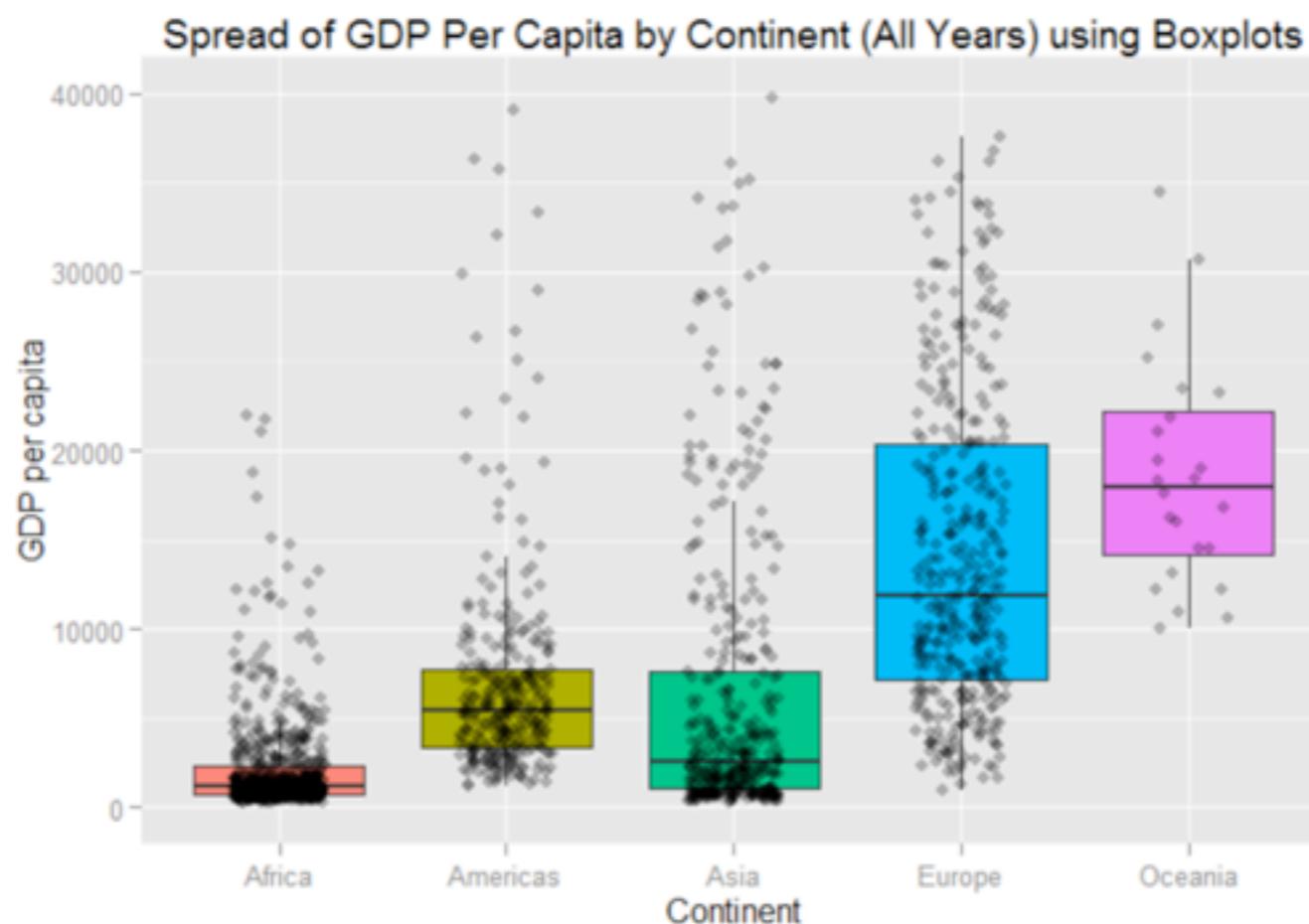
```
plot(x, y)
abline(a, b, col = "blue", lwd = 2)
```



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")
```



R Markdown: Integrating A Reproducible Analysis Tool into Introductory Statistics
Technology Innovations in Statistics Education, 8(1)

Baumer, Ben, Smith College

Cetinkaya-Rundel, Mine, Duke University

Bray, Andrew, Smith College

Loi, Linda, Smith College

Horton, Nicholas J., Amherst College

Publication Date:

2014

Permalink:

<https://escholarship.org/uc/item/90b2f5xh>

How do I show the world all these awesome dynamic HTML reports I'm creating?

Easiest: Rpubs

Or do whatever you usually do to get HTML on the web.

Or use GitHub

Big picture, so far:

web-friendly is good

various hosting platforms make it easy to share web-ready products with minimal effort

embedding analysis and logic in source document for a report is good

- huge win for reproducibility
- also excellent for communication and documentation

(R) Markdown + knitr (+ RStudio) make it very easy to author dynamic reports that are ready for the web

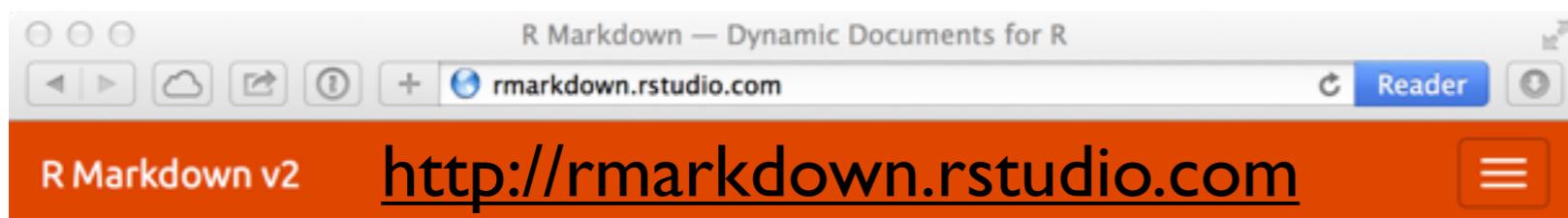
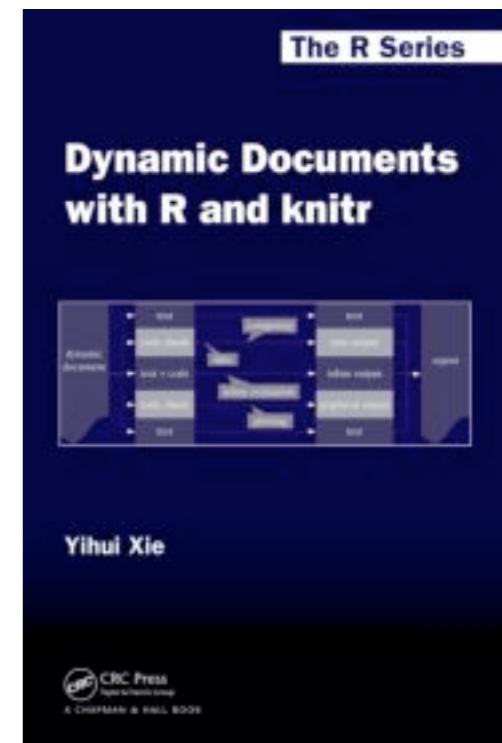
disclaimer:

knitr is **not limited** to executing R code
knitr is **not limited** to processing R Markdown

I just chose to focus on R and R Markdown

Read more in the book or on the web:

Dynamic documents with R and knitr by Yihui Xie,
part of the CRC Press / Chapman & Hall R
Series (2013). ISBN: 9781482203530.



R Markdown — Dynamic Documents for R

R Markdown is an authoring format that enables easy creation of dynamic documents, presentations, and reports from R. It combines the core syntax of **markdown** (an easy-to-write plain text format) with embedded R code chunks that are run so their output can be included in the final document. R Markdown documents are fully *reproducible* (they can be automatically regenerated whenever underlying R code or data changes).

OK you've got a collection of ...

R scripts

R package

R Markdown files

input data

intermediate results

figures

output tables

compiled reports

all evolving over time

how do you keep track of this?

how do I put my stuff on the web?

for the world or select collaborators?

Advice to preserve sanity:

Stop doing this via email, attachments, and tracking changes in Word. Get that stuff into plain text, put it under version control and get it out on the web.

"FINAL".doc



FINAL.doc!



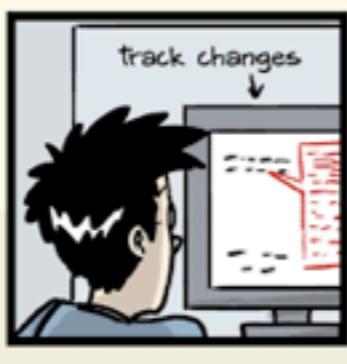
FINAL_rev.2.doc



↑
FINAL_rev.6.COMMENTS.doc



↑
FINAL_rev.8.comments5.
