Week 3 Discussion Sections

COGS 108 Fall 2024

Due dates

- A1: this friday (April 19)
- D2: next monday (April 22)
- Q3: next monday (April 22)

Data wrangling

Data wrangling deals with several functionalities:

1. Data exploration: In this process, the data is studied, analyzed and understood by visualizing representations of data.

2. Dealing with missing values: Most of the datasets having a vast amount of data contain missing values of NaN, they are needed to be taken care of by replacing them with mean, mode, the most frequent value of the column or simply by dropping the row having a NaN value.

Data wrangling

Data wrangling deals with several functionalities:

1. Data exploration

2. Dealing with missing values

3. Reshaping data: In this process, data is manipulated according to the requirements, where new data can be added or pre-existing data can be modified.

4. Filtering data: Some times datasets are comprised of unwanted rows or columns which are required to be removed or filtered

pandas and numpy

import pandas as pd

import numpy as np

pandas	numpy
When we have to work on Tabular data, we prefer the pandas module.	When we have to work on Numerical data, we prefer the NumPy module.
Pandas have a 2D table object called DataFrame.	Numpy is capable of providing multi-dimensional arrays.
The powerful tools of pandas are DataFrame and Series.	the powerful tool of NumPy is Arrays.
Pandas consume more memory.	Numpy is memory efficient.
Indexing of the Pandas series is very slow as compared to Numpy arrays.	Indexing of Numpy arrays is very fast.

https://www.geeksforgeeks.org/difference-between-pandas-vs-numpy/

pandas operation

read csv files into a pandas df: pd.read_csv(“link”)

Programming

This course assumes basic programming knowledge

- But not much!

Programming

Resources:

- Codeacademy
- Start Here: https://github.com/COGS108/Tutorials/blob/master/01-Python.ipynb
- Python in detail: https://jakevdp.github.io/PythonDataScienceHandbook/
- Pandas: https://www.dataschool.io/python-pandas-tips-and-tricks/
- Git: https://guides.github.com/activities/hello-world/

