Project Immortality

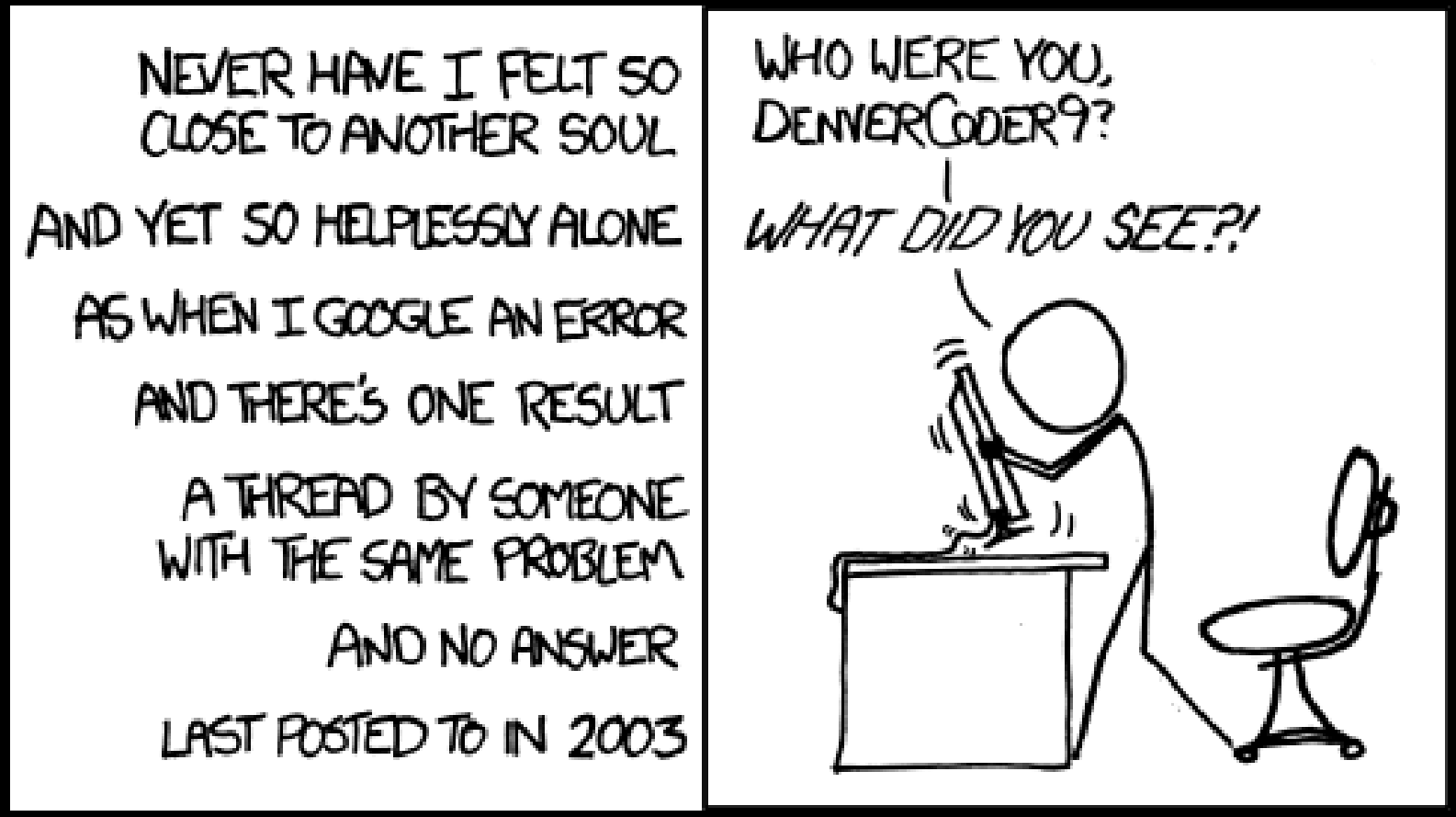
Using GitHub To Make Your Work Live Forever

Tan Ho [@_TanHo]

https://tanho.ca/project-immortality

The Life and Times of a Data Science Project

- Import some data 📖
- Do some wrangling 🤠
- Throw it into a model 🧠
- Tweet a plot 🎉
- {...most projects die here} 💀



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

ffopportunity

- Imports nflverse play by play data
- Wrangles some features
- Trains an xgboost model
- Predicts fantasy points

...now what?