CORRECTIONS.doc



FINAL_rev.18.comments7.
corrections9.MORE.30.doc



FINAL_rev.22.comments49.
corrections.10.#@\$%WHYDID
ICOMETOGRADSCHOOL????.doc

JORGE CHAM © 2012

Version control systems (VCS) were created to help groups of people develop software

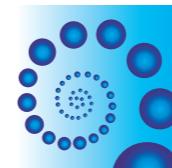
Git, in particular, is being “repurposed” for activities other than pure software development ... like the messy hybrid of writing, coding and data wrangling



“Git, provides a lightweight yet robust framework that is ideal for managing the full suite of research outputs such as datasets, statistical code, figures, lab notes, and manuscripts.”

“... this tool can be leveraged to make science more reproducible and transparent, foster new collaborations, and support novel uses.”

Ram *Source Code for Biology and Medicine* 2013, **8**:7
<http://www.scfbm.org/content/8/1/7>



SOURCE CODE FOR
BIOLOGY AND MEDICINE

BRIEF REPORTS

Open Access

Git can facilitate greater reproducibility and increased transparency in science

Karthik Ram

GitHub repository for this paper: https://github.com/karthik/smb_git

Ram: Git can facilitate greater reproducibility and increased transparency in science. *Source Code for Biology and Medicine* 2013 8:7. doi:10.1186/1751-0473-8-7

collaboration = the “killer app” of version control

Learning Git has been -- and continues to be -- painful. But not nearly as crazy-making as the alternatives:

- documents as email attachments
- uncertainty about which version is “master”
- am I working with the most recent data?
- archaeological “digs” on old email threads
- uncertainty about how/if certain changes have been made or issues solved
- hair-raising ZIP archives containing file salad

