## Easy package development Hadley Wickham

Assistant Professor / Dobelman Family Junior Chair Department of Statistics / Rice University





### HELLO my name is

## Hadley



Winston Chang
Research Associate
Department of Statistics
Rice University

#### Caveats

- Opinionated advice not everyone agrees that this is the right way to build packages.
- Don't look at ggplot2 (yet)
- This class only scratches the surface many more details to learn.

### Outline

- Package basics: devtools
- Documentation: roxygen2
- Testing: testthat
- Releasing your package: devtools

### Getting Started

```
# Check that you're ready to go
library(devtools)
has_devel()
# If not:
    * on windows: install Rtools
#
    * on mac: install xcode
#
    * on linux: install gcc, make
library(testthat)
library(roxygen2)
```

### Packages to look at

- devtools, evaluate, lubridate, reshape2, stringr, testthat: source code for my packages
- coin, Matrix: illustrate other important parts of packages that I don't use
- one-file, minimal, ...: various stages of a package you'll be developing today

## Learn from others

Read the source of other packages!

https://github.com/hadley/plyr

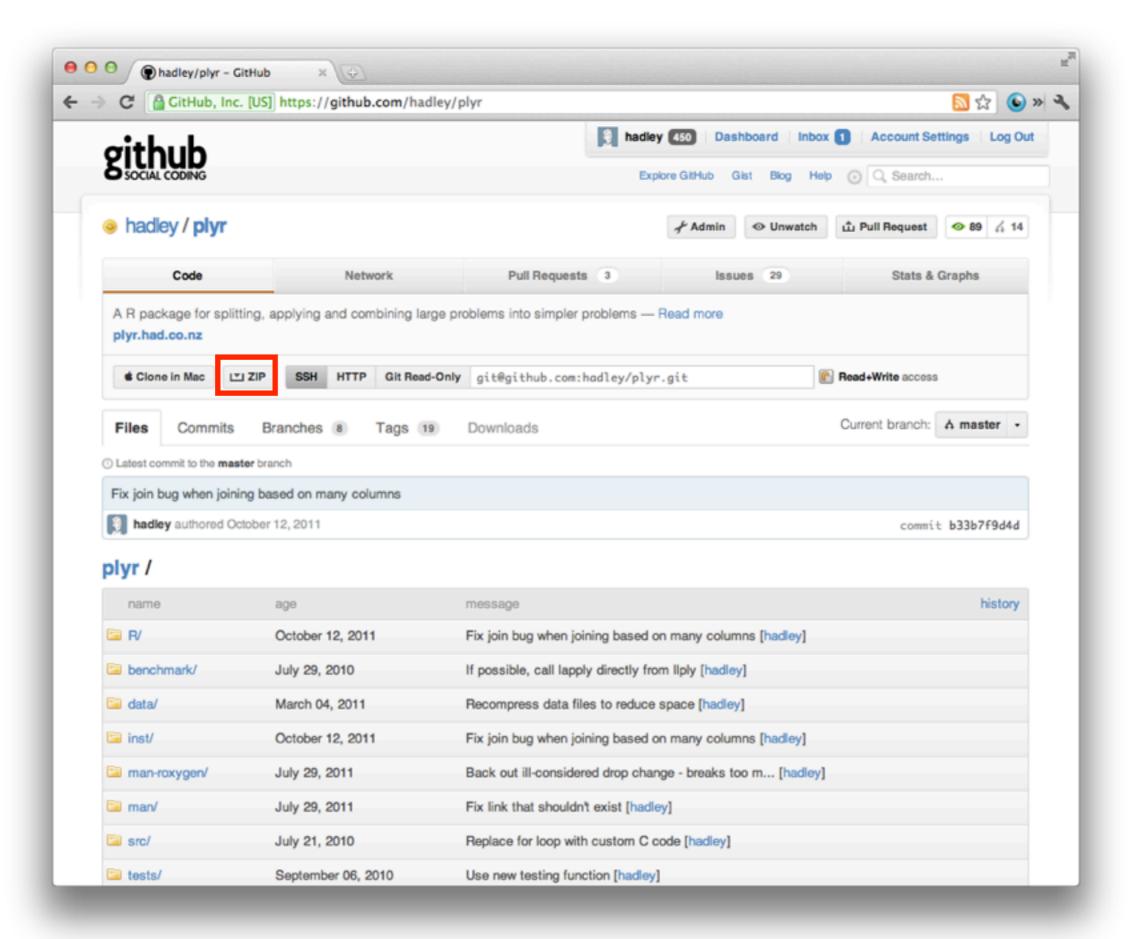
https://github.com/hadley/stringr

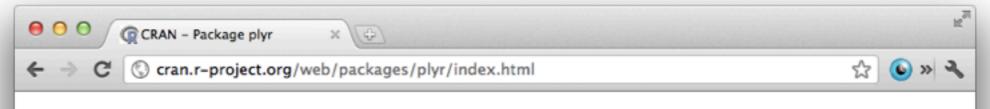
https://github.com/hadley/devtools

https://github.com/hadley/lubridate

https://github.com/hadley/evaluate

https://github.com/hadley/reshape





#### plyr: Tools for splitting, applying and combining data

plyr is a set of tools that solves a common set of problems: you need to break a big problem down into manageable pieces, operate on each pieces and then put all the pieces back together. For example, you might want to fit a model to each spatial location or time point in your study, summarise data by panels or collapse high-dimensional arrays to simpler summary statistics. The development of plyr has been generously supported by BD (Becton Dickinson).

Version: 1.6

Depends:  $R (\geq 2.11.0)$ 

Imports: <u>itertools</u>, <u>iterators</u>

Suggests:  $\underline{abind}$ ,  $\underline{testthat}$  ( $\geq 0.2$ ), tcltk,  $\underline{foreach}$ 

Published: 2011-07-29

Author: Hadley Wickham

Maintainer: Hadley Wickham < h.wickham at gmail.com>

License: MIT

URL: <a href="http://had.co.nz/plyr">http://had.co.nz/plyr</a>
Citation: <a href="plyr">plyr citation info</a>

CRAN checks: plyr results

#### Downloads:

Package source: plyr 1.6.tar.gz

MacOS X binary: plyr 1.6.tgz
Windows binary: plyr 1.6.zip

