

Project Immortality

Using GitHub To Make Your Work Live Forever

Tan Ho

July 27, 2022

The Life and Times of a Machine Learning Project

Import some data 

Do some wrangling 

Throw it into a model 

Tweet a plot 

{...most projects die here} 

NEVER HAVE I FELT SO
CLOSE TO ANOTHER SOUL
AND YET SO HELPLESSLY ALONE
AS WHEN I GOOGLE AN ERROR
AND THERE'S ONE RESULT
A THREAD BY SOMEONE
WITH THE SAME PROBLEM
AND NO ANSWER
LAST POSTED TO IN 2003

WHO WERE YOU,
DENVERCODER9?

WHAT DID YOU SEE?!



Sometime later...



Who This
Talk Is
For





About Me

Self-taught R developer

Hobby NFL analysis ►
data science career

Maintain nflverse/ffverse
R packages and data
pipelines

ffopportunity

- Uses NFL play by play data
- Expected Fantasy Points
 - Measures the value of player opportunities in fantasy football

github.com/ffverse/ffopportunity





Imports nflverse play by play data

Does some data wrangling

Trained an xgboost model

Generated some predictions

...now what?

How can I make fportunity **live on**?

Can someone run my code and build on this project?

Can new predictions be automatically generated?

Where can I store data?





Can someone
run my code
and build on it?



Make It a Package

- Not (necessarily) for CRAN
- Add a DESCRIPTION
- Wrap code in functions

Can new
predictions be
automatically
generated?



Add GitHub Actions

```
# .github/workflows/update.yml
on:
  schedule:
    # At 4:05 every Mon, Tues, and Fri from Sep-Jan
    - cron: 5 4 * 9-12,1 5,1,2
  workflow_dispatch:

name: ep-update-data

jobs:
  ep-update-data:
    runs-on: ubuntu-latest
    env:
      GITHUB_PAT: ${ secrets.GITHUB_TOKEN }
    steps:
      - uses: actions/checkout@v2
      - uses: r-lib/actions/setup-r@v2
      - uses: r-lib/actions/setup-r-dependencies@v2
        with:
          extra-packages: piggyback, arrow, readr
      - name: Run data update
        run: |
          source("update/ep_update.R")
        shell: Rscript
```



Where can I
store the data?



Three Common Problems with GitHub Data Repositories

File Size Limits

Inefficient binary data storage

Commit history chaos

Blind version control is bad!



Tan
@_TanHo

blindly versioning binary files is a recipe for getting rekt by your own automated data repo

data: 60 MB uncompressed

.git version storage: 3.7 GB compressed and growing daily (60x size!)

innocent maintainer trying to update code locally:

