Reproducible, Open-Source, Transparent Data Analysis (and more) in Python and R

Rich Pauloo

PhD Candidate in Hydrogeology

University of California, Davis

About me

Hydrogeology PhD (expected March 2020),

University of California Davis *[Advisor: Dr. Graham Fogg]*

Domain interests: Groundwater quantity/quality, numerical modeling of groundwater flow and contaminant transport, impact of drought and climate change on disadvantaged populations, remote sensing.

Method interests: AI (machine & deep learning), spatial statistics, data science, reproducible research, interactive web application development, sensor networks, automation.

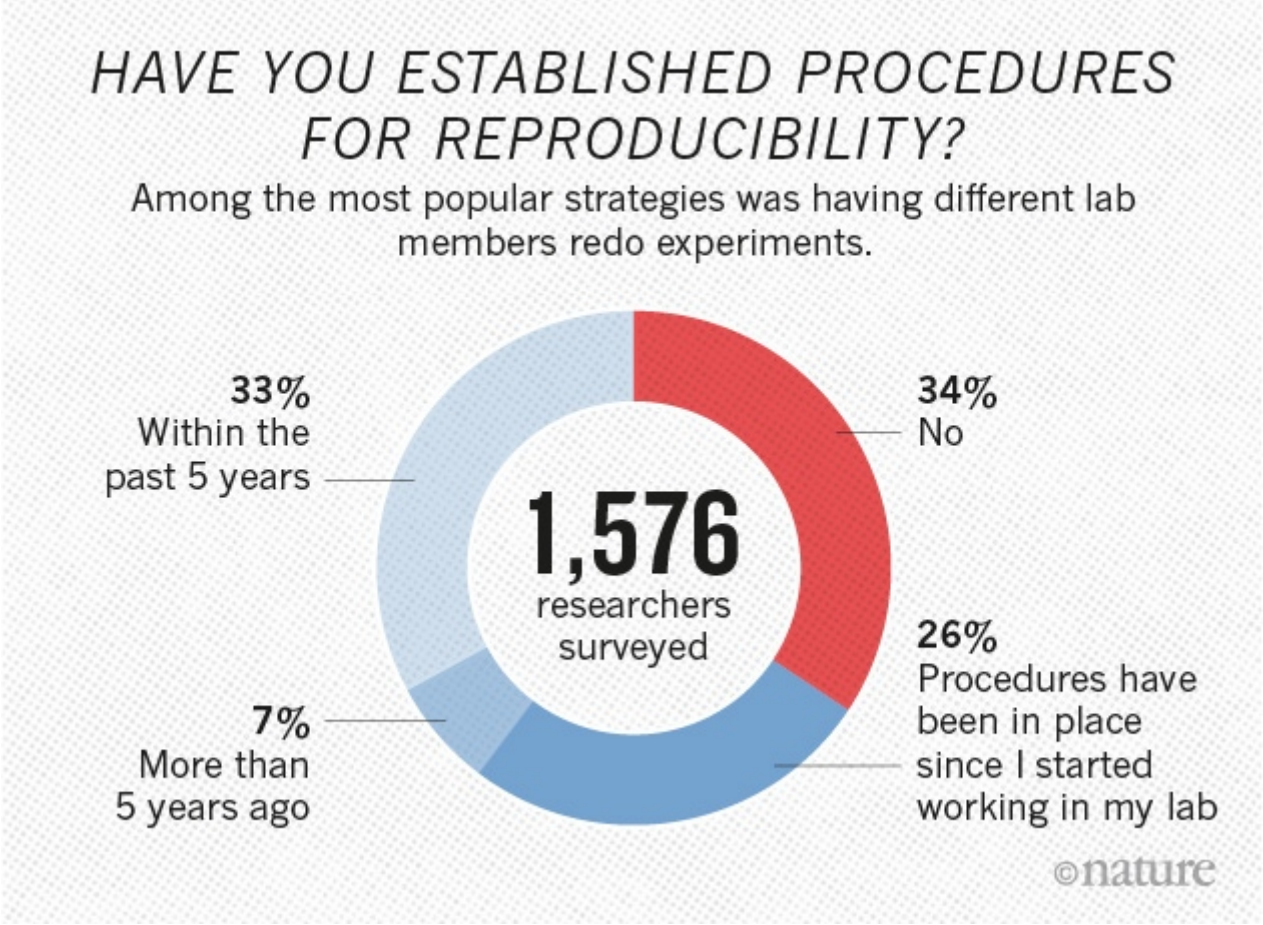
richpauloo.github.io @RichPauloo rpauloo@ucdavis.edu

Agenda

- Motivation
- Why Python and R?
- Jupyter and Github for sharing analysis and version control
- (and more) What else can I do with Python and R?
- Live Coding

Motivation

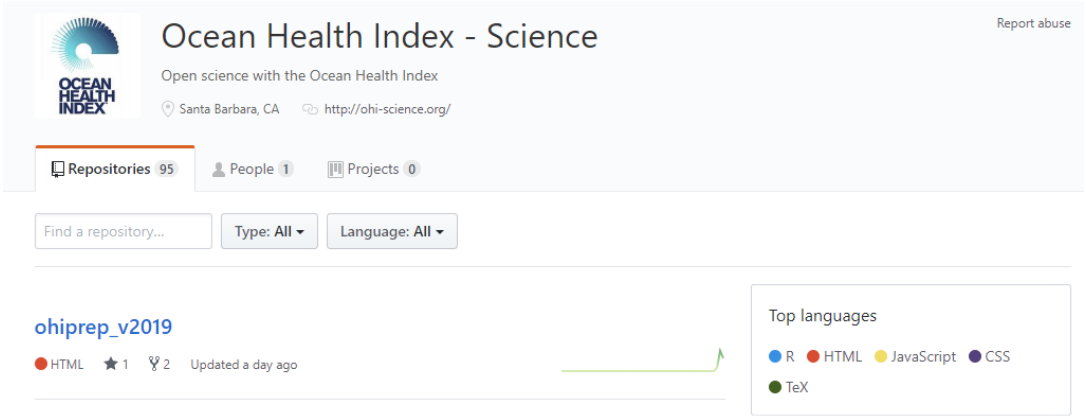
- Reproducibility crisis in science *(Baker, 2016; Munafò, 2017).*
  - More than 70% of researchers have tried and failed to reproduce another scientist's experiments, and more than half have failed to reproduce their own experiments.
- Towards reproducible, transparent, open-source analysis *(Kanterakis et al, 2018, Lowndes, 2017).*
- Ease of reproducibility and collaboration (including self).



