DSA101.1
Introduction to Data Analytics and Python

Prof. Alexandria Deliz Liang

Prof. Meredith C. Schuman, meredithchristine.schuman@uzh.ch

Dr. Domitille Coq--Etchegaray

Who am I?

Meredith C. Schuman (Merry)

Born in Minnesota, USA

2012 PhD in Chemical Ecology, FSU, Jena, DE

2012-2019 Group Leader, Max Planck Institute for Chemical Ecology, Jena, DE

Since 2019 Assistant Professor for Spatial Genetics, UZH, CH

University of Zurich UZH
Departments of Geography and Chemistry

We want to know *why* living things differ and *where* those differences matter for ecology and for people.

remote sensing

spatial chemical sampling

genetic associations

ecological phenotypes
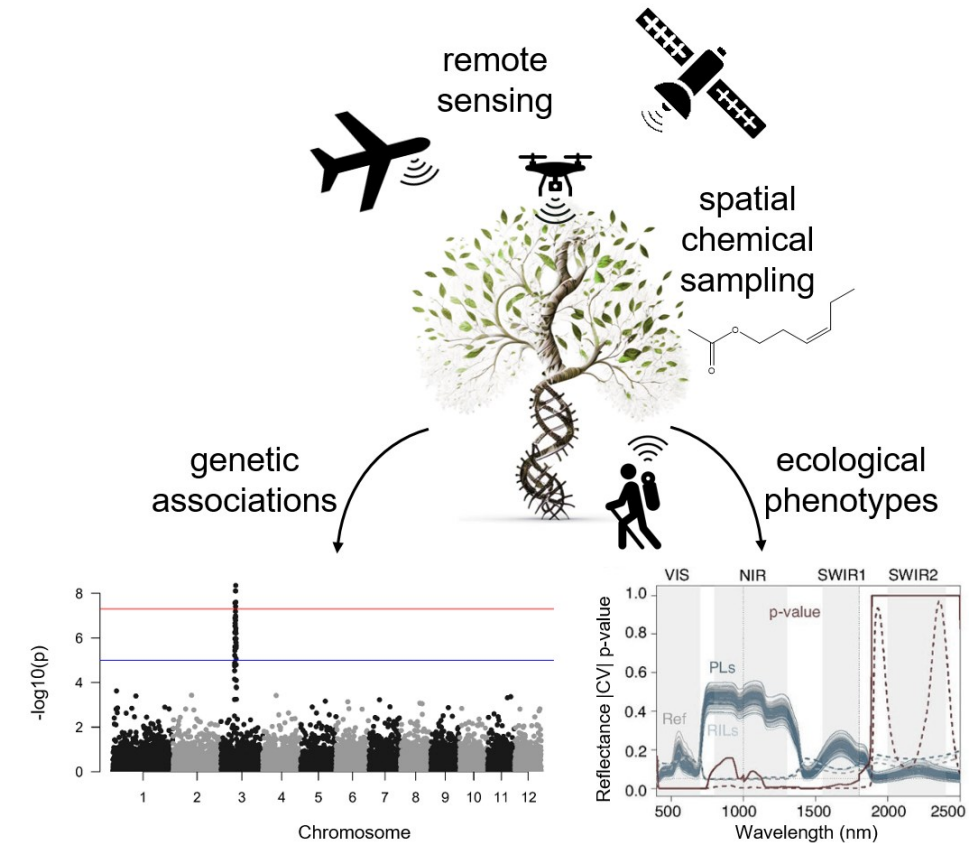
"tree DNA white background": gencraft, spectra: C. Li et al. (2023) CC-BY 4.0

Who are you?

Motivation

"METRIC, ENGLISH, WHATEVER..."
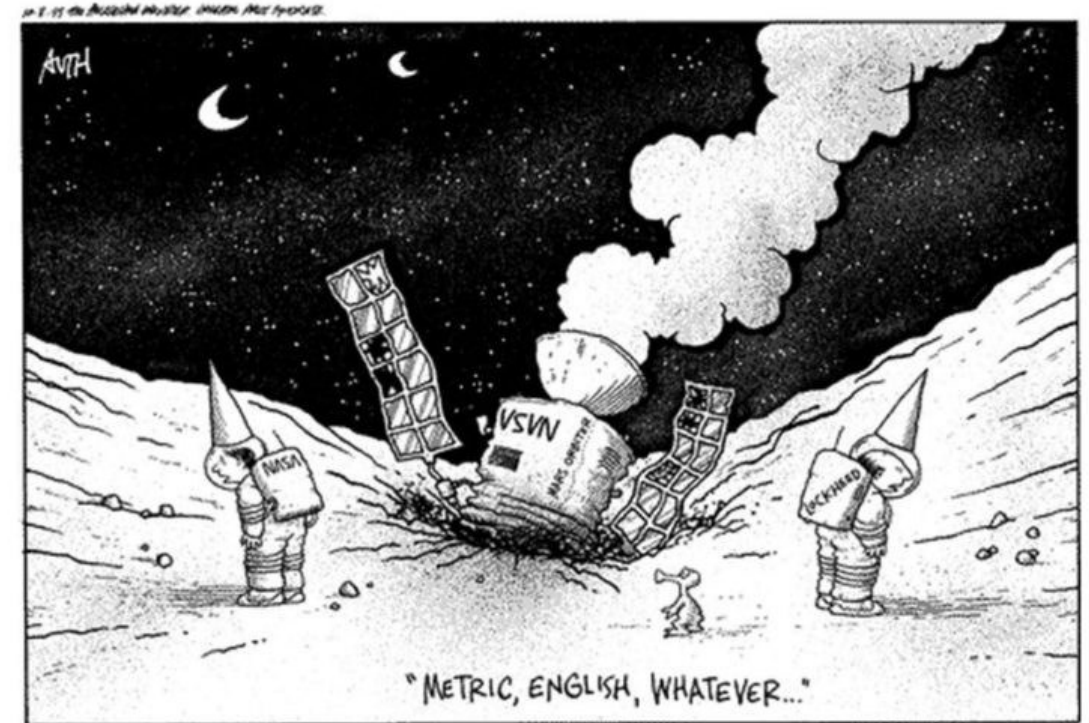
A Boeing Delta II 7425 expendable launch vehicle lifts off with NASA's Mars Climate Orbiter on Dec. 11, 1998
https://www.simscale.com/blog/nasa-mars-climate-orbiter-metric/

Motivation

The retracted paper analyzed excitatory synapse number and neurotransmitter release in mouse hippocampal neurons missing a presynaptic cell-adhesion molecule called neurexin-2. Südhof, professor of molecular and cellular physiology and neurosurgery at Stanford University, received the Nobel Prize for some of his previous work on neurexin proteins.