How can I make ffopportunity **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} has this covered!

```
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
  - Extract function (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

```yaml
# .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?

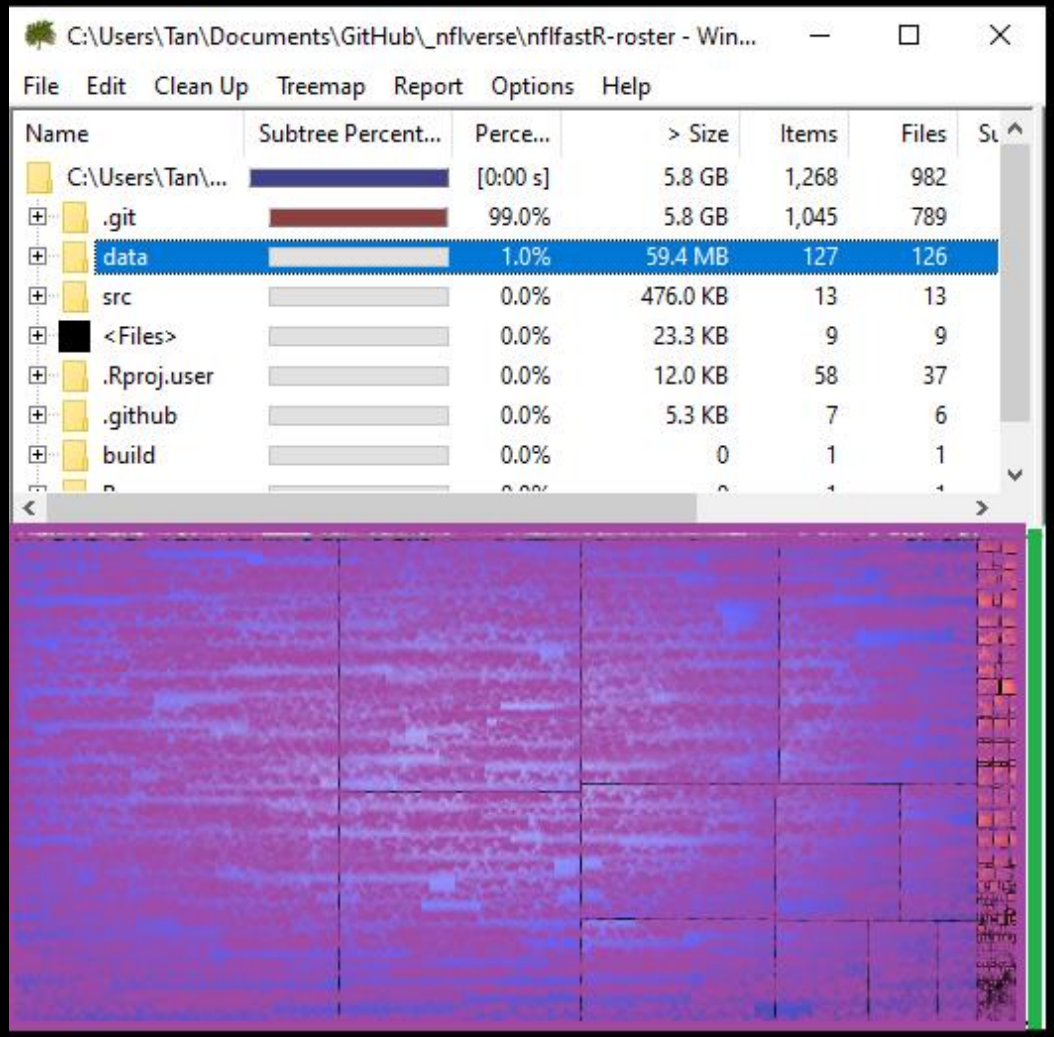
Three Common Problems with GitHub Data Repositories