(ಠ_ಠ)ಽ┐┌

WinDirStat - WinDirStat

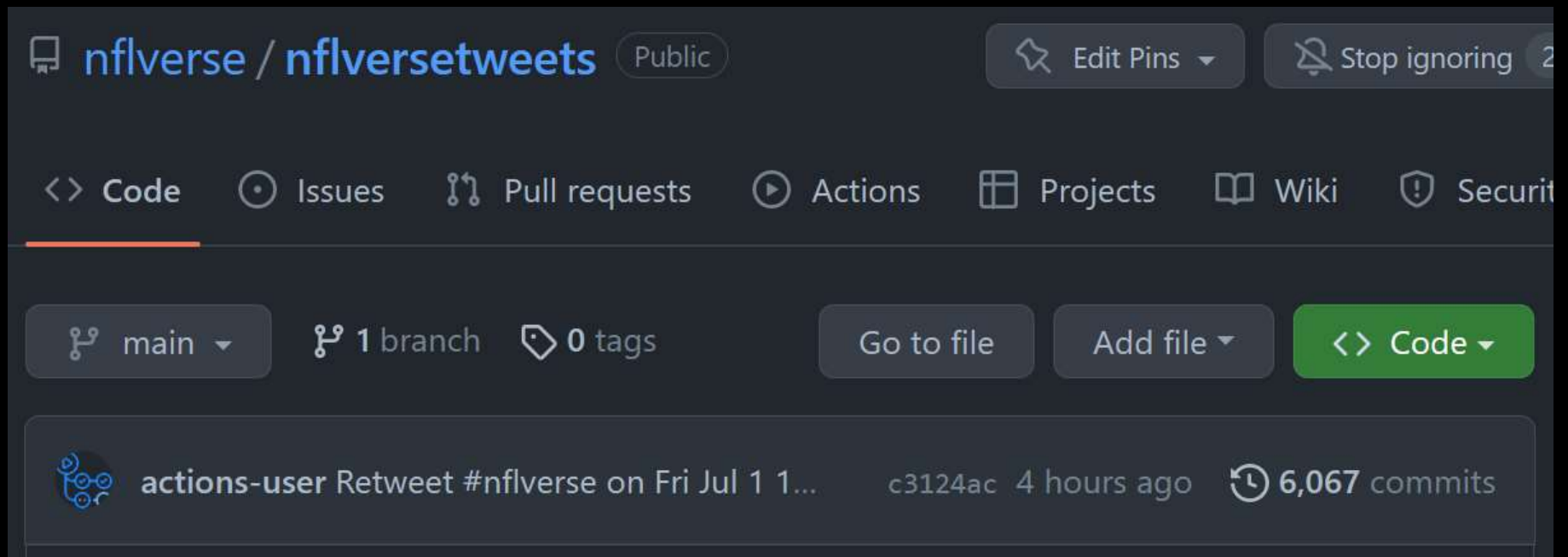
Report Options Help

	Subtree Percent...	Perce...	> Size	Items	Files	Subdirs
\GitHub\nflfastR-roster	<div></div>	[0:00 s]	3.8 GB	227	191	36
	<div></div>	98.5%	3.7 GB	47	32	15
	<div></div>	1.5%	59.1 MB	127	126	1
	<div></div>	0.0%	470.0 KB	10	10	0
	<div></div>	0.0%	7.7 KB	22	9	13
	<div></div>	0.0%	5.5 KB	8	8	0
	<div></div>	0.0%	5.2 KB	6	5	1
	<div></div>	0.0%	0	1	1	0