The authors retracted the study "because re-analysis of the original raw data for Figs. 2, 4 and 6 revealed that, although our analyses of the original data are supportive of the conclusions of the paper, unresolvable differences exist between these raw data and the published data source file that cannot be corrected by a simple erratum," according to the 5 March retraction notice.
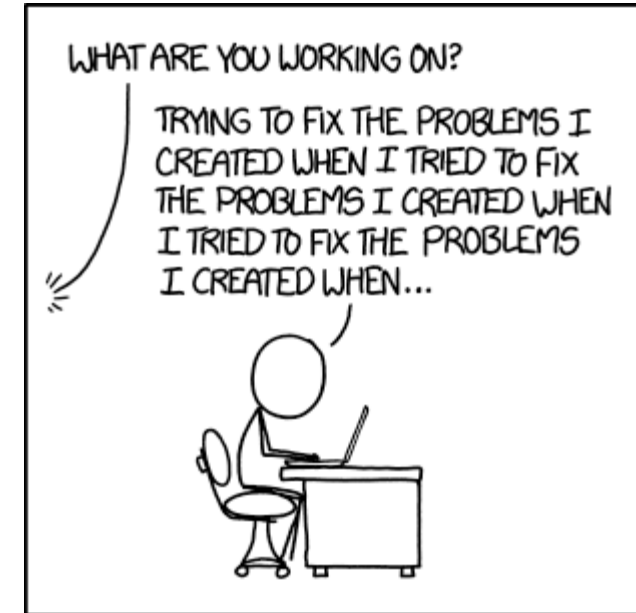
We want to solve a problem using a computer.

There will be many similar, but not identical problems.

Did we solve the original problem correctly?

What if…
- The data change?
- The assumptions change?
- We have 10x as much data? (Or 0.1x as much data?)
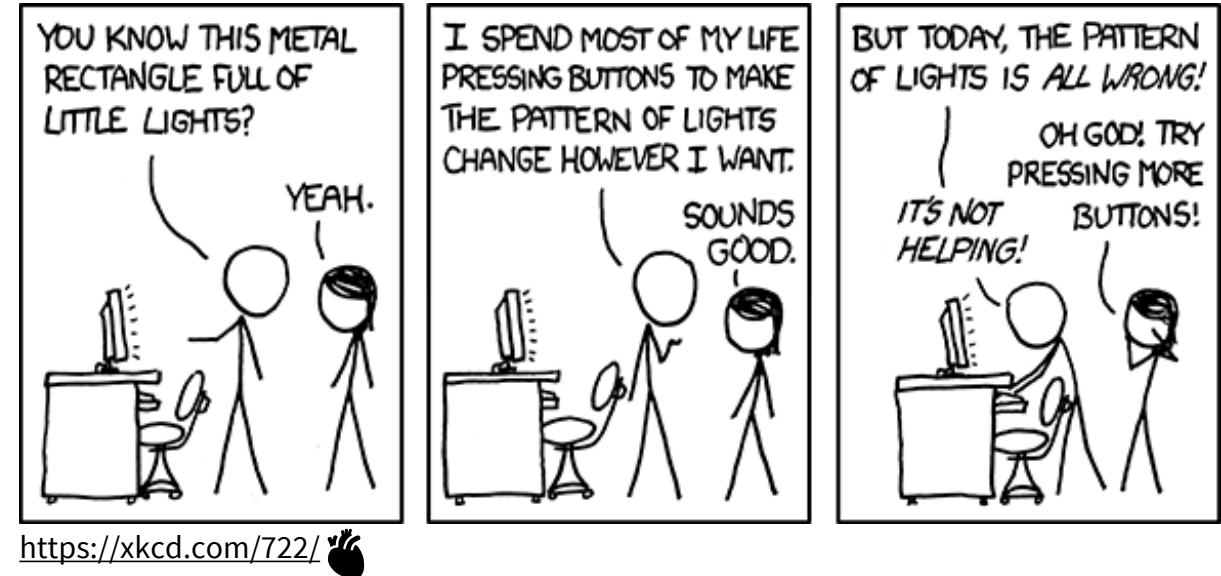- We pass this on to a colleague, or to ourselves six months or six years from now?

https://www.xkcd.com/1739/

We want to solve a problem using a computer.

If we made a mistake, what happens?

- Program crashes?
- Everything is *increeeedibly* slow?
- It still works for *this specific case*?
- The output is a *little bit* wrong?
- The output is wrong *sometimes*?

https://xkcd.com/722/

We want to solve a problem using a computer.

If we made a mistake, what happens?

Clear data science ranking:

- Program crashes?
- Everything is *increeeedibly* slow?
- It still works for *this specific case*?
- The output is a *little bit* wrong?
- The output is wrong *sometimes*?