- File Size Limits
- Inefficient binary data storage
- Commit history chaos

GitHub
File Size Limits

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

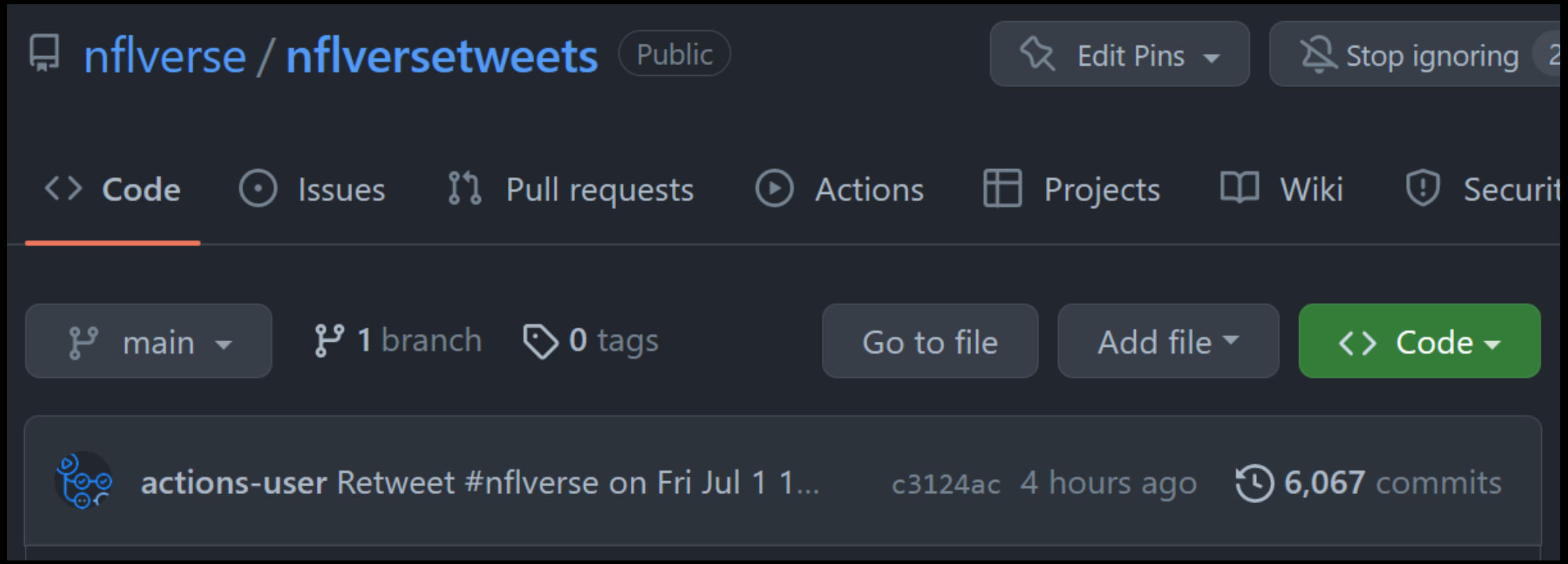
Version control is bad at binary files!



actual data: 60MB

git: 6GB

GitHub Data Repo Problems: Commit History Chaos!