(Baker, 2016)

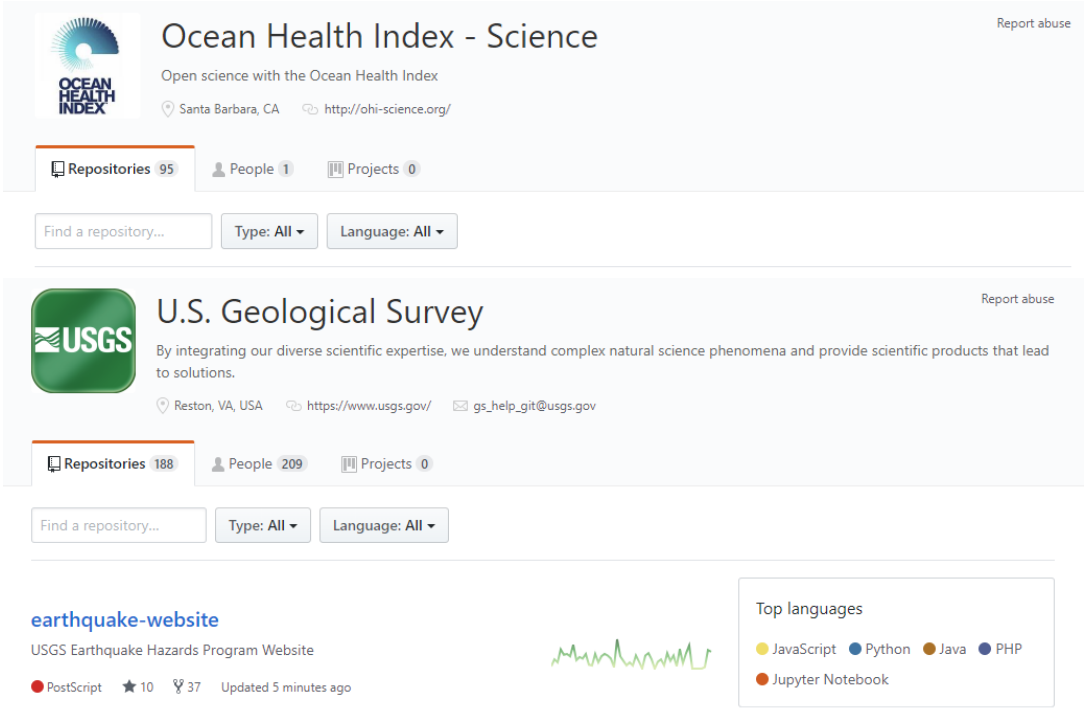
Motivation

- Reproducibility crisis in science *(Baker, 2016; Munafò, 2017).*
  - More than 70% of researchers have tried and failed to reproduce another scientist's experiments, and more than half have failed to reproduce their own experiments.
- Towards reproducible, transparent, open-source analysis *(Kanterakis et al, 2018, Lowndes, 2017).*
- Ease of reproducibility and collaboration (including self).