OK

DANGER

https://xkcd.com/722/
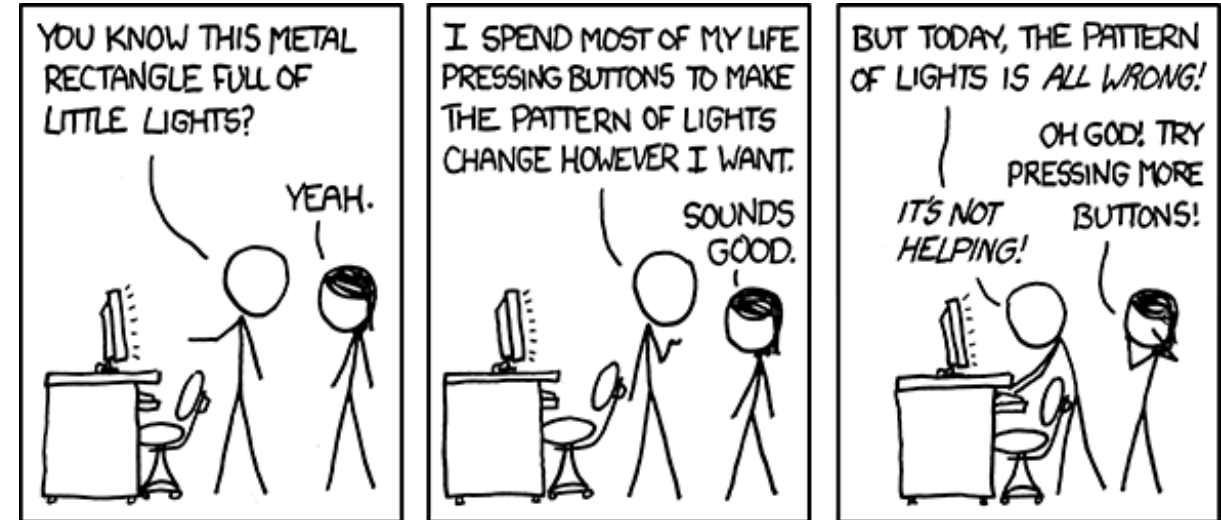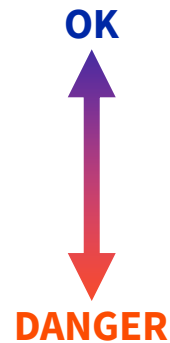
LIKELIHOOD YOU WILL GET CODE WORKING BASED ON HOW YOU'RE SUPPOSED TO INSTALL IT:

VERY LIKELY

APP STORE
OR PACKAGE
MANAGER

GITHUB LINK

SOURCEFORGE LINK

GEOCITIES/TRIPOD LINK

COPY-AND-PASTE
EXAMPLE CODE FROM
PAPER'S APPENDIX

ANYTHING THAT "REQUIRES
ONLY MINIMAL CONFIGURATION
AND TWEAKING"

UNLIKELY

https://xkcd.com/1742/

What is Git?

Git is a program for version control.

What is version control and why is it important?

- You change a file → change can be tracked
- A collaborator changes a file → changes can be merged
- You can go back and forth in time of your project
- You can branch out, test something and either integrate or revert changes
- **You can hardly break anything**

https://docs.gitlab.com/ee/topics/git/

The program

Free individual accounts, North American interface used worldwide
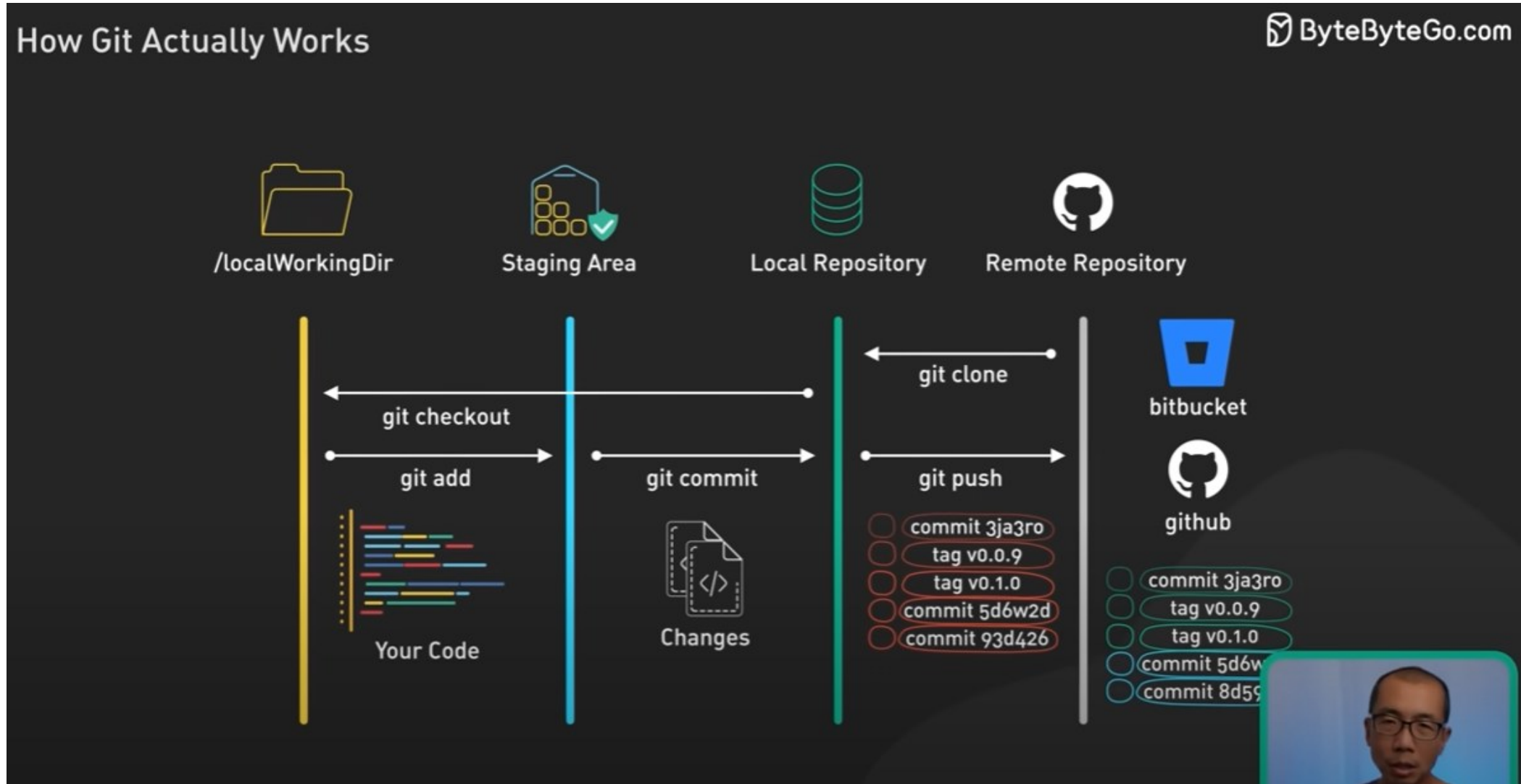
Institutional accounts, European interface

