

Project Immortality

Using GitHub To Make Your Work Live Forever

Tan Ho [@_TanHo]

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...



How can **your project**
help this person?



Who This
Talk Is
For





About Me

Self-taught R:
hobbyist football analysis

Led to a data science career

Maintain public NFL data &
nflverse/ffverse R packages



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

Wrangles some features

Trains an xgboost model

Predicts fantasy points

...now what?

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

Can someone use (and improve on) my model?

Can new predictions be automated every week?

Where can I make predictions accessible?





Can someone use
(and improve on)
my model?



Make It a Package

Not (necessarily) for CRAN

- Add a DESCRIPTION
- Wrap code in functions



Add a DESCRIPTION

- Specifies a LICENSE
 - Easy dependency install
- usethis does everything!



```
library(usethis)
```

```
use_description() # use template
```

```
use_mit_license() # add license of choice
```

```
use_package() # add dependencies
```

```
use_latest_dependencies() # specify version of dependencies you used
```



Wrap logic into functions

- Convert hardcoded variables to arguments
 - (Ctrl-Alt-X ⚡)
- Add some usage notes

Goal: Make it easy for users to run

Can new
predictions be
automated
every week?



Automate Predictions with 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 make
predictions
accessible?



Three Common Problems with GitHub Data Repositories

File Size Limits

Inefficient binary data storage

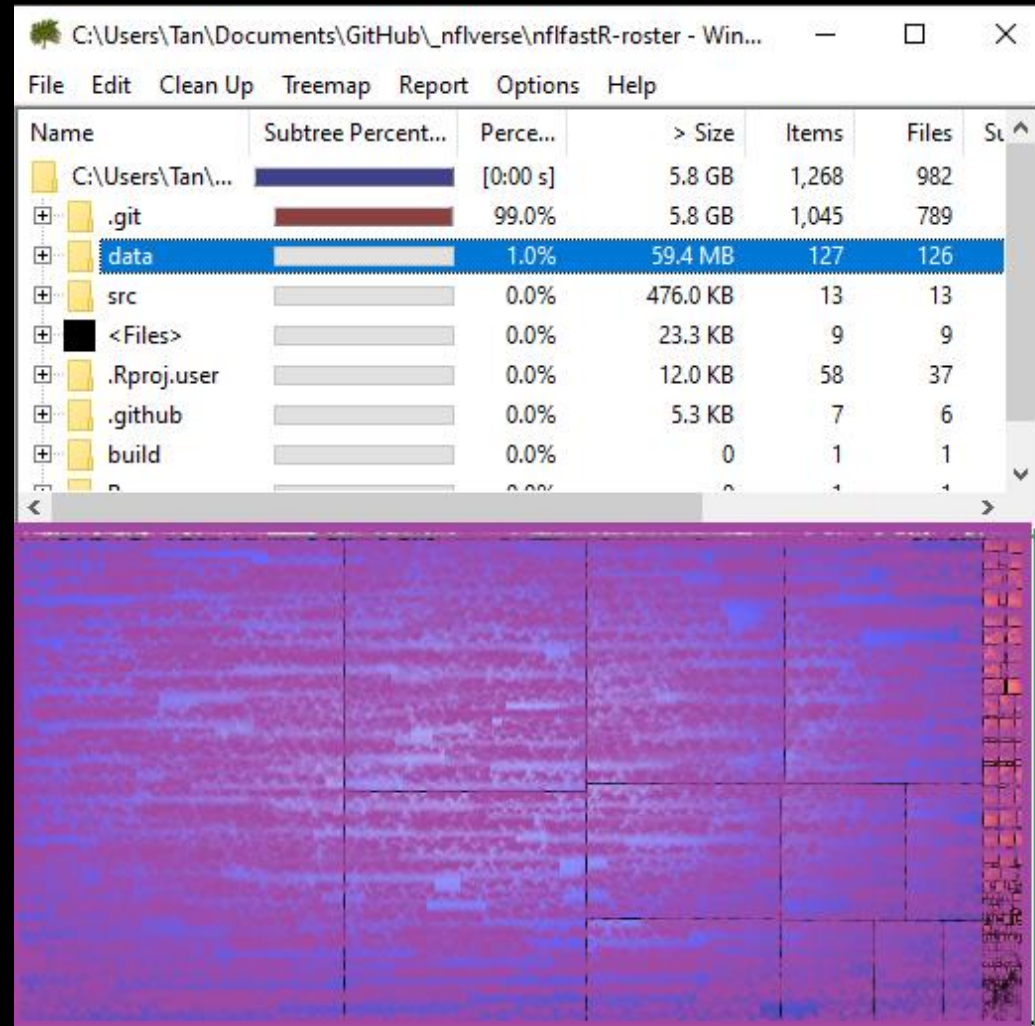
Commit history chaos

GitHub File Size Limits

- **75MB** per file
- **2GB** per repo



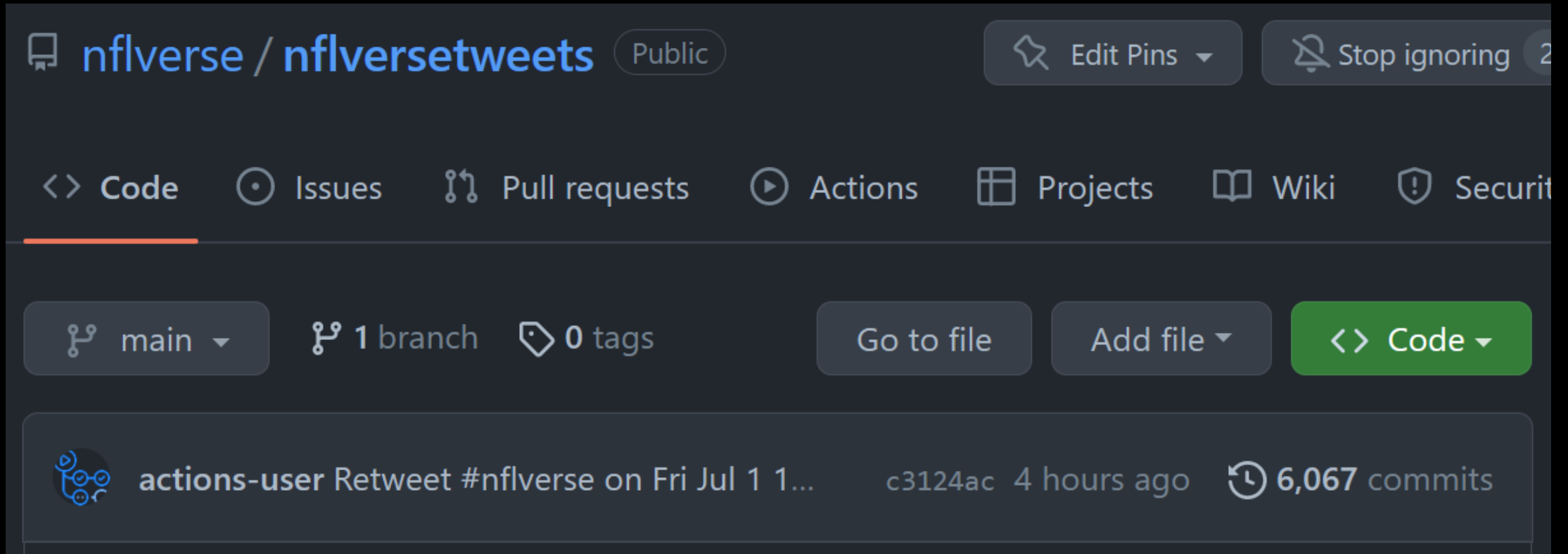
Version control is bad at binary files!



actual
data:
60MB

git: 6GB

GitHub Data Repo Problems: Commit History Chaos!