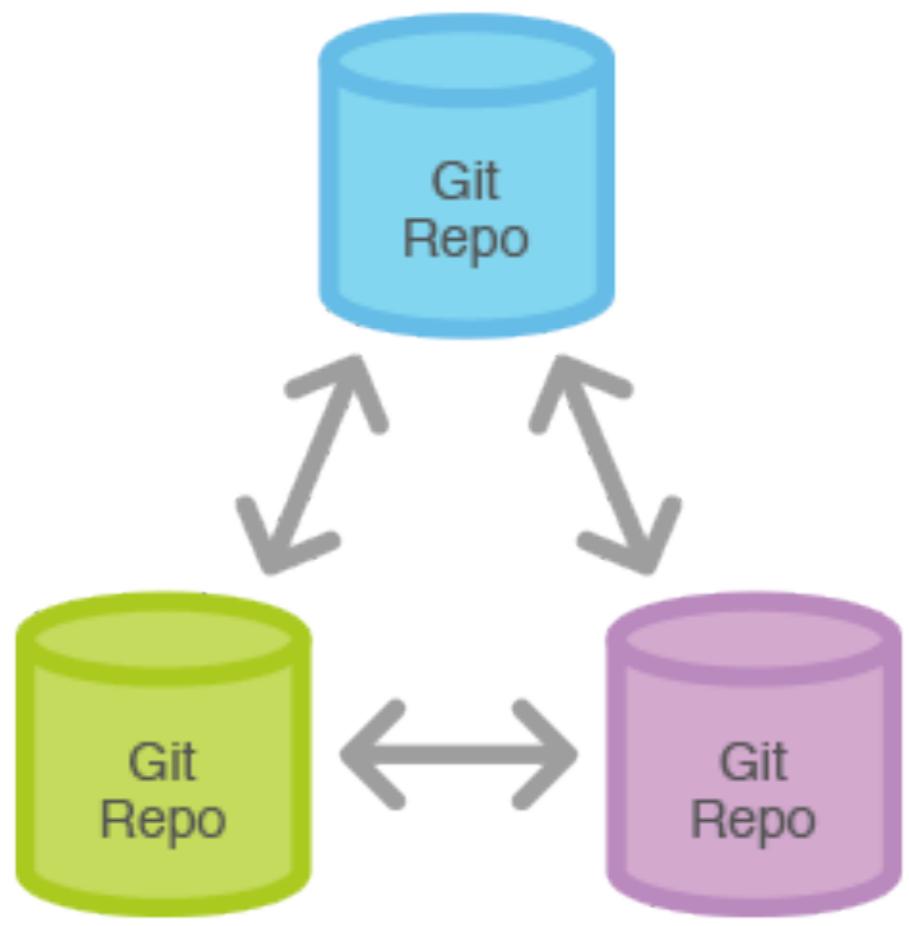
Git **repository** = a bunch of files you want to manage in a sane way

repo = repository

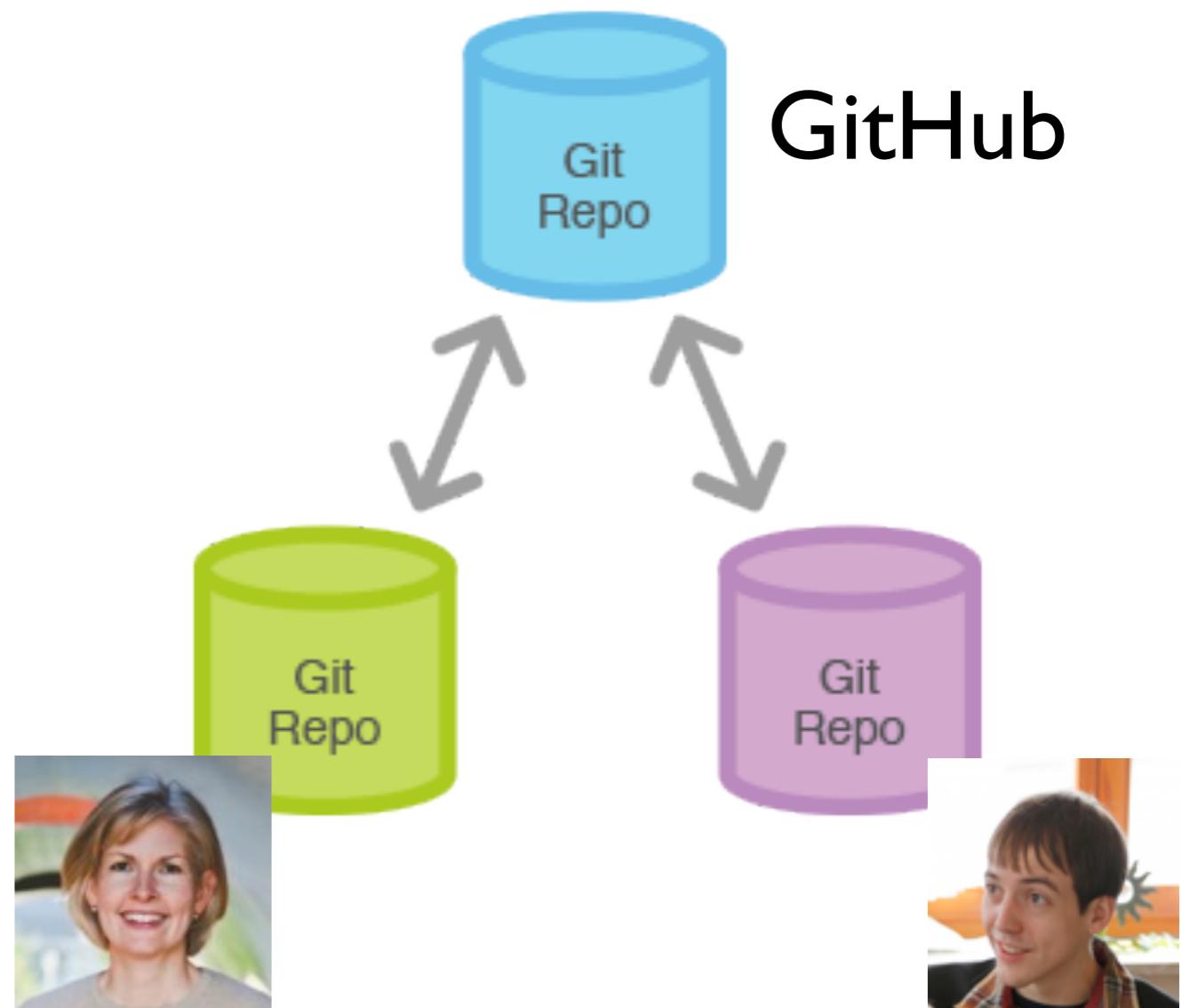
you can set up repo ... then start your work

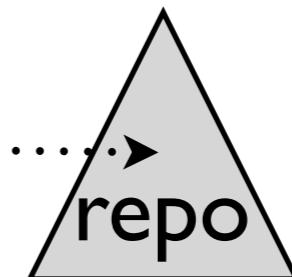
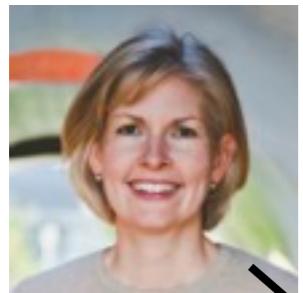
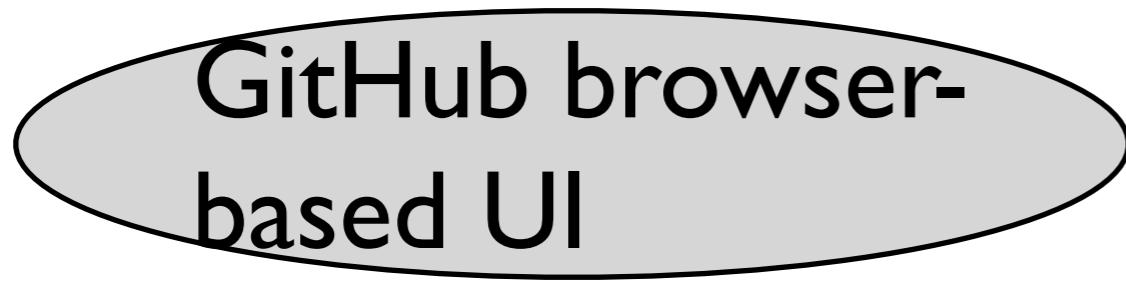
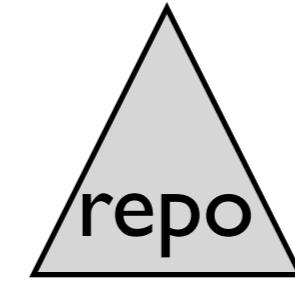
or you can make a set of existing files and make them into a repo

in theory

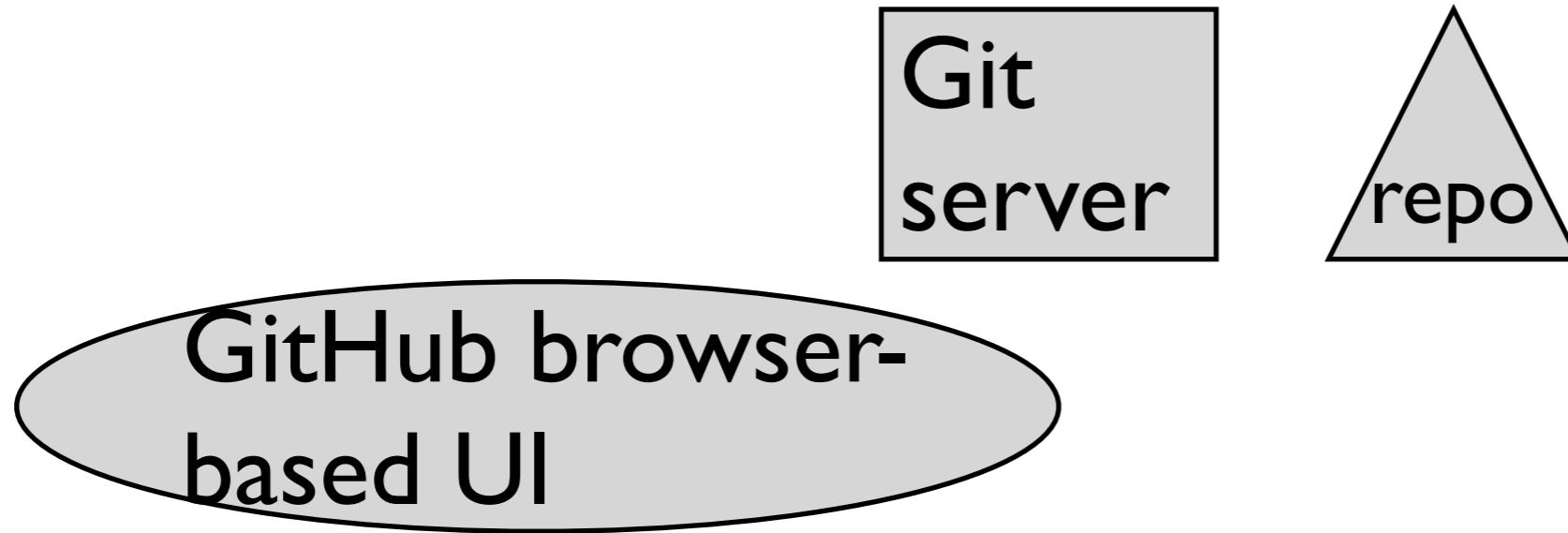


more typical

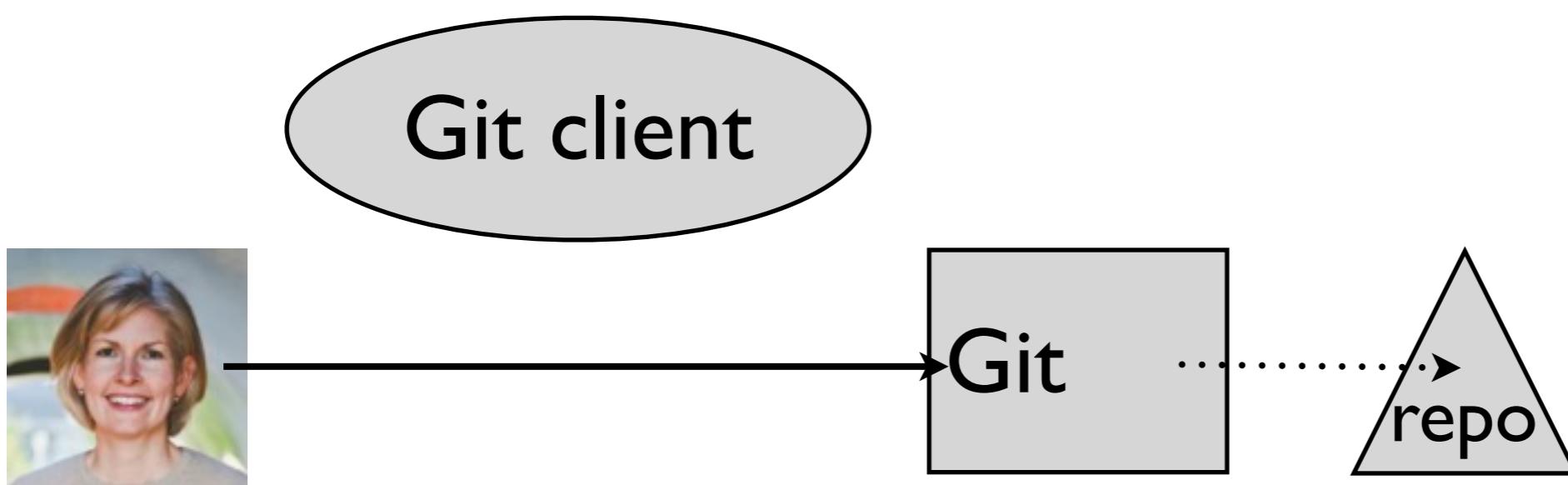




Using a Git client to commit
(a local operation)



Using Git at the command line to commit (a local operation)





GitHub browser-based UI

Git client

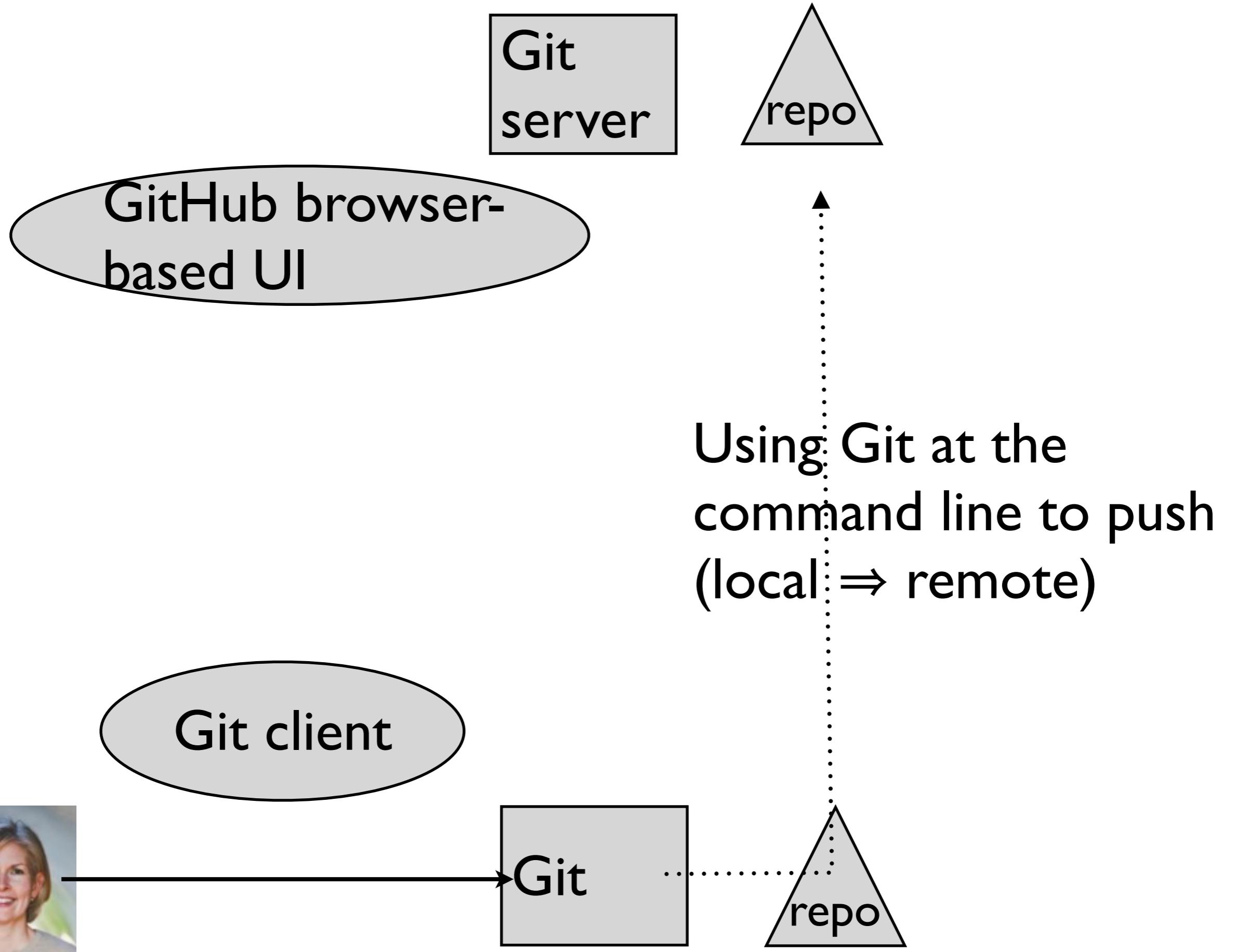
Git server

repo

Using a Git client to push (local \Rightarrow remote)

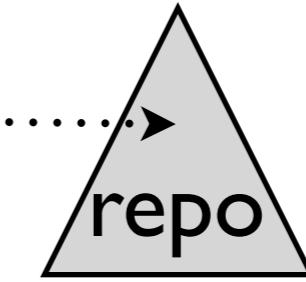
Git

repo



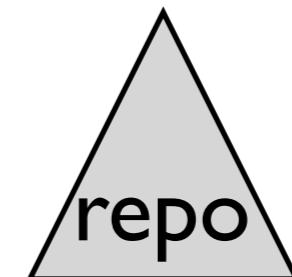


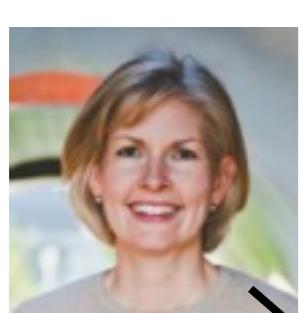
GitHub browser-based UI



Operating on a Git repo via GitHub in the browser

Git client





GitHub browser-based UI

Git client

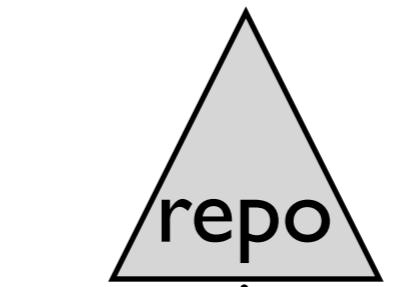
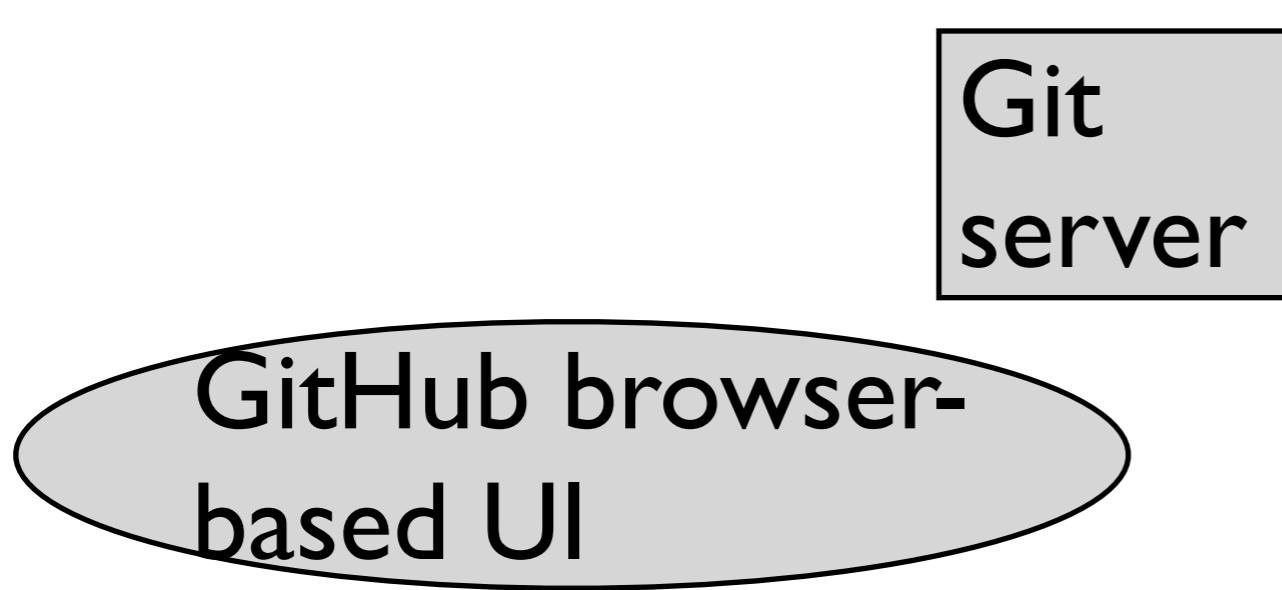
Git server

repo

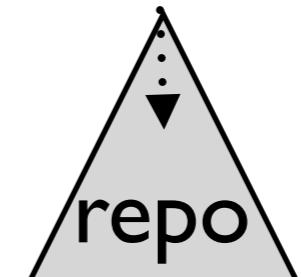
Git

repo

Using a Git client to
pull (remote \Rightarrow local)



Using Git at the
command line to pull
(remote \Rightarrow local)



GitHub = a place to host Git repositories on the web

GitHub ≠ Git

The screenshot shows a GitHub repository page for `STAT545-UBC/zz_derek_chiu-coursework`. The repository is private, has 259 commits, 1 branch, 0 releases, and 1 contributor. The master branch is selected. The repository contains files like `class_meetings`, `data`, `homeworks`, `.gitignore`, `README.md`, and `STAT545A.Rproj`. A commit by `dchiu911` from 14 days ago prepended class_meeting files with "cmXXX". The repository is described as a STAT 545A Homework Repository.