How git actually works

https://youtu.be/e9lnsKot_SQ?si=g0d_pQD2eUPpP8Ae

How git actually works
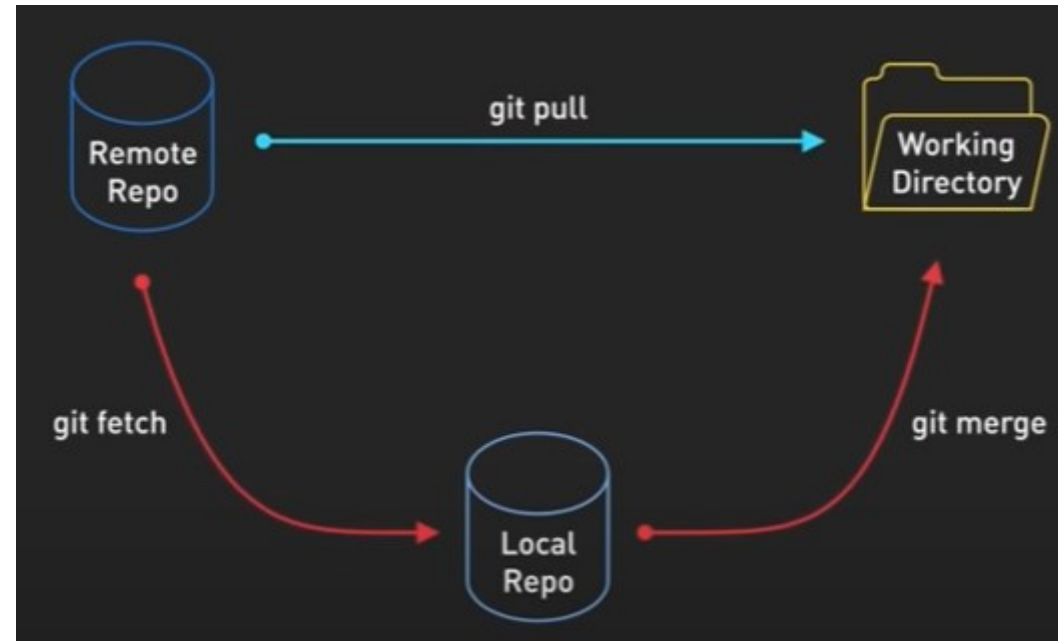
https://youtu.be/e9lnsKot_SQ?si=g0d_pQD2eUPpP8Ae

How git actually works
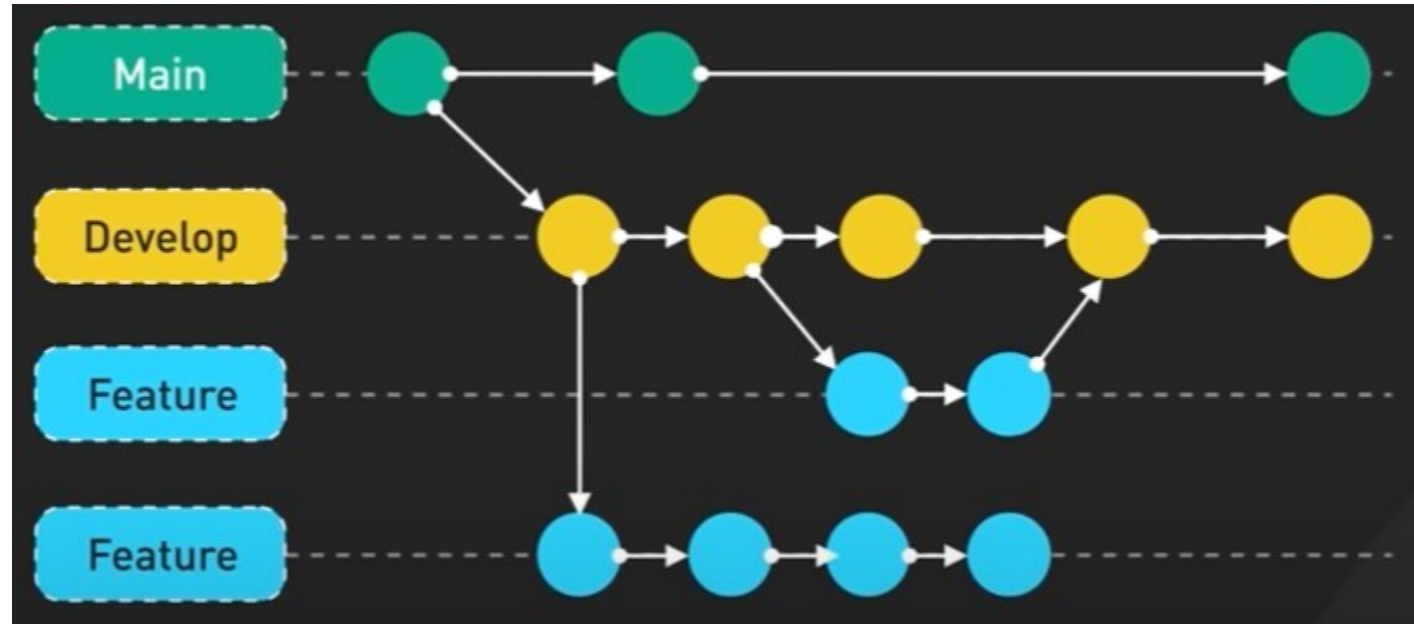
https://youtu.be/e9lnsKot_SQ?si=g0d_pQD2eUPpP8Ae

Git resources

The Git Cheat Sheet

Basic guides

- Git - the simple guide, https://rogerdudler.github.io/git-guide/
- Git tutorial and quiz, https://www.w3schools.com/git/
- Github hello world guide, https://docs.github.com/en/get-started/quickstart/hello-world

Advanced guides

- Learn Enough Git to Be Dangerous, https://www.learnenough.com/git-tutorial/getting_started

Learn Enough Git to Be Dangerous

Michael Hartl

An Introduction to Version Control with Git

We will be working from a GitHub repo

Let's start!

(1) <u>Make a GitHub account</u>
    if you don't have one

Let's start!

(1) Make a GitHub account if you don't have one

(2) Install Git: there are many options!

Let's start!

(1) <u>Make a GitHub account</u> if you don't have one

(2) <u>Install Git</u>: there are many options!

— **Easiest**: an IDE like Visual Studio Code: interface to interact with code and Git

Visual Studio Code or another IDE

Let's start!