Reference manual: plyr.pdf

News/ChangeLog: NEWS

Old sources: plyr archive

### Your turn

Download and extract the source code for one R package that you use regularly. Read the source code for the function from that package that use most regularly. What did you learn?

# Working directory

### Why?

All paths in R are relative to the working directory. Life is much easier when you have it correctly set.

Usually want one project per directory. (See also Rstudio's project support)

Makes code easy to move between computers. **Never** use setwd() in a script.

### How?

Rstudio: Tools | Set working directory | Choose directory ... (^ 企K)

Windows: File | Change dir. For frequent use, make shortcut in that folder.

Mac: 郑D

Terminal: start R from the desired

directory

```
# Find out what directory you're in
getwd()

# List files in that directory
dir()
```

### Your turn

Make sure your working directory is set to the rv directory inside the code and data directory you downloaded. Use dir() to check you're in the right place.

### Projects

- RStudio's projects make this even easier: double click on Rproj file to set working directory and restore state
- Also enables project search

# Discrete random variables

### Definitions

A random variable is a random experiment with a numeric sample space. Usually given a capital letter like X, Y or Z.

(More formally a random variable is a function that converts outcomes from a random experiment into numbers)

The space (or support) of a random variable is the range of the function (cf. sample space)

$$P(a < X < b) = \sum_{x_i \in (a,b)} f(x_i)$$
$$P(X = x) = f(x)$$

probability pmf

To be a pmf, f must satisfy two conditions:

$$\sum_{x_i \in S} f(x_i) = 1$$

$$f(x_i) \geq 0, \forall x_i \in S$$

### Transformations

X	-1	0	1	2	3
f(x)	0.2	0.1	0.3	0.1	0.3

Let X be a discrete random variable with pmf f as defined above.

Write out the pmfs for:

$$A = X + 2$$
  $B = 3X$   $C = X^2$   $D = 0 * X$ 

### Mean & variance

The mean summarises the "middle" of the distribution. The variance summarises the "spread" of the distribution.

Mean = E(X) = "Sum" of all outcomes, weighted by their probability.

Variance =  $Var(X) = E[(X - E[X])^2)] =$ expected squared distance from mean

### Expectation

Expectation is a linear operator:

Expectation of a sum = sum of expectations (additive)

Expectation of a constant \* a function = constant \* expectation of function (homogenous)

Expectation of a constant is a constant.