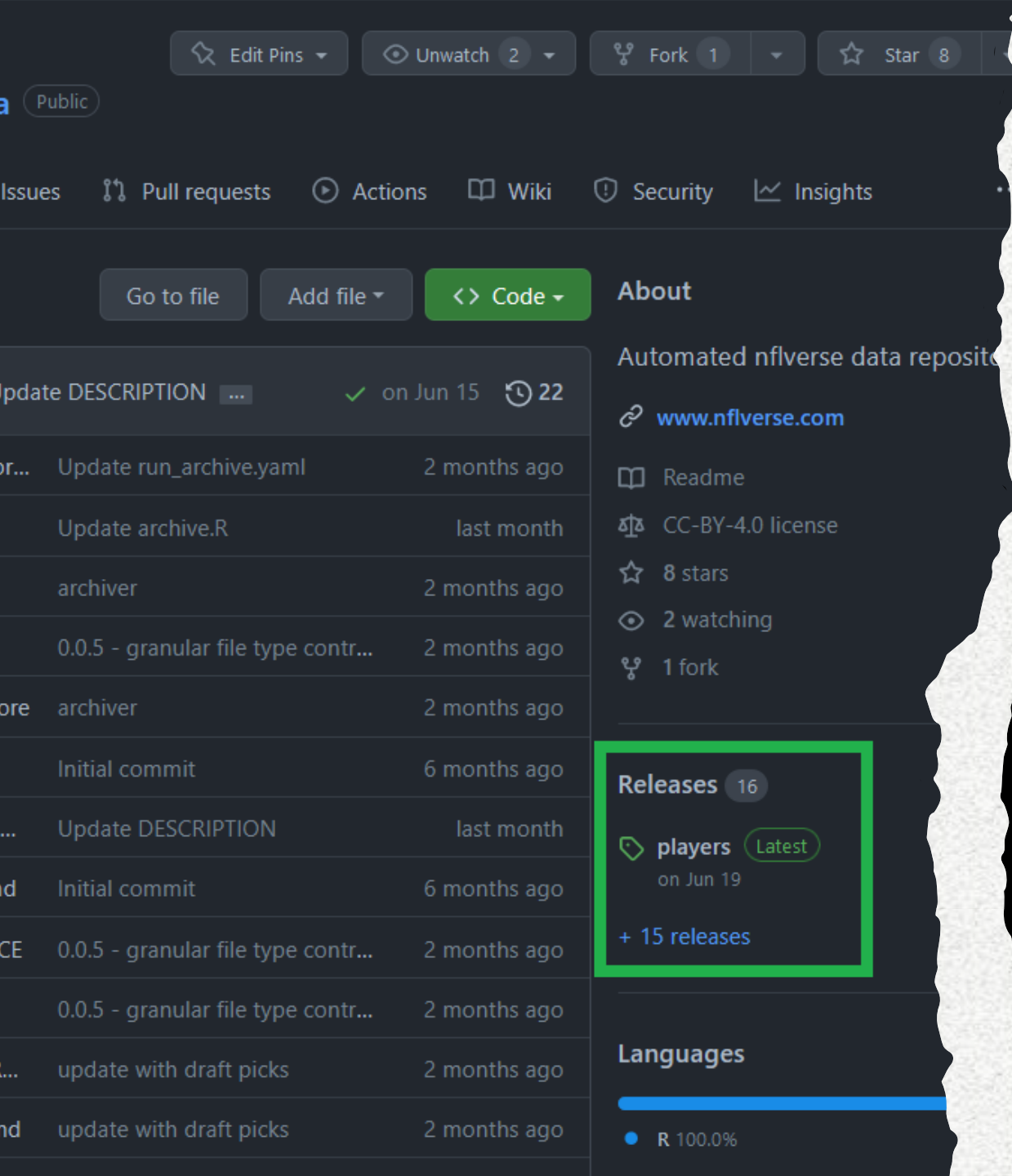
Meh Solutions

- Git Large File Storage (LFS) 🙄
- Amazon S3 buckets 📦
- Dropbox 🐼



Better Solution

GitHub Releases!



GitHub Releases

- Make a new release
- Upload files
- Update as desired

GitHub Releases

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





{piggyback}

```
library(piggyback)

repo <- "ffverse/ffoportunity"
tag <- "data-2022"

pb_release_create(repo = repo, tag = tag)

pb_upload(file = "expected_points_week01.rds",
          repo = repo,
          tag = tag)

pb_download(file = "expected_points_week01.rds",
            dest = ".",
            repo = repo,
            tag = tag)
```

Recap

How can **my project** help someone after I've moved on from it?

Project Immortality With GitHub



GitHub-hosted
package

Install, run,
license



GitHub
Actions

Automation



GitHub
Releases

Store data with
your project

Thank
you!



Resources

- Packages
 - R Packages book <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>