(1) Make a GitHub account if you don't have one

(2) Install Git: there are many options!

- **Easiest**: an IDE like Visual Studio Code: interface to interact with code and Git

- **Second easiest**: GitHub Desktop, a GUI that prompts you to use Git correctly

Visual Studio Code or another IDE
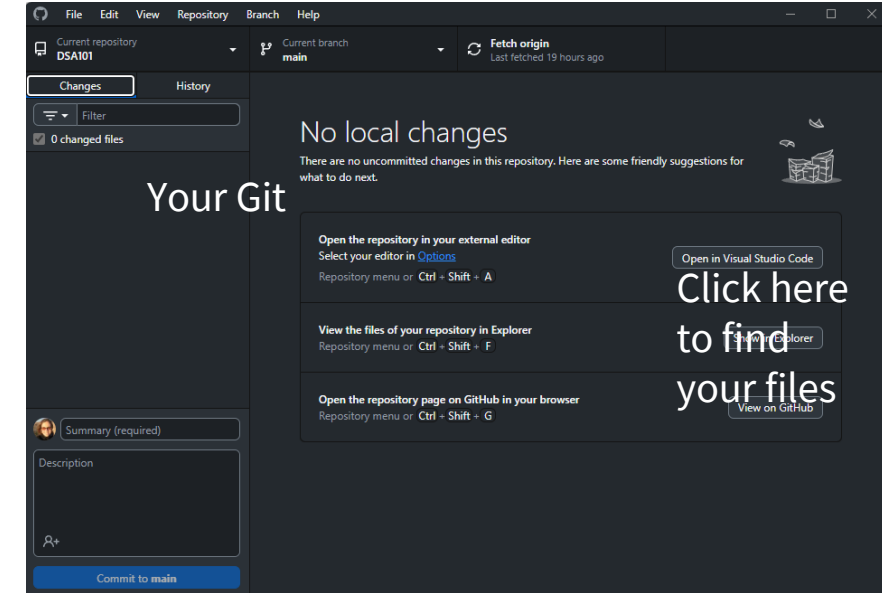
Your Git

Your files

GitHub Desktop

Your Git

Click here to find your files
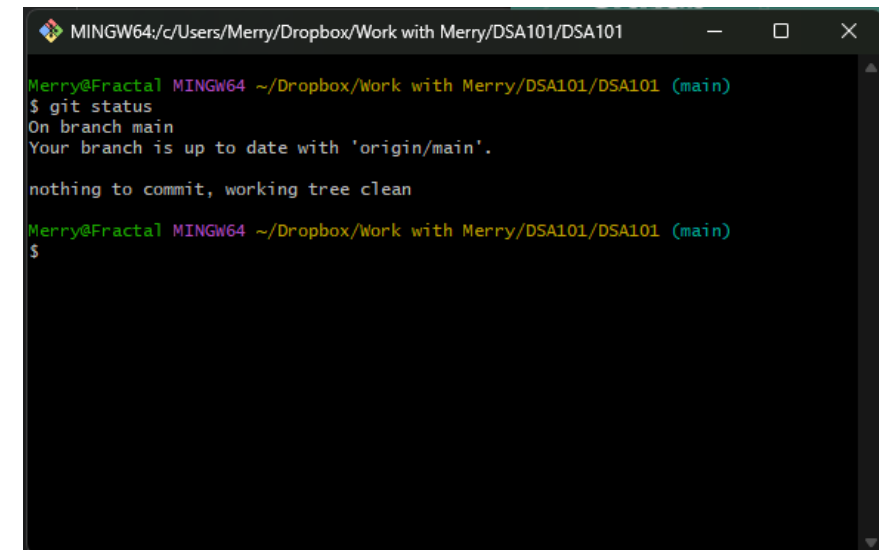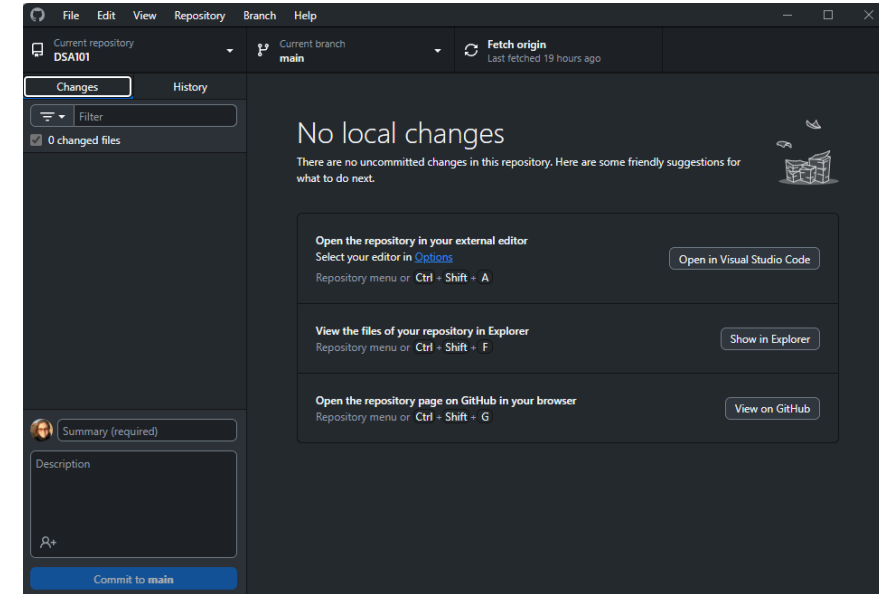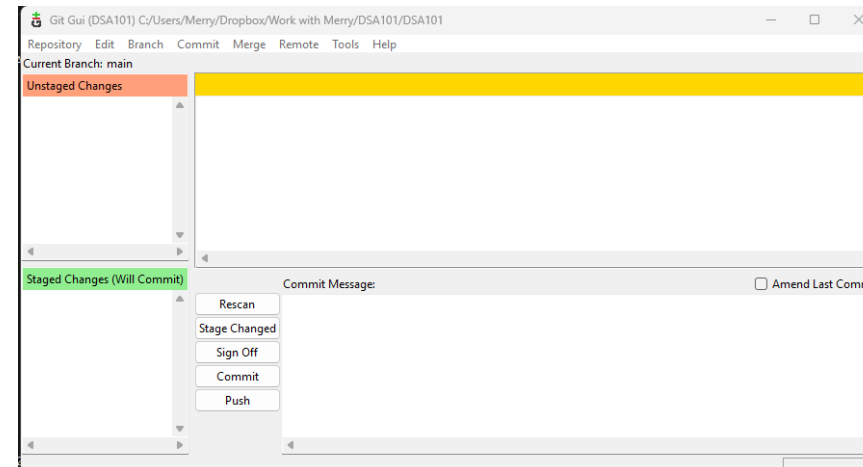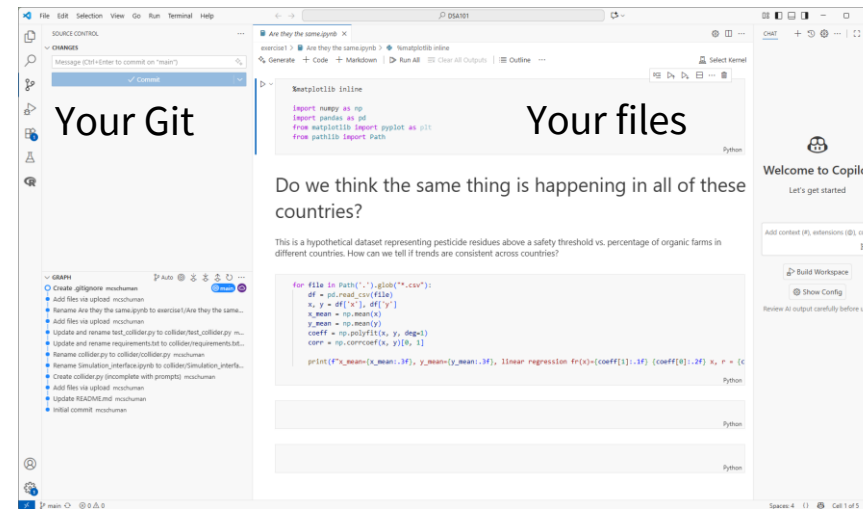
Let's start!