coursework created for zz_derek_chiu — Edit

259 commits 1 branch 0 releases 1 contributor

branch: master zz_derek_chiu-coursework / +

README for /homeworks

dchiu911 authored 14 days ago latest commit 3ccf25c6e0

File	Description	Time
<code>class_meetings</code>	Prepended class_meeting files with "cmXXX"	14 days ago
<code>data</code>	README for /data	14 days ago
<code>homeworks</code>	README for /homeworks	14 days ago
<code>.gitignore</code>	Changed .Rprofile and .gitignore	3 months ago
<code>README.md</code>	Updated README for hw07-12	14 days ago
<code>STAT545A.Rproj</code>	Changed makefilepath	14 days ago

HTTPS clone URL: <https://github.com/>

You can clone with [HTTPS](#), [SSH](#), or [Subversion](#).

Clone in Desktop Download ZIP

STAT 545A Homework Repository

Introduction

This is the repository of Derek Chiu. It contains all the homework submitted for the course STAT 545A. The main course webpage can be found at <http://stat545-ubc.github.io>. I am a first year M.Sc. in Statistic student with an option in Biostatistics. In my Co-op work terms, I have used `ggplot2` for

The image shows three GitHub repository pages side-by-side, each highlighted with an orange circle:

- Top Left (Circled):** hilaryparker / **cats**. An R package for cat-related functions. It has 10 commits. The branch is master. Contributors include karthik (author on Aug 1, 2014). The sidebar shows files like R, man, tests, .Rbuildignore, .travis.yml, DESCRIPTION, NAMESPACE, and README.md.