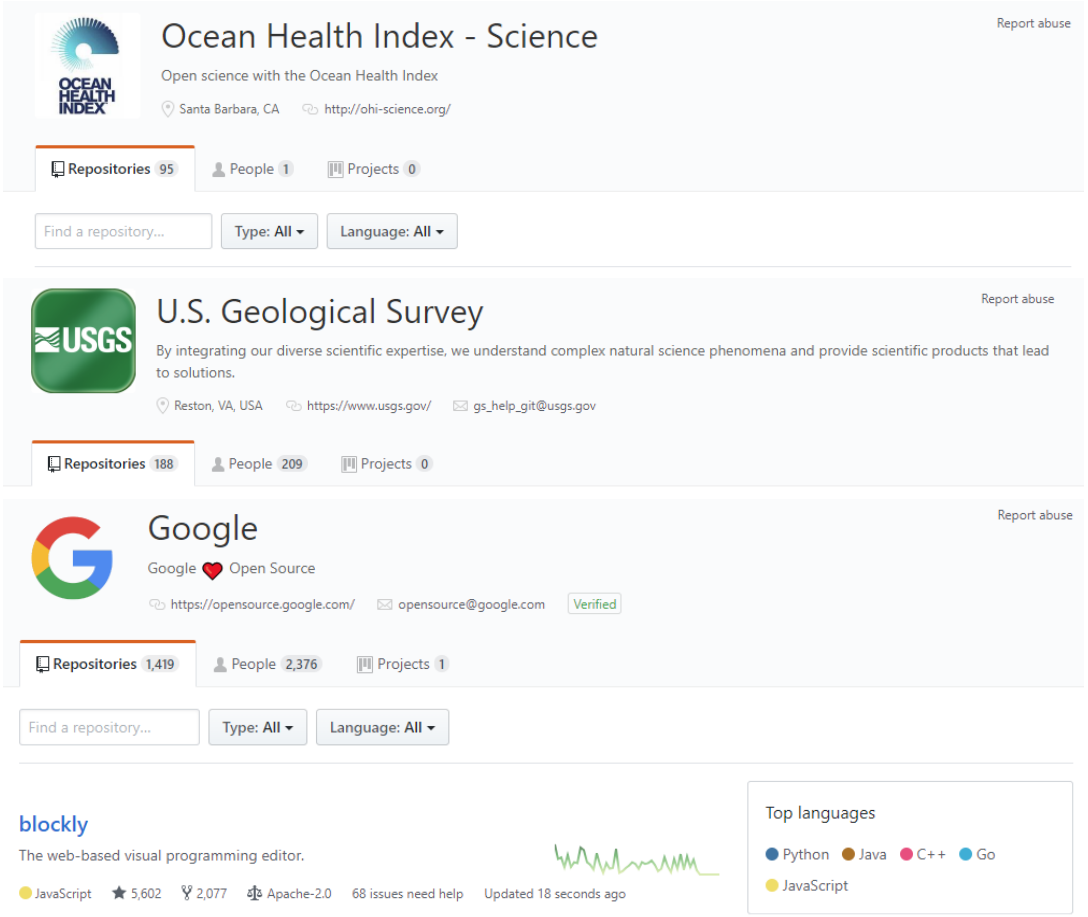
Motivation

- Reproducibility crisis in science *(Baker, 2016; Munafò, 2017).*
  - More than 70% of researchers have tried and failed to reproduce another scientist's experiments, and more than half have failed to reproduce their own experiments.
- Towards reproducible, transparent, open-source analysis *(Kanterakis et al, 2018, Lowndes, 2017).*
- Ease of reproducibility and collaboration (including self).



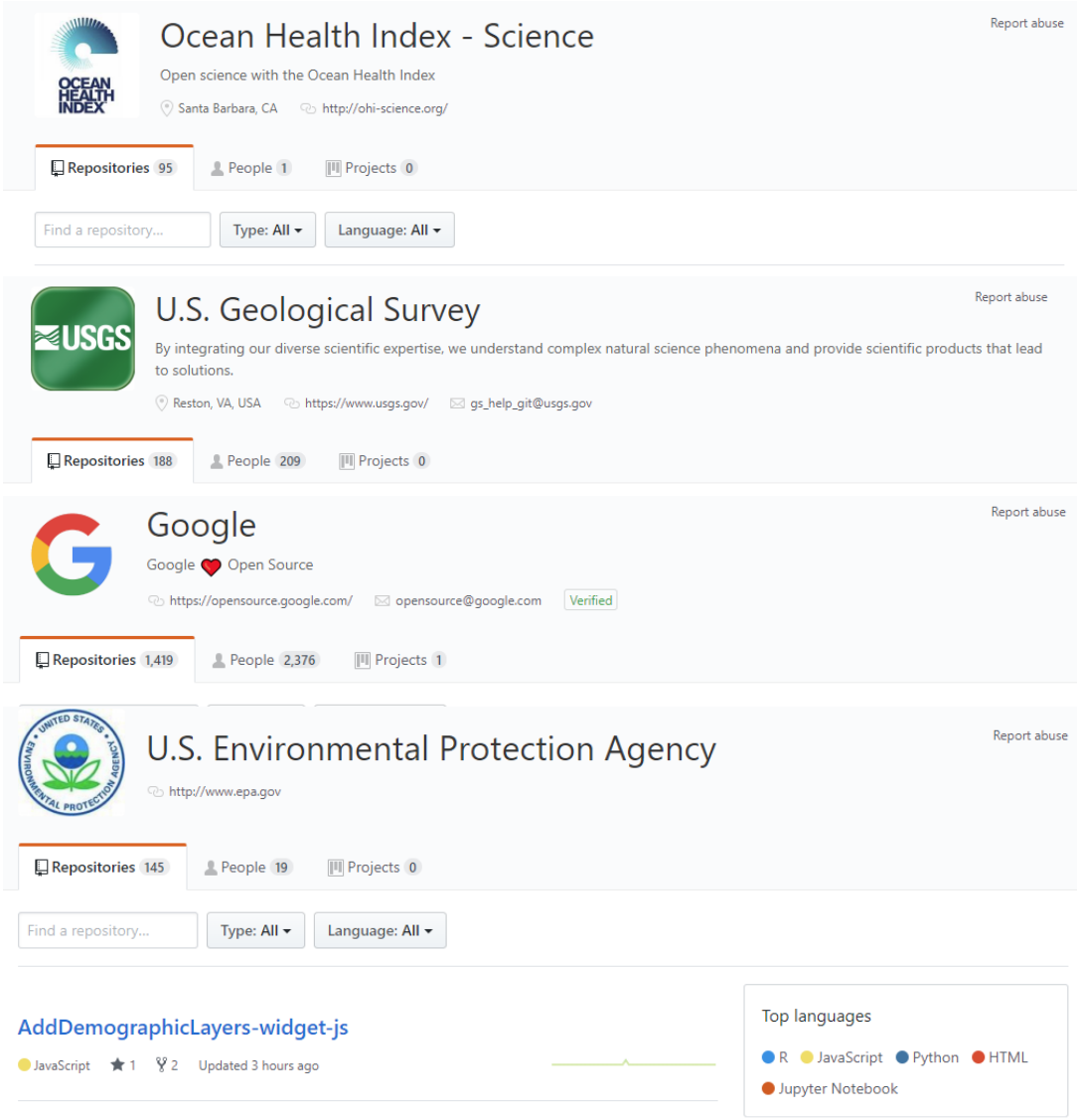
Motivation

- Reproducibility crisis in science *(Baker, 2016; Munafò, 2017).*
  - More than 70% of researchers have tried and failed to reproduce another scientist's experiments, and more than half have failed to reproduce their own experiments.
- Towards reproducible, transparent, open-source analysis *(Kanterakis et al, 2018, Lowndes, 2017).*
- Ease of reproducibility and collaboration (including self).