(1) <u>Make a GitHub account</u> if you don't have one

(2) <u>Install Git</u>: there are many options!

— **Easiest**: an IDE like Visual Studio Code: interface to interact with code and Git

— **Second easiest**: GitHub Desktop, a GUI that prompts you to use Git correctly

— **Develop your computer skills (still fairly easy)**: install with Git installer, then interact with Git using Git GUI or Git Bash (or Terminal)

Visual Studio Code or another IDE
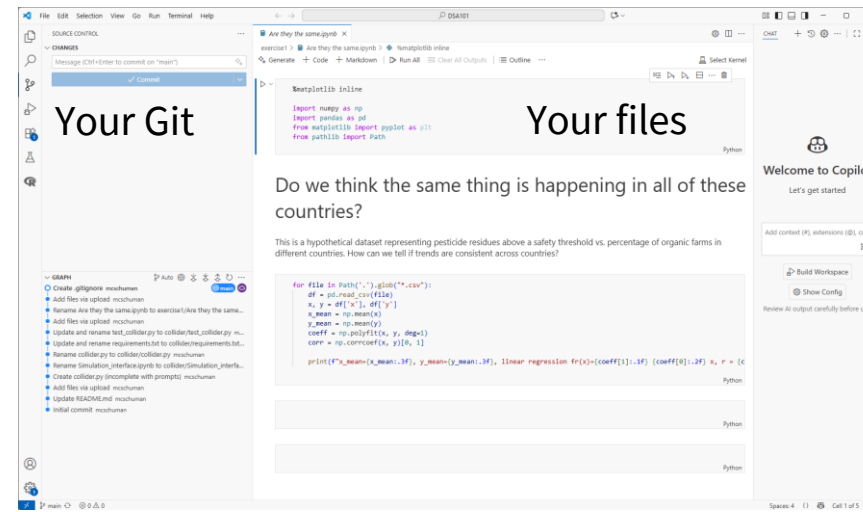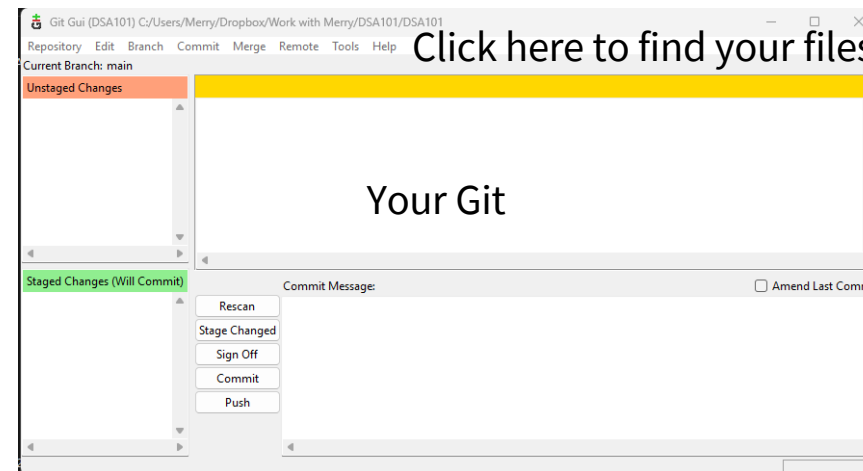
Your Git