- Middle (Circled):** hadley / **dplyr**. PLYR specialised for data frames: faster & with remote datastores. It has 2,178 commits, 14 branches, 10 releases, and 40 contributors. The branch is master.
- Bottom Right (Circled):** dgrtwo / **broom**. Convert statistical analysis objects from various packages into tidy, consistent objects. It has 89 commits. The branch is master. Contributors include dgrtwo (author 5 days ago). The sidebar shows files like R, man-roxygen, and man.

Text Box:

Many R packages are developed in the open on GitHub

Nice option when someone tells you to “read the source”!

Many government agencies, media outlets, academic labs, etc. put their stuff on GitHub

<https://github.com/WhiteHouse>

<https://github.com/chicago>

<https://github.com/fivethirtyeight>

<https://github.com/TheUpshot>

<https://github.com/propublica/>

<http://ncip.github.io> (NCI's informatics program)

<https://github.com/LSST> (Large Synoptic Survey Telescope)

<https://github.com/ctb> (Titus Brown lab)

<https://github.com/lh3> (Heng Li lab)

STAT 545 is an Organization on GitHub

all course materials are posted there (public repo)

all course development was done there (private
repo for instructors only)

each student had his/her own repo for coursework
(visible only within the Organization)

rough notes on set-up

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

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

The screenshot shows a GitHub repository page for 'hw03_dplyr-manipulation' at the 'master' branch. The top navigation bar includes icons for back, forward, refresh, and search, followed by the GitHub logo and the URL 'github.com/STAT545-UBC/zz_derek_chiu-coursework/tree/master/homework'. A 'Reader' button is also present. The main content area displays a commit history and a file tree. The commit history shows a single commit from 'dchiu911' 15 days ago, titled 'Gave names to code chunks with figures'. Below this, there are several files listed with their descriptions and times:

File	Description	Time
hw03_dplyr-manipulation_files/figure-html	Gave names to code chunks with figures	15 days ago
README.md	Fixed typos in links to hw03 README	16 days ago
hw03_dplyr-manipulation.R	Generated R script using knitr:::purl	15 days ago
hw03_dplyr-manipulation.html	Gave names to code chunks with figures	15 days ago
hw03_dplyr-manipulation.md	Gave names to code chunks with figures	15 days ago
hw03_dplyr-manipulation.rmd	Gave names to code chunks with figures	15 days ago

Below the commit history, there is a section for 'README.md' with a link to the file.

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 available in the `data` folder

Commits are how the files evolve

The screenshot shows a GitHub repository page for 'STAT545-UBC / STAT545-UBC.github.io'. The top navigation bar includes links for 'Explore', 'Gist', 'Blog', and 'Help'. On the right, there's a user profile for 'jennybc' with options to '+', 'Edit', and settings. Below the header, the repository name 'STAT545-UBC / STAT545-UBC.github.io' is displayed, along with statistics: 495 commits, 2 branches, 0 releases, and 8 contributors. A red circle highlights the '495 commits' link. To the right, there are links for 'Code', 'Issues' (3), 'Pull Requests' (1), 'Pulse', 'Graphs', and 'Settings'. The main content area shows a list of recent commits:

Commit	Message	Date
jennybc authored an hour ago	latest commit ce2a2b5198	an hour ago
automation01_slides	Automation: Delete Makefile.gv and Makefile.png	3 months ago
automation10_holding-area	Automation: Delete Makefile.gv and Makefile.png	3 months ago
block002_hello-r-workspace-w...	recompile block002 subsequent to merging pull request in 9feb137	2 months ago
block006_care-feeding-data_fi...	finish 2014-ification of basic care and feeding of data(frames)	5 months ago
block011_write-your-own-func...	partially done with block011_write-your-own-function	5 months ago
block012_function-regress-life...	write lifeExp ~ year function	5 months ago
block014_factors_files/figure-...	small updates to block014_factors	4 months ago

Commit message = short description of what/why changed

Commits · STAT545-UBC/STAT545-UBC.github.io

GitHub, Inc. github.com/STAT545-UBC/STAT545-UBC.github.io/commits/master Reader

jennybc authored on Jan 5

updating links to shiny material as referenced in STAT545-UBC/Instruc... [...](#) [ec5ab6b](#) [🔗](#)

joolia authored on Jan 5

add info about fall 2015 to faq [...](#) [95f04b2](#) [🔗](#)

jennybc authored on Jan 5

Commits on Jan 1, 2015

add a resource for difference btwn Depends and Imports in packages fr... [...](#) [1db68d3](#) [🔗](#)

daattali authored on Jan 1

Commits on Dec 30, 2014

Merge pull request #15 from zhaoy/master [...](#) [9feb137](#) [🔗](#)

BernhardKonrad authored on Dec 30, 2014

Commits on Dec 29, 2014

Removed redundant getwd() [...](#) [8af9cc6](#) [🔗](#)

zhaoy authored on Dec 29, 2014

Fixing typo [...](#) [f306f45](#) [🔗](#)

zhaoy authored on Dec 29, 2014

Commits on Dec 21, 2014

link to tutorial on sending email with r and gmailr [...](#) [ea4cf0f](#) [🔗](#)

jennybc authored on Dec 21, 2014

Go to "<https://github.com/STAT545-UBC/STAT545-UBC.github.io/commit/1db68d3c95b55fe0c1b9cc13f840a04989c8aea8>"

“diffs” compare a file then vs. now

add info about fall 2015 to faq · 95f04b2 · STAT545-UBC/STAT545-UBC.github.io

GitHub, Inc. github.com/STAT545-UBC/STAT545-UBC.github.io/commit/95f04b2739ecac5bdfd2e84d3a1fef11d Reader

10 faq.md

10 8 @@ -8,7 +8,11 @@ output:
8 8
9 9
10 10
11 -### Course facts
11 +### When is the course next offered?
12 +
13 +September - Dec 2015 *to be confirmed, but very likely*
14 +
15 +### Course facts for Sept - Dec 2014 run
12 16
13 17 | STAT 545A | STAT547M |
14 18 |-----|-----|-----|
10 @@ -31,12 +35,12 @@ Up-to-date info on [office hours](https://github.com/STAT545-UBC/Discussion/issues/35)
31 35
32 36 For several years, I have taught STAT 545A as a 1.5 credit course. I -- and many students -- have felt there was a lot of great, relevant content that could go into an additional 1.5 credits.
33 37
34 -Therefore, in 2014/2015, we will pilot a full semester on data exploration, visualization, and all-around data wrangling. It is structured as two half courses for various reasons, such as allowing STAT 545A alums to register for STAT 547M and get the "missing half" of the course!
38 +Therefore, in 2014/2015, we piloted a full semester on data exploration, visualization, and all-around data wrangling. It was structured as two half courses primarily so that STAT 545A alums could register for STAT 547M and get the "missing half" of the course. We're still figuring out the long term plan re: 2 courses of 1.5 credits vs. 1 course of 3 credits.
35 39
36 40 ### Am I allowed to register in ...?
37 41
38 42 * I have taken STAT 545A for 1.5 credits in the past. Can I take STAT 547M?
39 - - YES. But you will want to follow along during STAT 545A (at least online), so you get some new content. Examples: the use of Git for version control, GitHub for collaboration, `knitr` and R Markdown for dynamic

GitHub repositories can have *issues*: think discussion forum.

The screenshot shows a GitHub repository named "STAT545-UBC / Discussion". The page title is "Issues · STAT545-UBC/Discussion". The top navigation bar includes links for "Explore", "Gist", "Blog", and "Help". The user "jennybc" is logged in. The repository has 23 issues, 1 star, and 1 fork. The "Issues" tab is selected, showing 19 open issues and 64 closed issues. A search bar filters the results by "is:issue is:open".