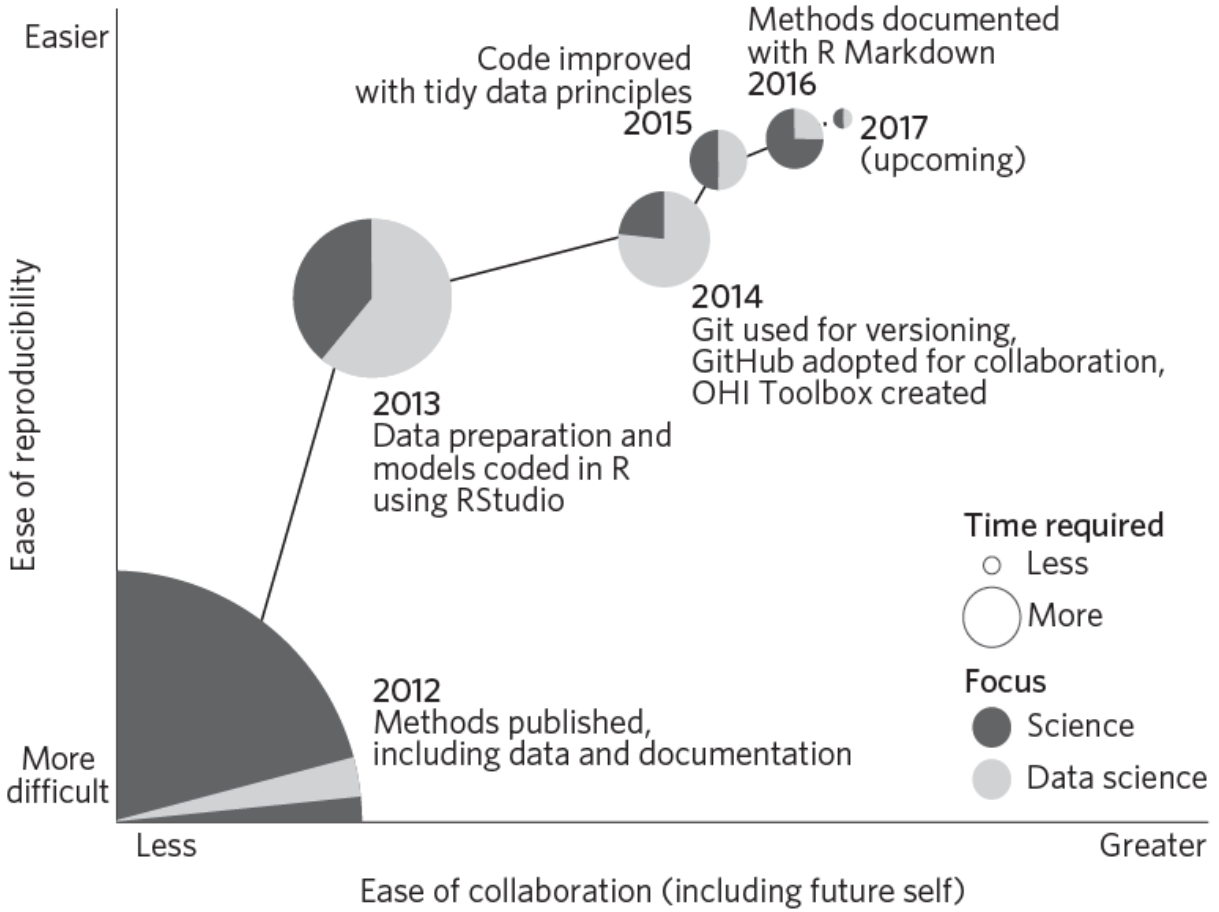
Motivation

- Reproducibility crisis in science *(Baker, 2016; Munafò, 2017).*
  - More than 70% of researchers have tried and failed to reproduce another scientist's experiments, and more than half have failed to reproduce their own experiments.
- Towards reproducible, transparent, open-source analysis *(Kanterakis et al, 2018, Lowndes, 2017).*
- Ease of reproducibility and collaboration (including self).



Motivation

- Reproducibility crisis in science *(Baker, 2016; Munafò, 2017).*
  - More than 70% of researchers have tried and failed to reproduce another scientist's experiments, and more than half have failed to reproduce their own experiments.
- Towards reproducible, transparent, open-source analysis *(Kanterakis et al, 2018, Lowndes, 2017).*
- Ease of reproducibility and collaboration (including self).



(Lowndes, 2017)

Why Python and R?