Your files

Git GUI

Click here to find your files

Your Git

GitHub Desktop

Your Git

Click here to find your files

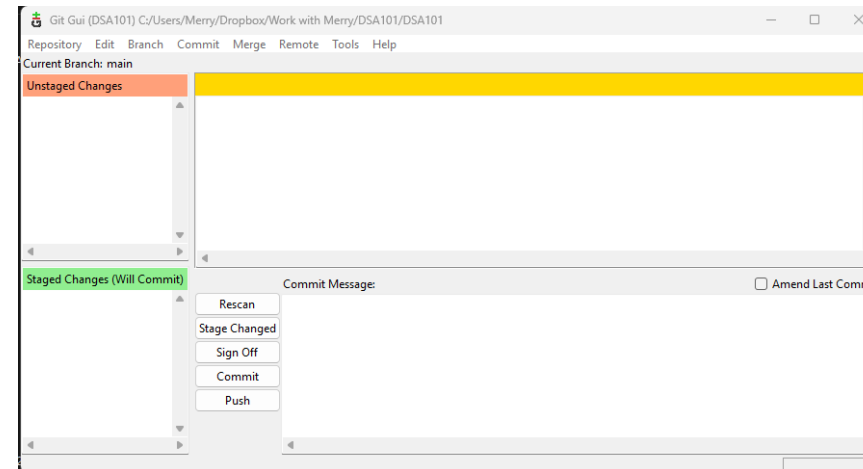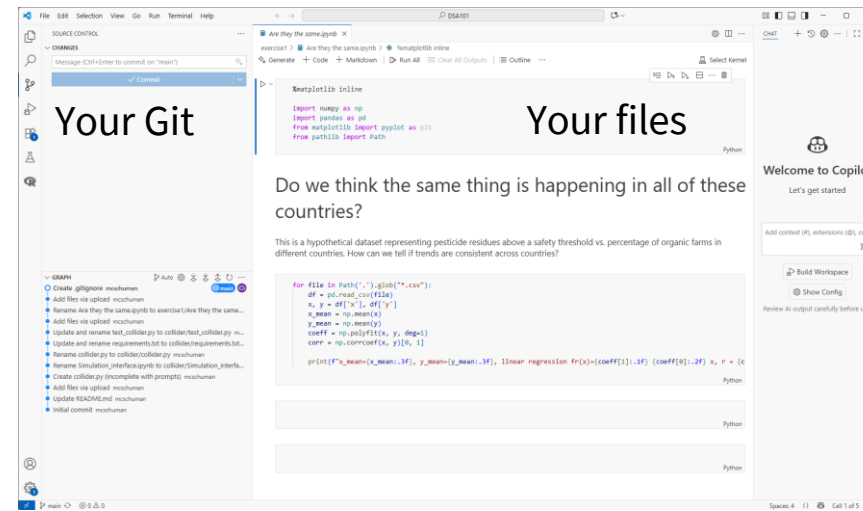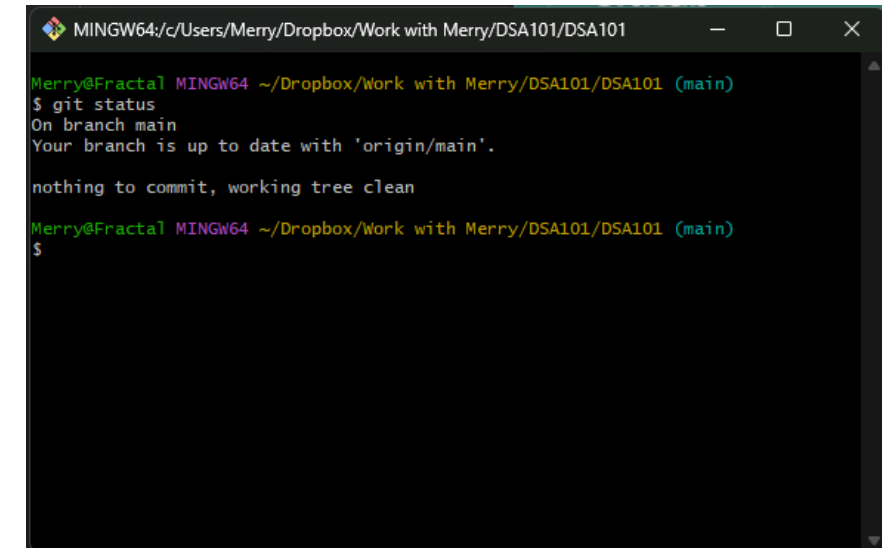## Git Bash

Your Git

Your filepath

Let's start!

(1) Make a GitHub account if you don't have one

(2) Install Git: there are many options!

(3) Fork the DSA101 repository

Edit Pins ▾   ⊙ Watch 2.4k ▾   ⑂ Fork 62.3k ▾   ☆ Star 15k ▾

🔍 Go to file   t   Add file ▾   <> Code ▾   **About**

·collaborator... ••• ✓ 7e310f0 · 1 hour ago ⏱ 44,459 Commits   The open-source repo for docs.github.com

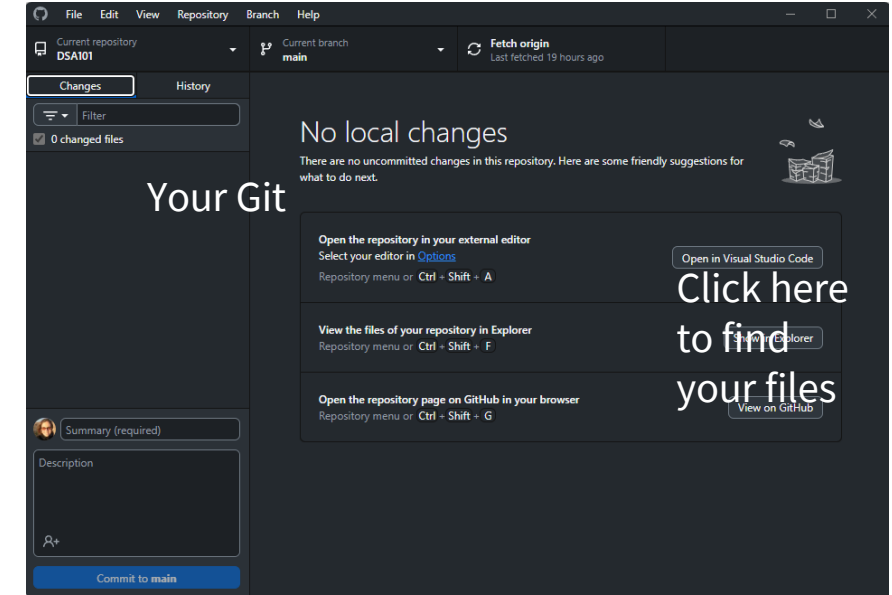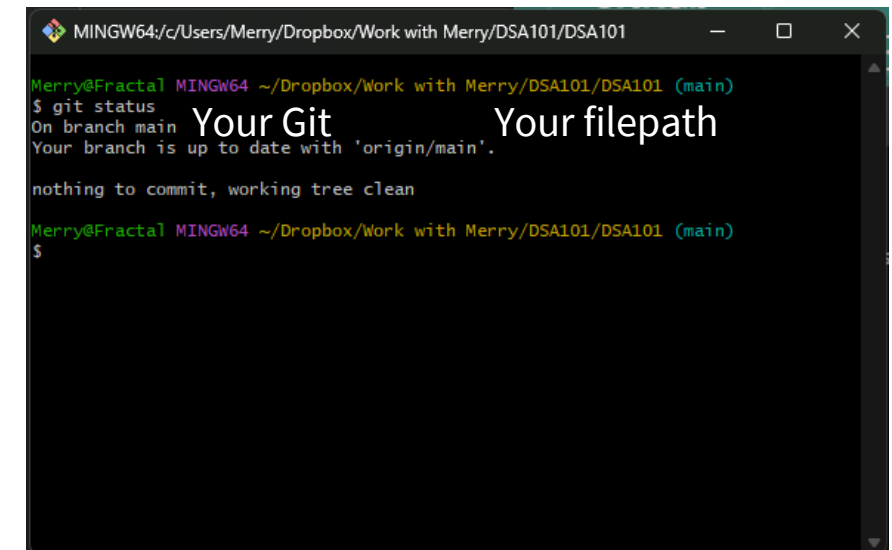**3** Under "Owner," select the dropdown menu and click an owner for the forked repository.