Issue Number	Title	Author	Labels	Milestones	Assignee	Comments
#83	Future plans for your STAT 545 / 547 coursework repository	jennybc				3
#82	Worked example: when two data.frames are almost, but not quite, the same	jennybc				1
#81	Using `rplots` in RMarkdown without including API key inside RMarkdown?	spencerfrei				5
#80	Remixing data -- `dplyr` bug	aammd				0
#78	Small detail: Indented Rstudio code not displaying nicely with github's tab setting of 8 spaces	joolia	Mac OS Windows			0
#76	Deploying our shiny apps to the UBC stats server	daattali				2

GitHub repositories can have *issues*: think “to do list”

The screenshot shows a GitHub repository page for "STAT545-UBC / zz_derek_chiu-coursework". The repository is private, with 2 unwatched, 0 stars, and 1 fork. The user "jennybc" is logged in. The "Issues" tab is selected, showing 33 closed issues. A search bar filters the results to "is:issue is:closed". The issues listed are:

- Mark homework 12 of Derek-Chiu (#33) - opened on Dec 5, 2014 by dchiu911
- Peer review of derek_chiu's hw11 by beryl_zhuang (#32) - opened on Dec 1, 2014 by jennybc
- Peer review of derek_chiu's hw11 by abrar_wafa (#31) - opened on Dec 1, 2014 by jennybc
- Mark homework 11 of Derek-Chiu (#30) - opened on Dec 1, 2014 by dchiu911
- Peer review of derek_chiu's hw10 by omar_alomeir (#29) - opened on Nov 25, 2014 by jennybc

GitHub repositories can have *issues*: think “bug tracker”

The screenshot shows the GitHub Issues page for the `hadley/dplyr` repository. The page has a header with navigation icons, a search bar, and user profile information. Below the header, the repository name `hadley / dplyr` is displayed, along with metrics for 132 issues, 656 stars, and 239 forks. A navigation bar at the top includes tabs for **Issues**, **Pull requests**, **Labels**, and **Milestones**. A filter bar shows the current filter as `is:issue is:open`. A green button for **New issue** is visible. The main content area lists 144 open issues, each with a title, a brief description, the number of comments (indicated by a speech bubble icon), and the date it was opened. The issues listed are:

- dplyr crashing in RStudio Server on Linux** #989 opened 3 days ago by pyftime
- unexpected result when arranging after grouping by multiple variables** #986 opened 3 days ago by justmarkham
- Performance issues with joins** #984 opened 5 days ago by pimentel
- rename() should generate a subquery in SQL translation** #983 opened 5 days ago by zozlak
- SQL translation of "x %in% 1" fails** #982 opened 5 days ago by zozlak
- mutate segfaults on many groups for some functions (though summarise works OK)** #979 opened 5 days ago by andrewblim

Markdown files are automatically rendered nicely in GitHub repositories

The screenshot shows a GitHub repository interface. At the top, there's a header bar with icons for file operations and a URL: `zz_derek_chiu-coursework/homeworks/hw03_dplyr-manipulation at master · STAT545-UBC/zz_derek_chiu-coursework`. Below the header is a toolbar with navigation buttons (back, forward, refresh, etc.), a GitHub logo, and a 'Reader' button.

The main area displays a list of files and their descriptions:

File	Description	Last Commit
<code>hw03_dplyr-manipulation_files/figure-html</code>	Gave names to code chunks with figures	15 days ago
<code>README.md</code>	Fixed typos in links to hw03 README	16 days ago
<code>hw03_dplyr-manipulation.R</code>	Generated R script using knitr::purl	15 days ago
<code>hw03_dplyr-manipulation.html</code>	Gave names to code chunks with figures	15 days ago
<code>hw03_dplyr-manipulation.md</code>	Gave names to code chunks with figures	15 days ago
<code>hw03_dplyr-manipulation.rmd</code>	Gave names to code chunks with figures	15 days ago

