Introduction to python and git

Katie Markovich, EPS 522/ ENVS 423L (Fall 2025)

This course will be making use of two open-source tools that are becoming standard practice for hydrologists:

1. python: an object-oriented, interpreted programming language
2. git: a version control system.

These tools empower hydrologists to do collaborative, reproducible, and transparent data-driven/modeling analyses (and they look pretty good on a resume)

Why scientific computing?

python is a vehicle for scientific computing.

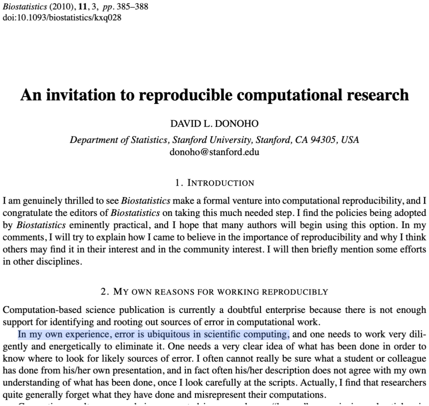
Physical Hydrology involves a lot of quantitative analysis.

Scientific computing empowers the user to show what they did, how they did it, and enables it to be reproduced.

Plus, analysis, postprocessing, and visualization can all occur in the same place!

Ubiquity of Error

Scientific computing does not solve all problems related to quantitative analysis, in fact it may exacerbate human error.



Biostatistics (2010), **11**, 3, *pp*. 385–388
doi:10.1093/biostatistics/kxq028

An invitation to reproducible computational research

DAVID L. DONOHO
Department of Statistics, Stanford University, Stanford, CA 94305, USA
donoho@stanford.edu

1. INTRODUCTION

I am genuinely thrilled to see *Biostatistics* make a formal venture into computational reproducibility, and I congratulate the editors of *Biostatistics* on taking this much needed step. I find the policies being adopted by *Biostatistics* eminently practical, and I hope that many authors will begin using this option. In my comments, I will try to explain how I came to believe in the importance of reproducibility and why I think others may find it in their interest and in the community interest. I will then briefly mention some efforts in other disciplines.

2. MY OWN REASONS FOR WORKING REPRODUCIBLY

Computation-based science publication is currently a doubtful enterprise because there is not enough support for identifying and rooting out sources of error in computational work.

In my own experience, error is ubiquitous in scientific computing, and one needs to work very diligently and energetically to eliminate it. One needs a very clear idea of what has been done in order to know where to look for likely sources of error. I often cannot really be sure what a student or colleague has done from his/her own presentation, and in fact often his/her description does not agree with my own understanding of what has been done, once I look carefully at the scripts. Actually, I find that researchers quite generally forget what they have done and misrepresent their computations.

Check your work and don’t feel bad when you make errors!

Why python?

...instead of R or MATLAB or Excel?

1. python is the most commonly used programming language in the Hydrologic Sciences.
2. There are abundant resources for learning python as it applies to Hydrologic Sciences.
3. There are numerous libraries available for data retrieval, data analysis, hydrologic modeling, etc.

`https://github.com/raoulcollenteur/Python-Hydrology-Tools`

Intro to git

git is a version-control system (VCS) that keeps track of files and all revisions to those files.

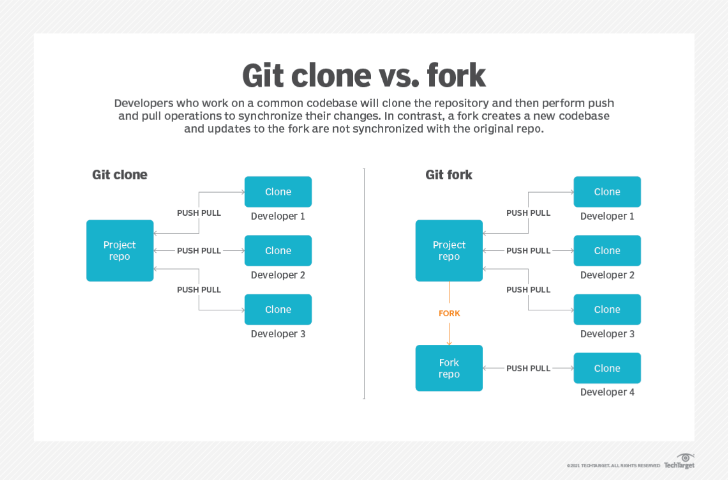
GitHub is a free, cloud-hosting service for git repositories.

Using GitHub allows any of the developers to see:

1. Which changes were made
2. By who
3. When
4. and why they were made

Shared repository

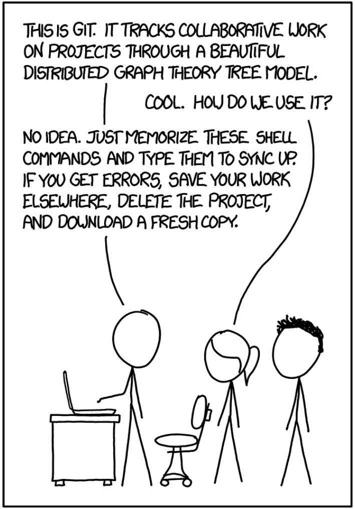
A whole class could be taught on how to team-develop software/workflows using GitHub, but there is basically two approaches: shared repository and fork and pull.



Git clone vs. fork

Developers who work on a common codebase will clone the repository and then perform push and pull operations to synchronize their changes. In contrast, a fork creates a new codebase and updates to the fork are not synchronized with the original repo.

Let's clone!


THIS IS GIT. IT TRACKS COLLABORATIVE WORK ON PROJECTS THROUGH A BEAUTIFUL DISTRIBUTED GRAPH THEORY TREE MODEL.
COOL. HOW DO WE USE IT?
NO IDEA. JUST MEMORIZE THESE SHELL COMMANDS AND TYPE THEM TO SYNC UP. IF YOU GET ERRORS, SAVE YOUR WORK ELSEWHERE, DELETE THE PROJECT, AND DOWNLOAD A FRESH COPY.

What is conda?

Wait, you said we'd only be learning 2 tools!?

conda is a free, open-source package and environment manager for python. We will be using it to install python as well as our python packages to ensure we are all "on the same page"

Assuming you have conda installed...

What is jupyter notebook?

A web-application that we will be using to run and visualize or python code for the homeworks.

jupyter notebook shortcuts

`shift+enter`: execute cell
`ctrl+/`: comment out line (note: CMD for Mac)
`ESC`: enter into "edit mode"
if in edit mode:
`A`: add cell above
`B`: add cell below
`D,D`: delete cell

at last, a python tutorial

We will be covering:

- python basics
- os, sys, path
- functions and scripts
- arrays and math with NumPy
- dataframes with pandas
- plotting with matplotlib

House Keeping

I will be pushing your first homework problem set to our class repository today, please be sure to `git pull` (if using shared repository approach) or `git fetch` (if using fork and pull approach).

It will be due (uploaded to Canvas) by midnight on **Tuesday, August 26th**.

Our next lecture will dive into The Climate System. If you are feeling eager, you can start reading Chapter 2 in Dingman.