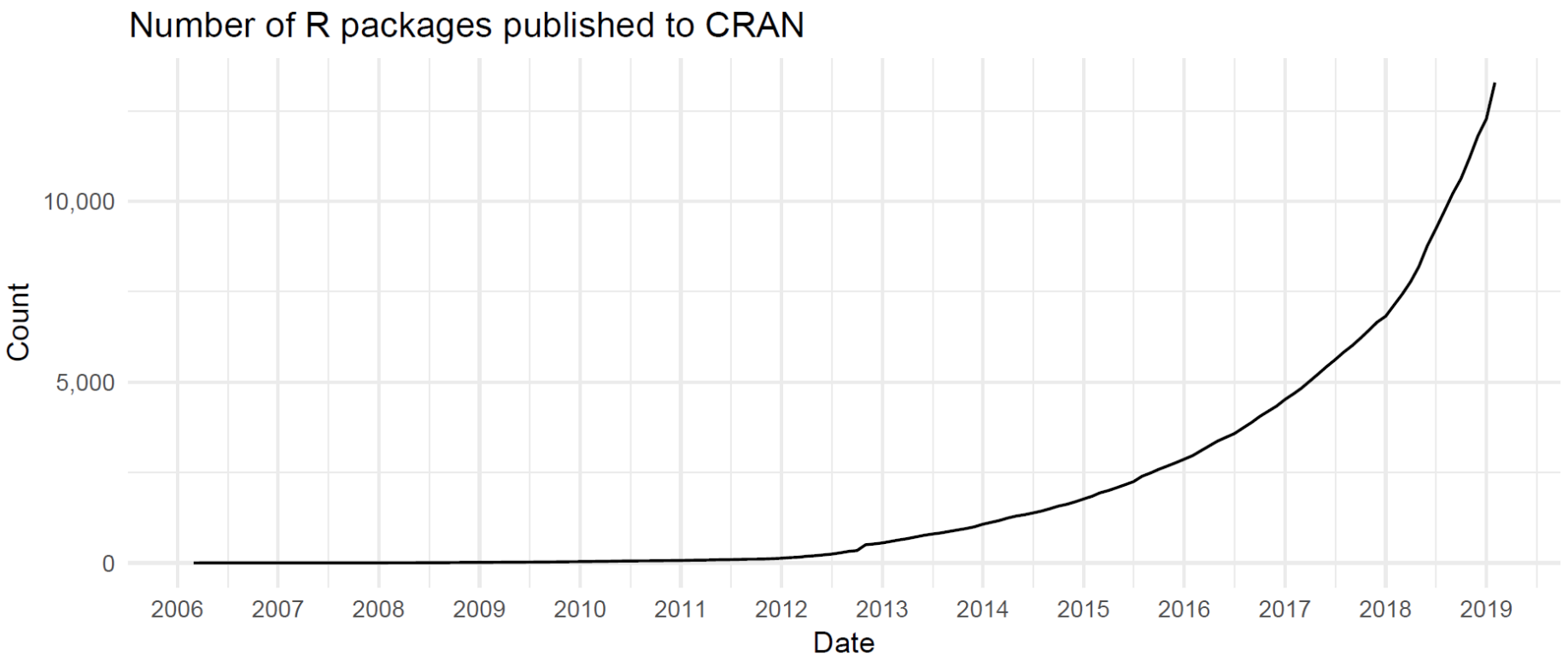
- REPRODUCIBLE = programmatic workflow
- OPEN-SOURCE = free software
- TRANSPARENT = contains end-to-end workflow:
  - Download
  - Cleaning / Transformation
  - Analysis / Modeling
  - Plotting / Map making
  - Report Writing
- Version control (i.e. - no need for “analysis_v12_RP_BV_FINAL.xlsx”)
- Scales to big data (i.e. - won’t fit in: a spreadsheet, your computer)
- User community:
  - Thousands of packages for many kinds of analysis
  - Active StackOverflow Q & A communities

```r
library(rvest)
library(magrittr)
library(ggplot2)
library(plotly)
library(dplyr)
library(zoo)
library(scales)

url <- "https://cran.r-project.org/web/packages/available_packages_by_date.html"

page <- read_html(url)
pkgs <- page %>% html_node("table") %>% html_table() %>%
  mutate(count = rev(1:nrow(.)), Date = as.Date(Date), Month = format(Date, format="%Y-%m")) %>%
  group_by(Month) %>% summarise(published = min(count)) %>% mutate(Date = as.Date(as.yearmon(Month)))

ggplot(pkgs, aes(Date, published)) + geom_line() +
  scale_x_date(date_breaks = "1 year", date_labels = "%Y") +
  scale_y_continuous(label = comma) +
  labs(title = "Exponential Growth of packages published to CRAN", y = "Count") +
  theme_minimal()
```



Questions tagged [r]

R is a free, open-source programming langu
visualization and general computing. Provid
result. Use dput() for data and specify all nc
indented code blocks. For statistics questio

Watch Tag Ignore Tag Le

275,334 questions

Questions tagged [python]

Python is a multi-paradigm, dynamically typed, multipurp
and to understand), and to enforce a clean and uniform s
in use, Python 2.7 and 3.x. For version-specific Python q
Python variant (i.e Jython, Pypy, etc...), please also tag th

Watch Tag Ignore Tag Learn more...

1,109,497 questions info Newe

```r
library(rvest)
library(magrittr)
library(ggplot2)
library(plotly)
library(dplyr)
library(zoo)
library(scales)

url <- "https://cran.r-project.org/web/packages/available_packages_by_date.html"

page <- read_html(url)
pkgs <- page %>% html_node("table") %>% html_table() %>%
  mutate(count = rev(1:nrow(.)), Date = as.Date(Date), Month = format(Date, format="%Y-%m")) %>%
  group_by(Month) %>% summarise(published = min(count)) %>% mutate(Date = as.Date(as.yearmon(Month)))

ggplot(pkgs, aes(Date, published)) + geom_line() +
  scale_x_date(date_breaks = "1 year", date_labels = "%Y") +
  scale_y_continuous(label = comma) +
  labs(title = "Exponential Growth of packages published to CRAN", y = "Count") +
  theme_minimal()
```


Questions tagged [r]

R is a free, open-source programming langu
visualization and general computing. Provid
result. Use dput() for data and specify all nc
indented code blocks. For statistics questio

Watch Tag | Ignore Tag | Le

275,334 questions

Questions tagged [python]

Python is a multi-paradigm, dynamically typed, multipurp
and to understand), and to enforce a clean and uniform s
in use, Python 2.7 and 3.x. For version-specific Python q
Python variant (i.e Jython, Pypy, etc...), please also tag th

Watch Tag | Ignore Tag | Learn more...