Programming

Cheatsheets

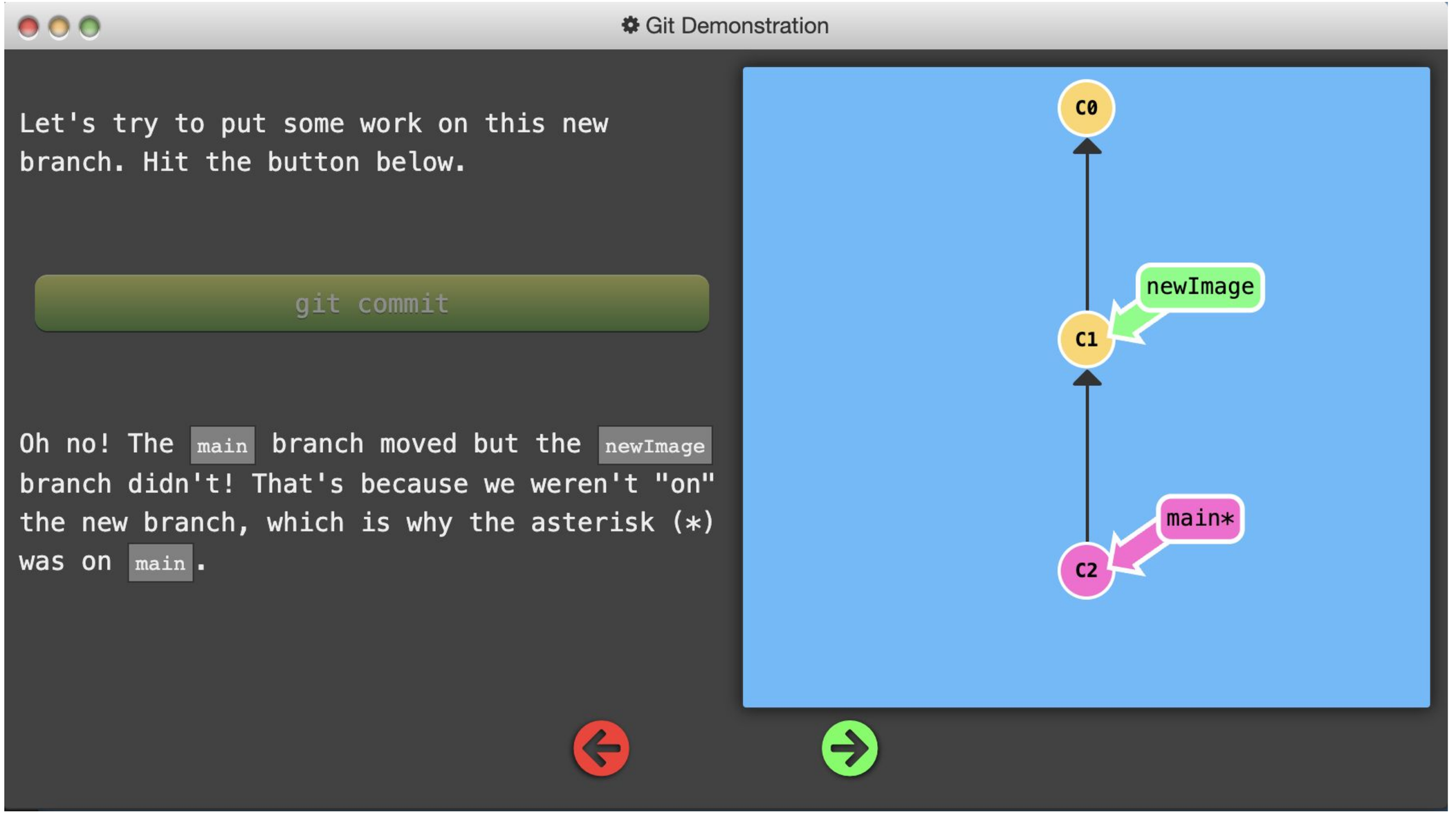
- Google: ‘python cheatsheet’, ‘pandas cheatsheet’, ‘git cheatsheet’ (find one that’s good for you)

Git

Version control system!

- Go to https://git-scm.com/downloads
- Choose your Operating System (Windows/OS X/Linux)
- Follow the steps specific to your OS
- Verify installation: In terminal type “git –-version”

learngitbranching.js.org



https://about.gitlab.com/images/press/git-cheat-sheet.pdf

A Git installation

For GNU/Linux distributions, Git should be available in the standard system repository. For example, in Debian/Ubuntu please type in the **terminal**:

```
$ sudo apt-get install git
```

If you need to install Git from source, you can get it from git-scm.com/downloads.

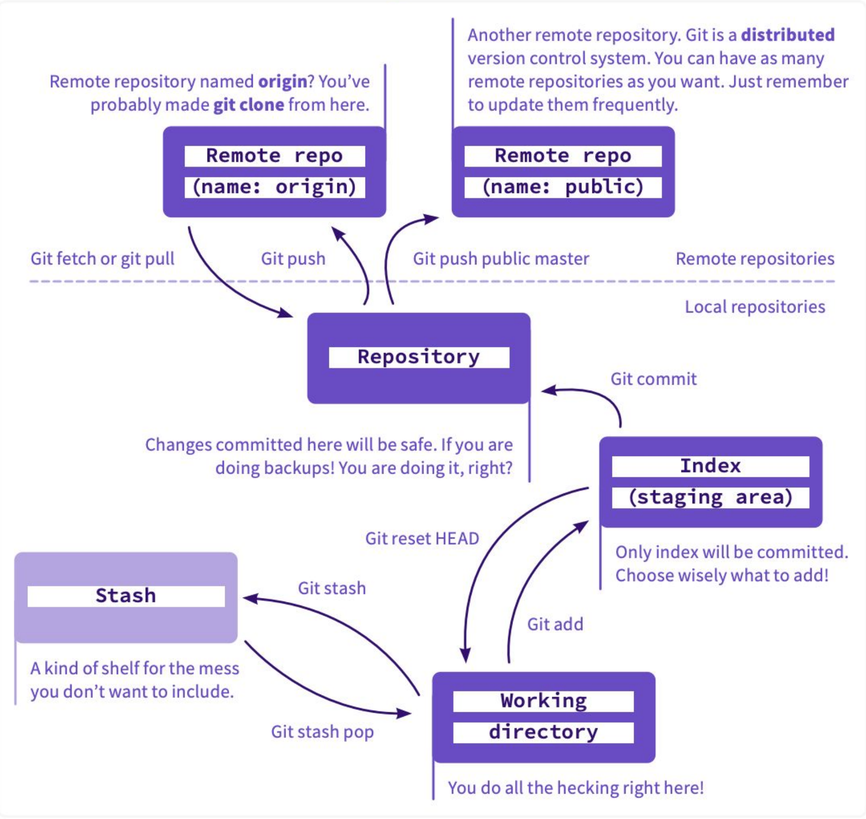
An excellent Git course can be found in the great **Pro Git** book by Scott Chacon and Ben Straub. The book is available online for free at git-scm.com/book.