ALT

10:50 AM · Mar 10, 2022 · Twitter Web App

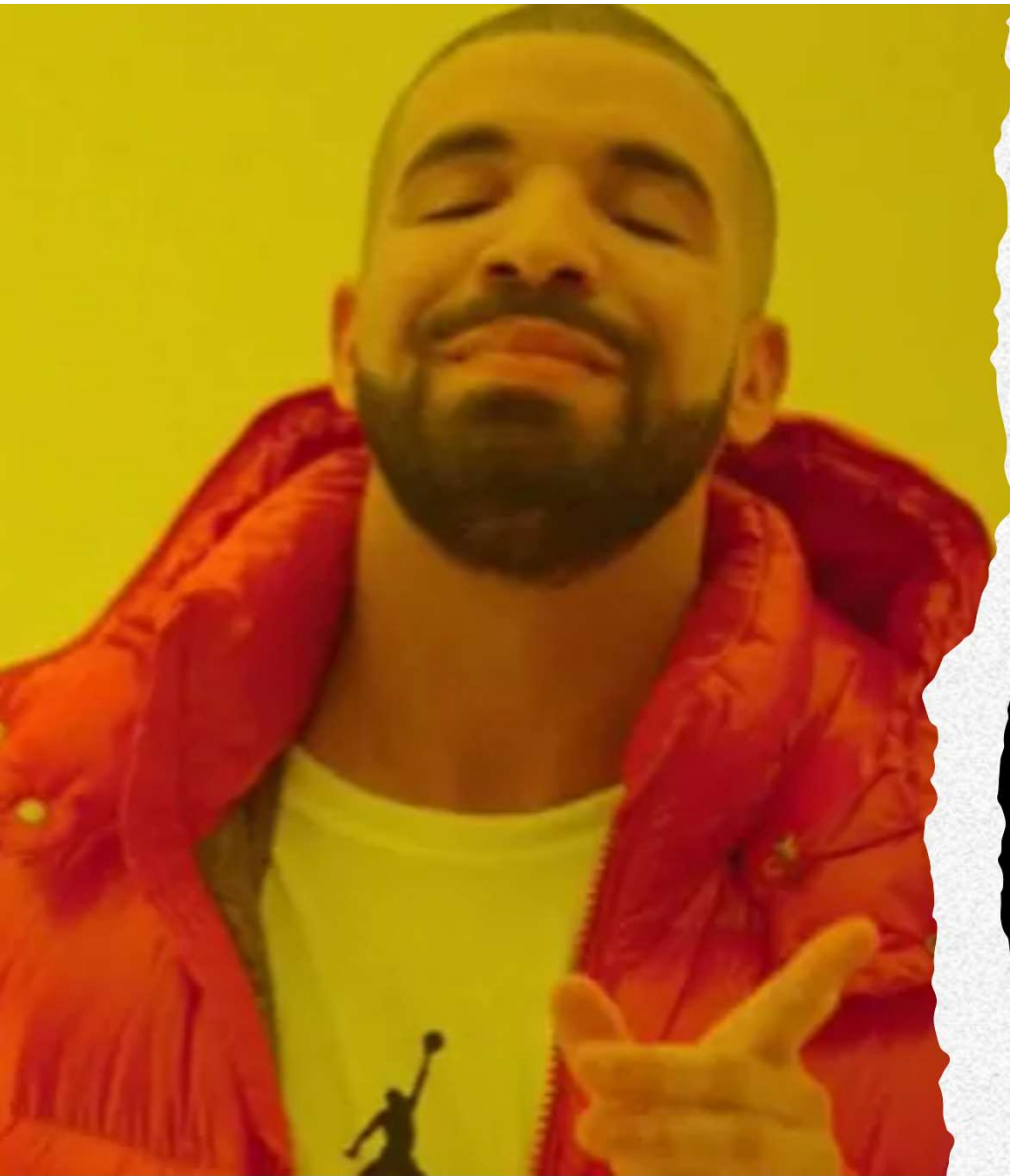
GitHub Data Repo Problems: Commit History Chaos!





Meh Solutions

- Git Large File Storage
- Amazon S3 buckets
- Dropbox (?)



Better Solution

GitHub Releases!



GitHub Releases

- Friendly file size limits
- Versioning by choice
- Keeps commit history clean
- Data stays with your project



{piggyback}

```
library(piggyback)
repo <- "ffverse/ffoportunity"
tag <- "data-2022"

pb_new_release(repo = repo,
               tag = tag)

pb_upload("ep_2022-week1.rds",
          repo = repo,
          tag = tag)

# download all data from release
pb_download(repo = repo,
            tag = tag)
```

Recap

The image features a solid black background. In the lower portion, there is a horizontal band of white, textured material that looks like torn paper or a rough cut. Below this white band is a dark grey, wavy, and irregular shape that resembles a stylized horizon line or a piece of fabric. The word "Recap" is written in a white, serif font in the upper left area of the black background.



Project Immortality With GitHub



GitHub
R Packages

Install and run



GitHub
Actions

Automation



GitHub
Releases

Store data with
your project

Resources

- Packages
 - R Packages book by Hadley Wickham & Jenny Bryan
<https://r-pkgs.org>
- GitHub Actions
 - r-lib/actions <https://github.com/r-lib/actions/>
 - GHA with R book https://orchid00.github.io/actions_sandbox/
- GitHub Releases
 - piggyback pkg <https://github.com/ropensci/piggyback>
- ffopportunity <https://github.com/ffverse/ffopportunity>