Below the file list, there's a section titled `README.md` which contains the following text:

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

Comma (.csv) and tab (.tsv) delimited files are automatically rendered nicely in GitHub repositories

Example: some Lord of the Rings data

jennybc / **lotr** Unwatch 1 Star 0 Fork 1

branch: master **lotr / lotr_clean.tsv** Raw

jennybc 2 months ago Add early exploration/cleaning

1 contributor

file | 684 lines (683 sloc) | 42.64 kb Open Edit Raw Blame History Delete

Search this file...

	Film	Chapter	Character	Race	Words
1	The Fellowship Of The Ring	01: Prologue	Bilbo	Hobbit	4
2	The Fellowship Of The Ring	01: Prologue	Elrond	Elf	5
3	The Fellowship Of The Ring	01: Prologue	Galadriel	Elf	460
4	The Fellowship Of The Ring	02: Concerning Hobbits	Bilbo	Hobbit	214
5	The Fellowship Of The Ring	03: The Shire	Bilbo	Hobbit	70
6	The Fellowship Of The Ring	03: The Shire	Frodo	Hobbit	128
7	The Fellowship Of The Ring	03: The Shire	Gandalf	Wizard	197
8	The Fellowship Of The Ring	03: The Shire	Hobbit Kids	Hobbit	10

Note the contributions to STAT 545 materials from one prof, 3 TAs, and one kind soul from the internet

Commits · STAT545-UBC/STAT545-UBC.github.io

github.com/STAT545-UBC/STAT545-UBC.github.io/commits/master

jennybc authored on Jan 5

updating links to shiny material as referenced in STAT545-UBC/Instruc... [...](#) [ec5ab6b](#) [🔗](#)

joolia authored on Jan 5

add info about fall 2015 to faq [...](#) [95f04b2](#) [🔗](#)

jennybc authored on Jan 5

Commits on Jan 1, 2015

add a resource for difference btwn Depends and Imports in packages fr... [...](#) [1db68d3](#) [🔗](#)

daattali authored on Jan 1

Commits on Dec 30, 2014

Merge pull request #15 from zhaoy/master [...](#) [9feb137](#) [🔗](#)

BernhardKonrad authored on Dec 30, 2014

Commits on Dec 29, 2014

Removed redundant getwd() [...](#) [8af9cc6](#) [🔗](#)

zhaoy authored on Dec 29, 2014

Fixing typo [...](#) [f306f45](#) [🔗](#)

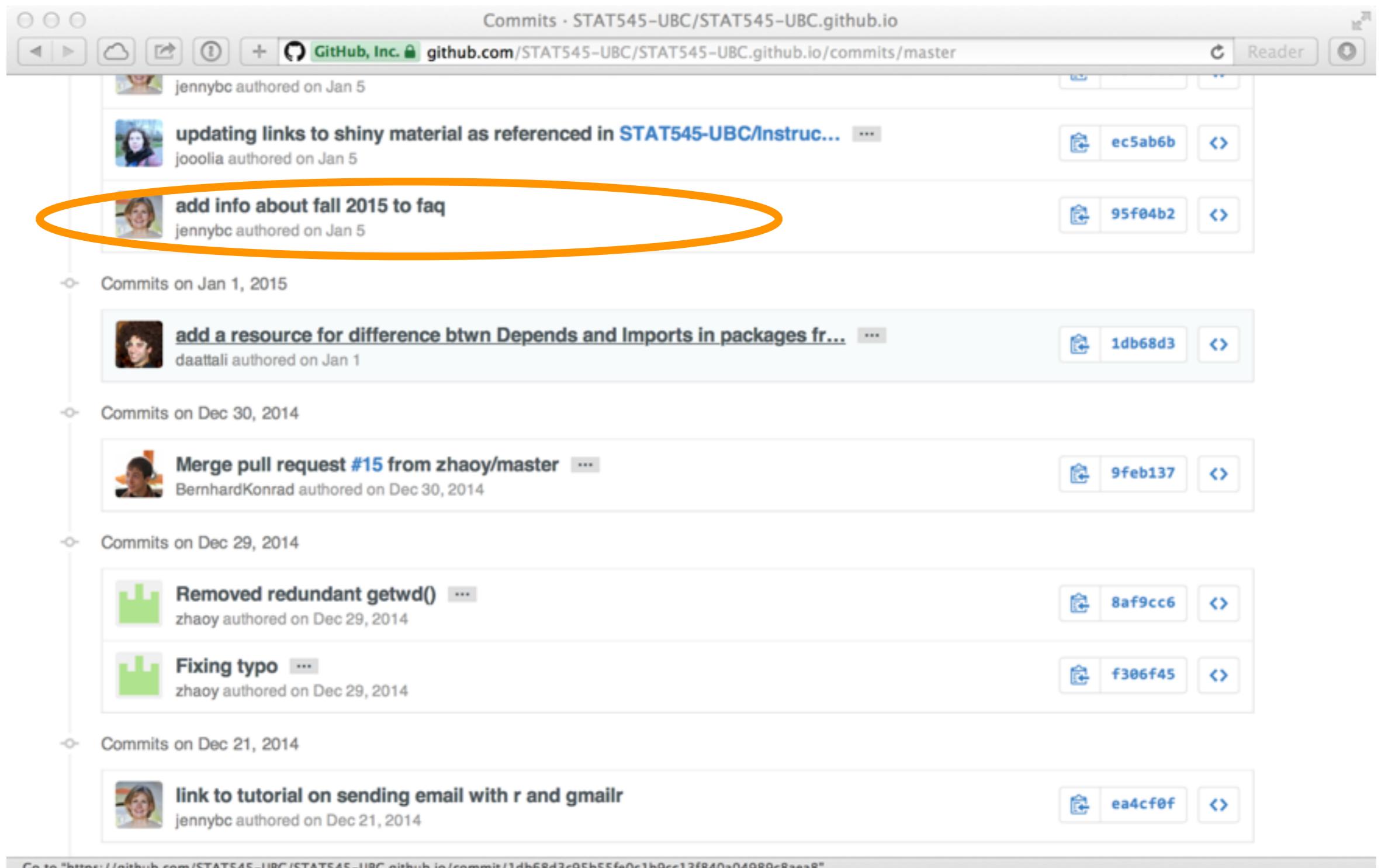
zhaoy authored on Dec 29, 2014

Commits on Dec 21, 2014

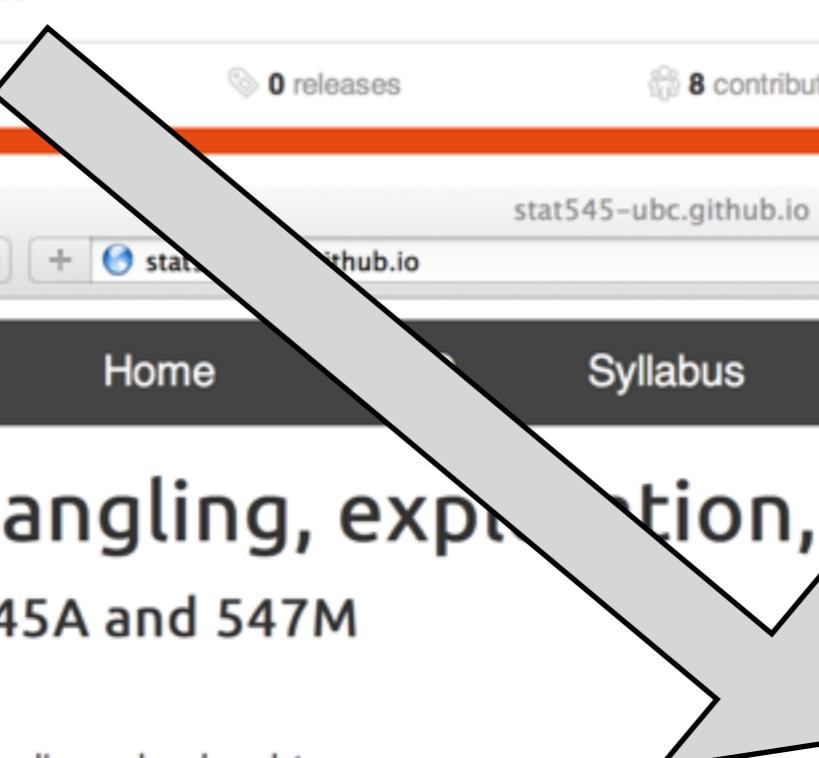
link to tutorial on sending email with r and gmailr [...](#) [ea4cf0f](#) [🔗](#)

jennybc authored on Dec 21, 2014

Go to "<https://github.com/STAT545-UBC/STAT545-UBC.github.io/commit/1db68d3c95b55fe0c1b9cc13f840a04989c8aea8>"



When prof or TA pushes to repo, website updates!



STAT545-UBC/STAT545-UBC.github.io

GitHub, Inc. github.com/STAT545-UBC/STAT545-UBC.github.io

This repository Search Explore Gist Blog Help jennybc + ⌂

STAT545-UBC / STAT545-UBC.github.io

Unwatch 16 Star 34 Fork

Main repository for STAT 545 @ University of British Columbia, a course in data wrangling, exploration, and analysis with R. <http://stat545-ubc.github.io> — Edit

495 commits 2 branches 0 releases 8 contributors

branch: master

typo/correction jennybc authored 2 hours ago

automation01_slides automation10_holding-area block002_hello-r-workspace

stat545-ubc.github.io

STAT 545 Home Syllabus Topics People

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

Selected topics

- Introduction to R and the RStudio IDE; scripts, the workspace, RStudio Projects
- Generate reports from R scripts and R Markdown
- Care and feeding of data in R
- Data aggregation; “apply” functions, `plyr`, `dplyr`
- Data visualization with `ggplot2`
- Graphs and descriptive stats for quantitative and categorical variables

Files in a Git repo, even one hosted on GitHub, still reside on your computer

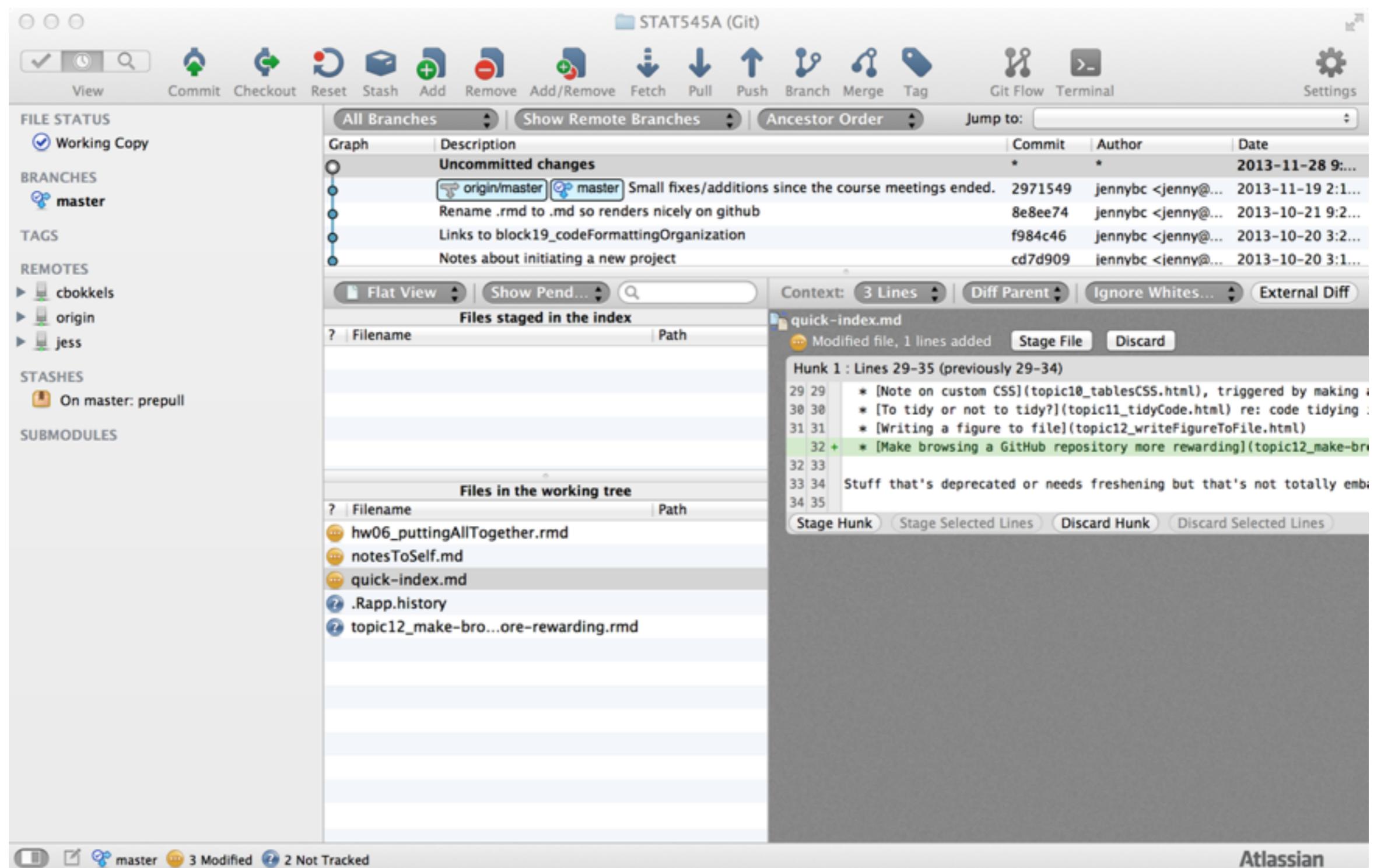
Browse and edit them all you want

Git has commands for communicating with the remote repository, e.g. the GitHub repo (push, pull, fetch, clone)

I highly recommend using a Git GUI on your computer for making commits, syncing with the remote, etc.