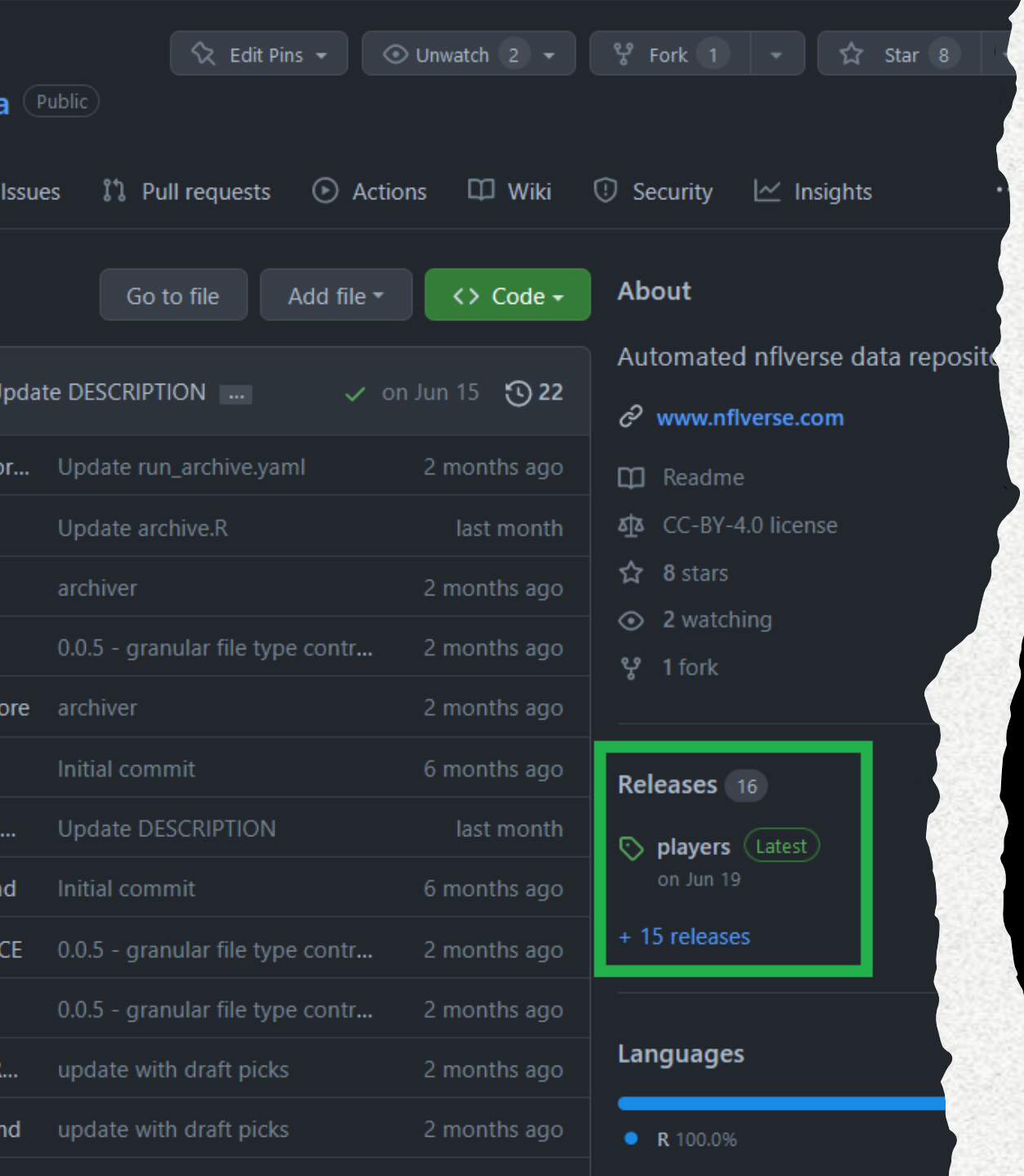


Meh Solutions

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

Better Solution

GitHub Releases!