1,109,497 questions

info | Newe

Jupyter

- Interactive development environment (IDE)
  - data analysis
  - easily share on Github

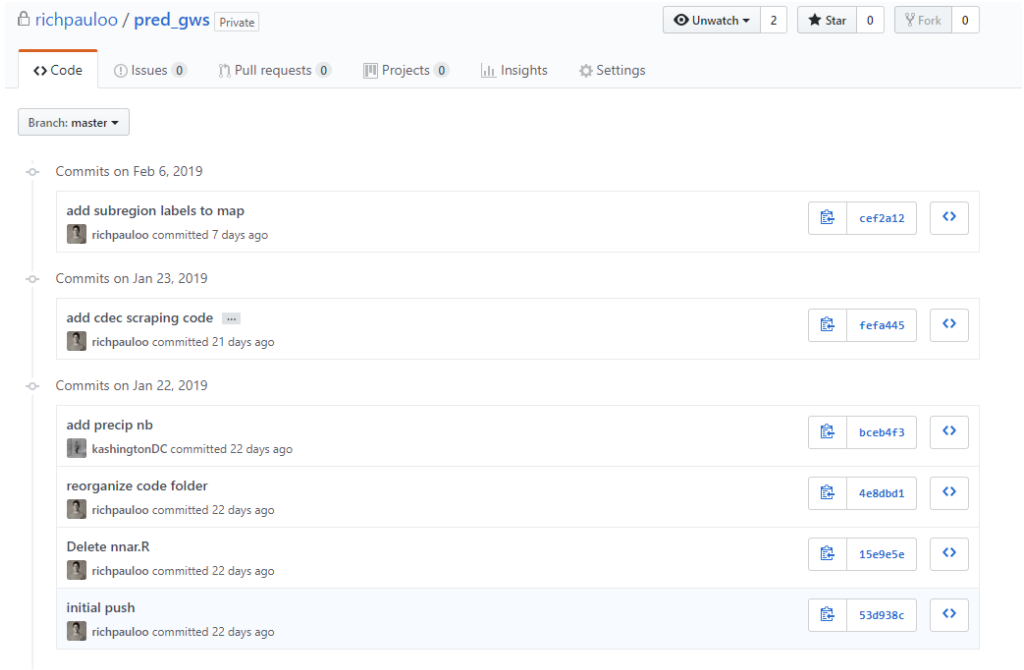
```
In [3]: # read water budget for entire cv
        d <- read_xlsx("C:/Users/rpauloo/Desktop/pred_gws/data/C2VSimFG-BETA_PublicRelease/Excel/C2VSimFG_GWBudget.xlsx",
                       sheet = 22,
                       skip = 4)

In [4]: head(d); tail(d)
```

Time	Percolation	Beginning Storage (+)	Ending Storage (-)	Deep Percolation (+)	Gain from Stream (+)	Recharge (+)	Gain from Lake (+)	Boundary Inflow (+)	Subsidence (+)	Subsurface Irrigation (+)
1973-10-31	817622.6	4575168754	4575555111	543864.7	-152956.8	86155.80	0	229874.4	0	0
1973-11-30	827360.9	4575555111	4578075790	2114095.9	391909.9	35303.47	0	227418.0	0	0
1973-12-31	1102666.2	4578075790	4581297108	2857326.3	244139.3	26615.45	0	226741.3	0	0
1974-01-31	1235043.9	4581297108	4585170011	3150737.9	496860.6	108291.29	0	230756.7	0	0
1974-02-28	701306.1	4585170011	4588180232	3126878.7	-107853.2	86778.34	0	233714.9	0	0
1974-03-31	867709.8	4588180232	4591566945	2827869.0	440654.7	136912.71	0	242601.6	0	0

Github

- Cloud-based version control software
  - creates “save points”
  - workflow for collaborating on code



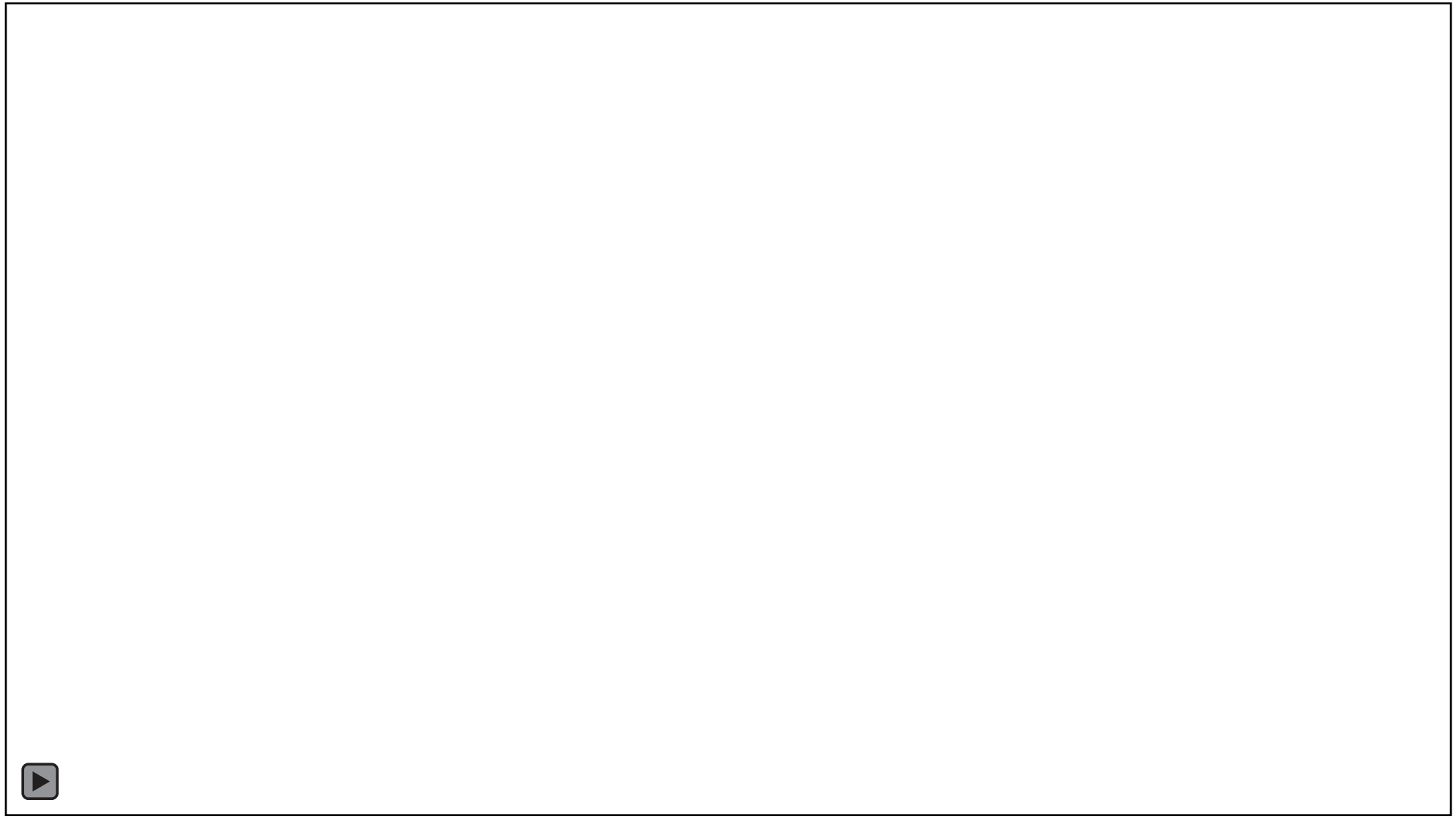
What Else can I do with Python & R?