4 By default, forks are named the same as their upstream repositories. Optionally, to further distinguish your fork, in the "Repository name" field, type a name.

5 Optionally, in the "Description" field, type a description of your fork.

6 Optionally, select **Copy the DEFAULT branch only**.

For many forking scenarios, such as contributing to open-source projects, you only need to copy the default branch. If you do not select this option, all branches will be copied into the new fork.

7 Click **Create fork**.

https://docs.github.com/en/pull-requests/collaborating-with-pull-requests/working-with-forks/fork-a-repo
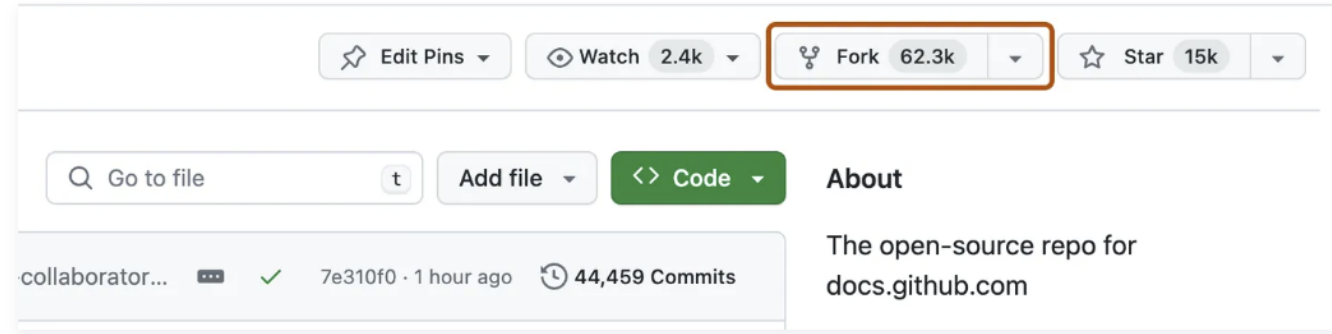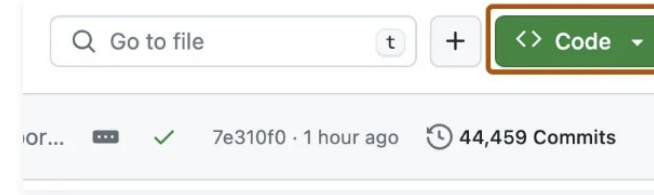
Let's start!

(1) <u>Make a GitHub account</u> if you don't have one

(2) <u>Install Git</u>: there are many options!

(3) <u>Fork</u> the DSA101 repository

(4) <u>Clone your fork</u>

I expect you to understand how Git works at this level for the exam.

③ Copy the URL for the repository.

- To clone the repository using HTTPS, under "HTTPS", click ⧉.
- To clone the repository using an SSH key, including a certificate issued by your organization's SSH certificate authority, click **SSH**, then click ⧉.
- To clone a repository using GitHub CLI, click **GitHub CLI**, then click ⧉.
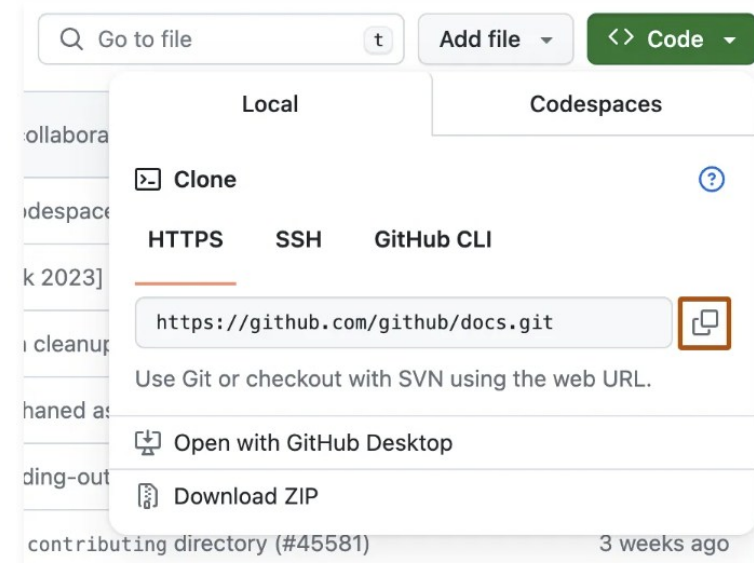
Next: try exercise1

Are they the same?

We have a fictitious dataset from 12 different countries documenting the relationship between proportion of organic farms, and proportion of test sites in which pesticide residues are above a regulatory threshold.

You will characterize this relationship for each country and answer the question, is the relationship essentially the same across all countries in our dataset?

I expect you to understand how to look at data to answer specific questions for the exam.

https://4.bp.blogspot.com/-1gcbXO2ePh4/UY8Vzp5khqI/AAAAAAAAJbk/i6HUCVq43GE/s1600/pesticide+spraying+man.jpg