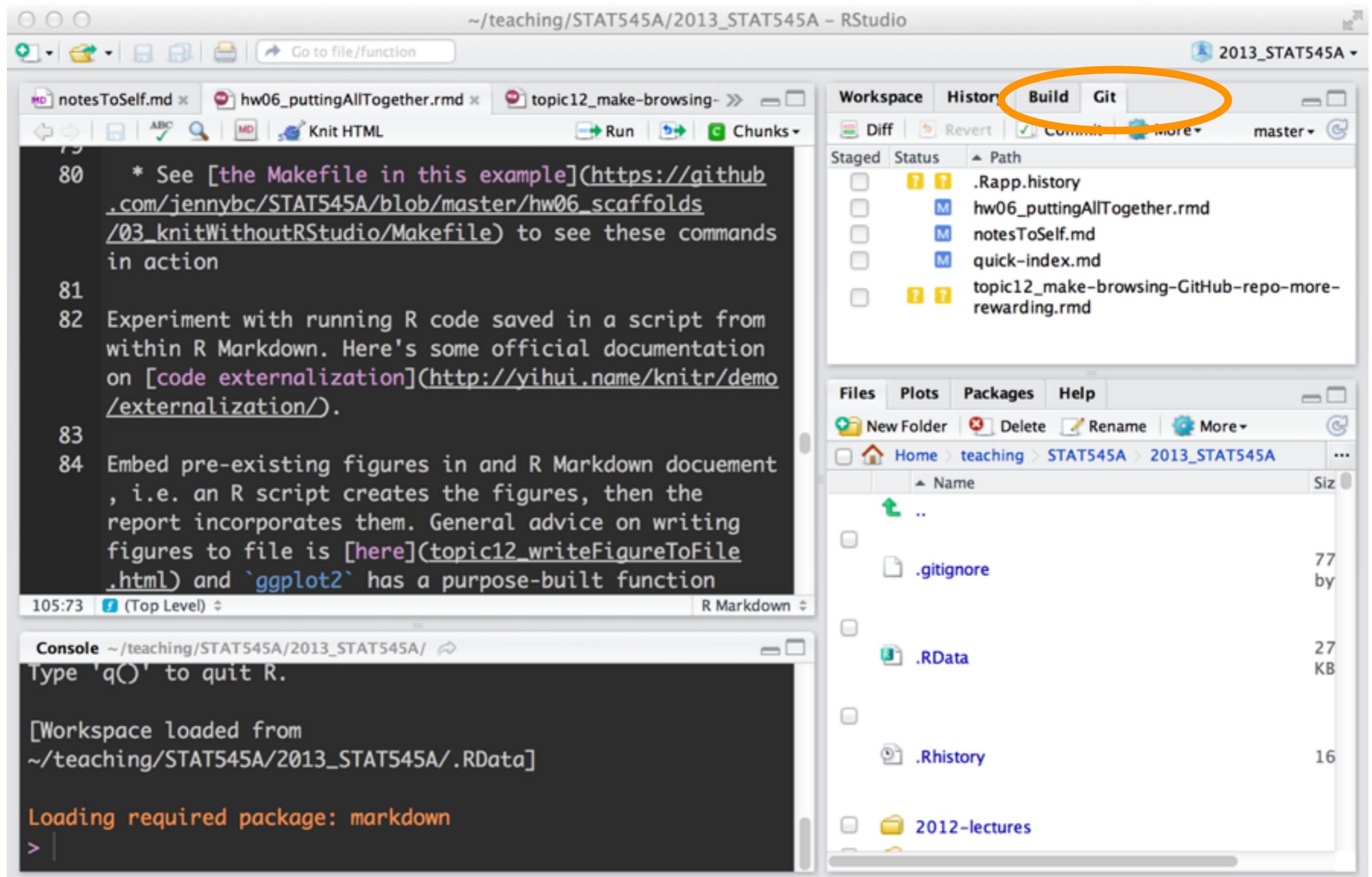
Reconciling and merging changes when two people make conflicting commits is not fun, but better than the alternatives

I recommend SourceTree, a free Git client for Windows and Mac.



RStudio can also act as your Git(Hub) client

http://www.rstudio.com/ide/docs/version_control/overview



Big picture, second half:

sane file and project management is good
that's what version control does

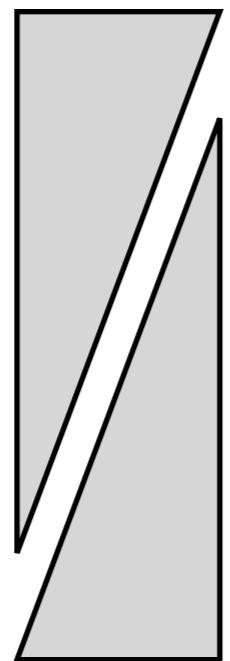
distributed file management is good
excellent for 2+ people collaborating

ability to browse something on the web is unreasonably
powerful

Git + GitHub provide a compelling solution for
collaborative file wrangling; (R) Markdown and RStudio
play well with Git(Hub)

STAT 545 = I semester, 3 contact hours/wk

R markdown



Git(Hub)

Data wrangling, cleaning, munging
↔
Visualization
(R chops, in general)

8 weeks



Automation & pipelines



R packages

4 weeks

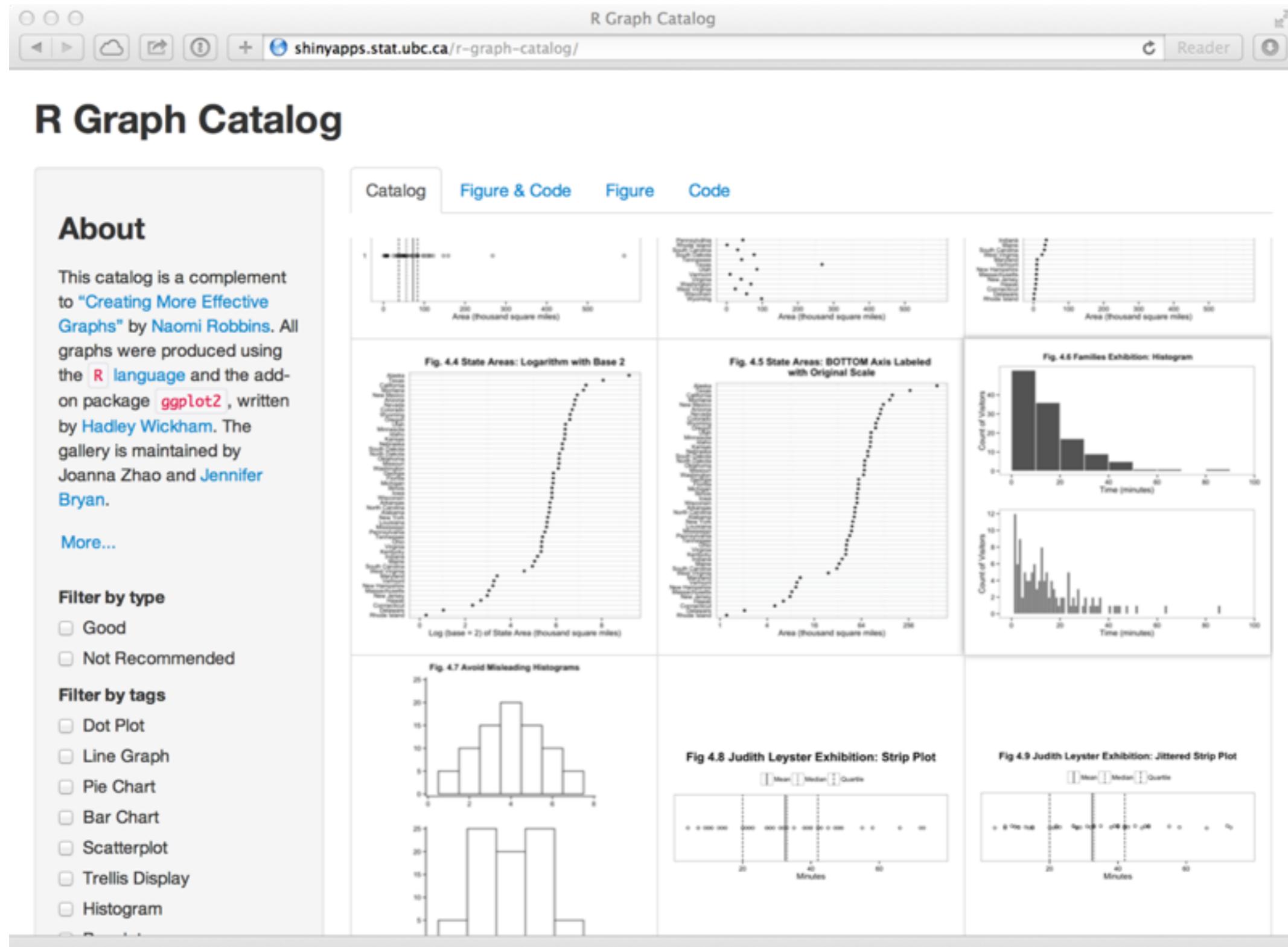


Shiny



Web APIs and scraping

<http://shinyapps.stat.ubc.ca/r-graph-catalog/>
<https://github.com/jennybc/r-graph-catalog>



Bottom line: do something deliberate that has a good hassle: result ratio for you.

Be open to upgrading your approach as time goes on.

Keep your eyes and ears open re: new developments.