- Query Databases (SQL, PostgreSQL, etc.)
- Web scraping
- Text Processing
- Image Processing
- Geospatial Data Processing
- Rapid Interactive App Development with Shiny (R) and Flask (Dash)
  - Django Web Development in Python
- R & Python interfaces to popular ML, AI libraries
- And more

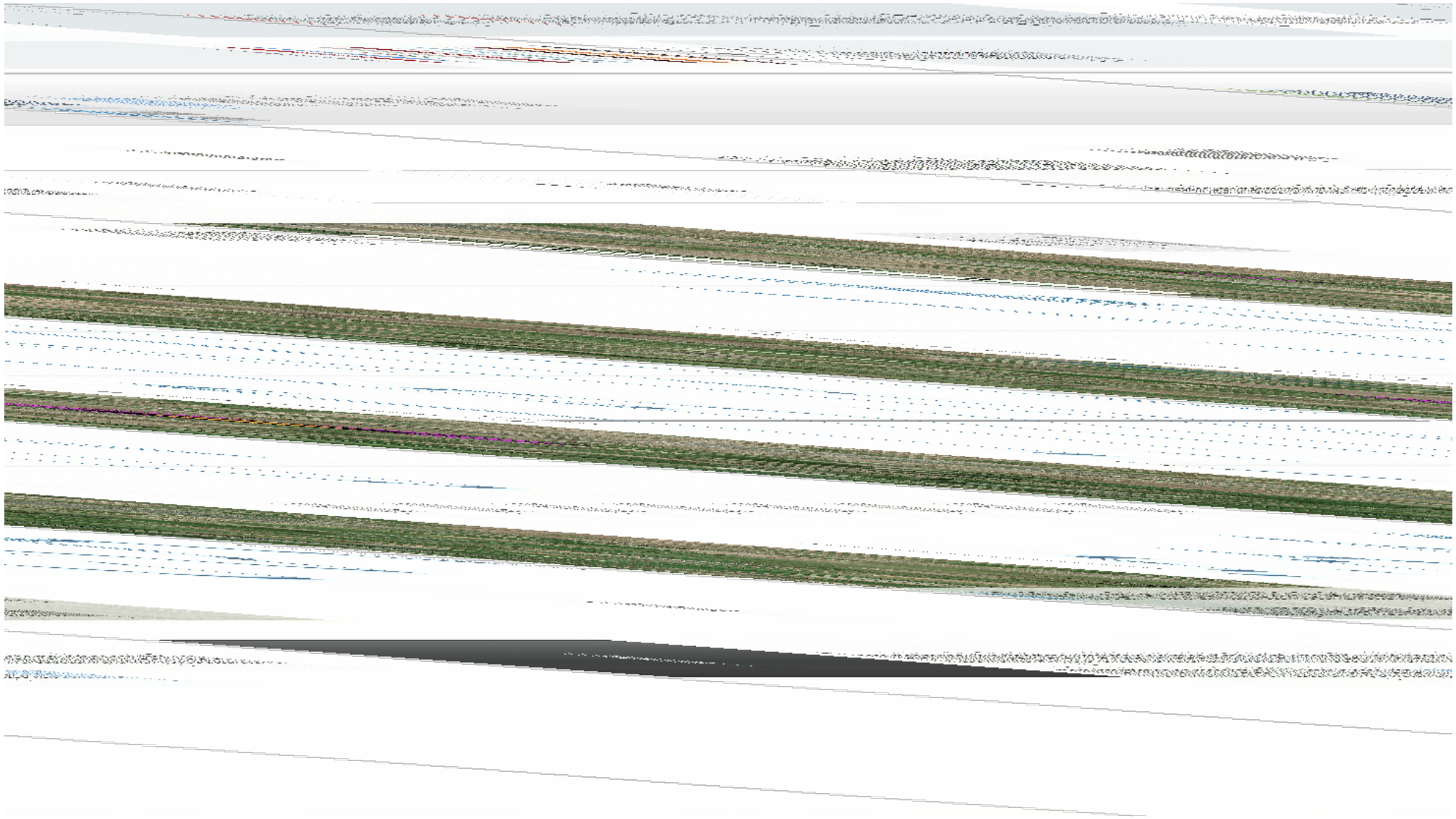
What Else can I do with Python & R?