GitHub Releases

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

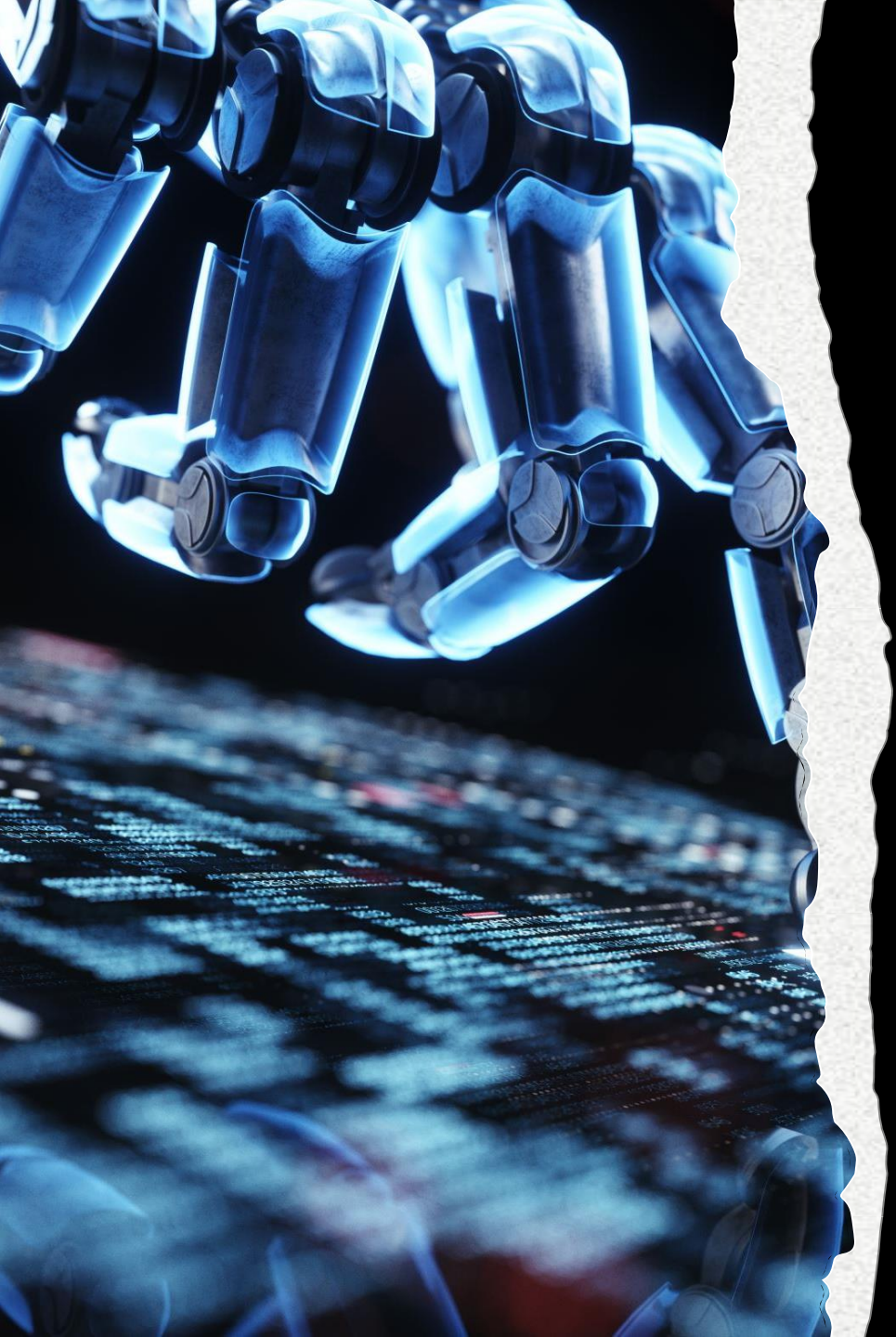
GitHub Releases

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

{piggyback}

```r
library(piggyback)

repo ← "ffverse/ffopportunity"
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



Three Tools for Project Immortality

- R Package Infrastructure
- GitHub Actions
- GitHub Releases

How can **your project** help others, now and into the future?

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