- Query Databases (SQL, PostgreSQL, etc.)
- Web scraping
- Text Processing
- Image Processing
- Geospatial Data Processing
- Rapid Interactive App Development with Shiny (R) and Flask (Dash)
  - Django Web Development in Python
- R & Python interfaces to popular ML, AI libraries
- And more

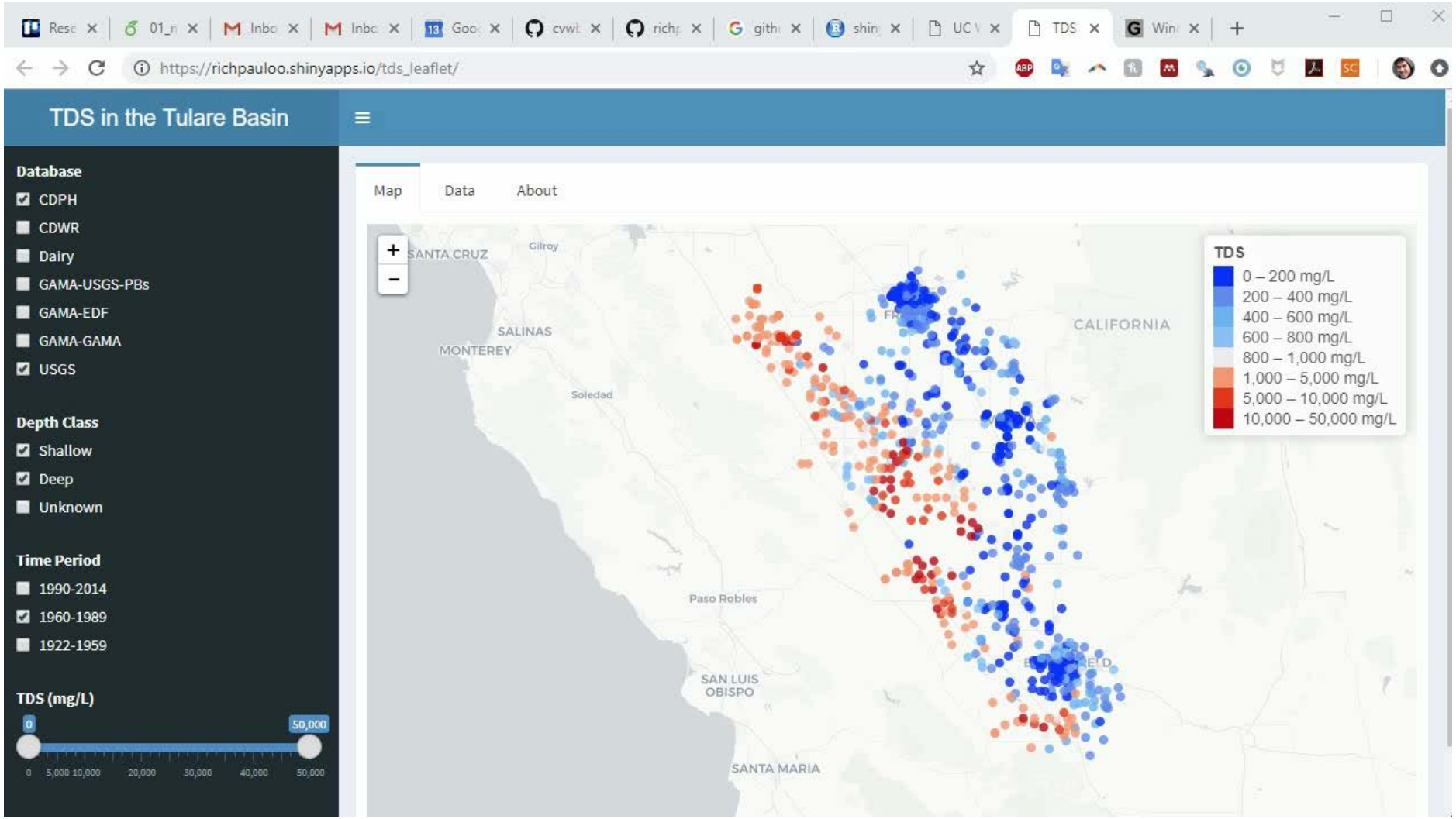
Example: R Shiny: ucwater.org/oswcr/



Example: R Shiny: ucwater.org/gw_obs/



Example: R Shiny: richpauloo.shinyapps.io/tds_leaflet/



Example: Webscrape CDEC Reservoir Storage

- It’s been raining recently.
- How has daily precipitation and reservoir storage at Folsom (FOL) and Shasta (SHA) changed?

Example: Postprocess C2VSim budget

- Geospatial data!
- What's the historical Central Valley, Hydrologic Region, and Subbasin-level change in storage?

Example: Image Process Google Earth Engine

- Tons of Satellite Data out there.
- How do we load just one image?
- What's the evapotranspiration over California according to MODIS?
  - Aakash Ahmed, Stanford University

Thanks for your attention!

Free Online Resources

- Python Data Science Handbook
- R for Data Science
- Datacamp: R and Python online lessons

richpauloo.github.io @RichPauloo rpauloo@ucdavis.edu

References

1. Baker, Monya. "Is there a reproducibility crisis? A Nature survey lifts the lid on how researchers view the 'crisis rocking science and what they think will help." *Nature*, vol. 533, no. 7604, 2016, p. 452+. *Academic OneFile*, Accessed 12 Feb. 2019.
2. 2. Munafò, M. R., Nosek, B. A., Bishop, D. V. M., Button, K. S., Chambers, C. D., Percie Du Sert, N., … Ioannidis, J. P. A. (2017). A manifesto for reproducible science. *Nature Human Behaviour*, *1*(1), 1–9. https://doi.org/10.1038/s41562-016-0021
3. Kanterakis, A., Potamias, G., Swertz, M. A., & Patrinos, G. P. (2018). *Creating Transparent and Reproducible Pipelines: Best Practices for Tools, Data, and Workflow Management Systems*. *Human Genome Informatics*. Elsevier Inc. https://doi.org/10.1016/B978-0-12-809414-3.00002-4
4. Lowndes, J. S. S., Best, B. D., Scarborough, C., Afflerbach, J. C., Frazier, M. R., O'Hara, C. C., … Halpern, B. S. (2017). Our path to better science in less time using open data science tools. *Nature Ecology and Evolution*, *1*(6). https://doi.org/10.1038/s41559-017-0160