B Ignoring Files

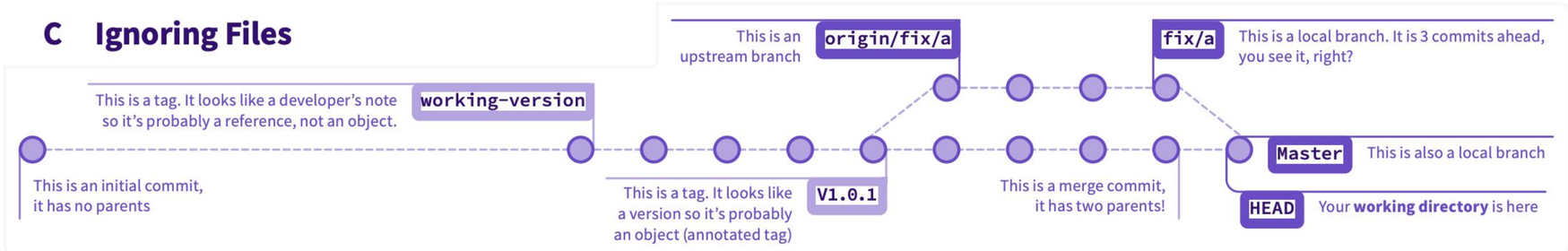
```
$ cat .gitignore
/logs/*
!logs/.gitkeep
/tmp
*.swp
```

Verify the .gitignore file exists in your project and ignore certain type of files, such as all files in **logs** directory (excluding the **.gitkeep** file), whole **tmp** directory and all files ***.swp**. File ignoring will work for the directory (and children directories) where **.gitignore** file is placed.

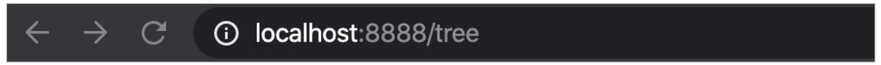
D The zoo of working areas



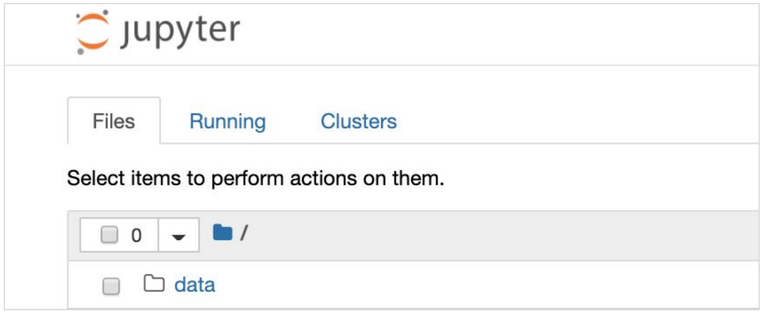
C Ignoring Files

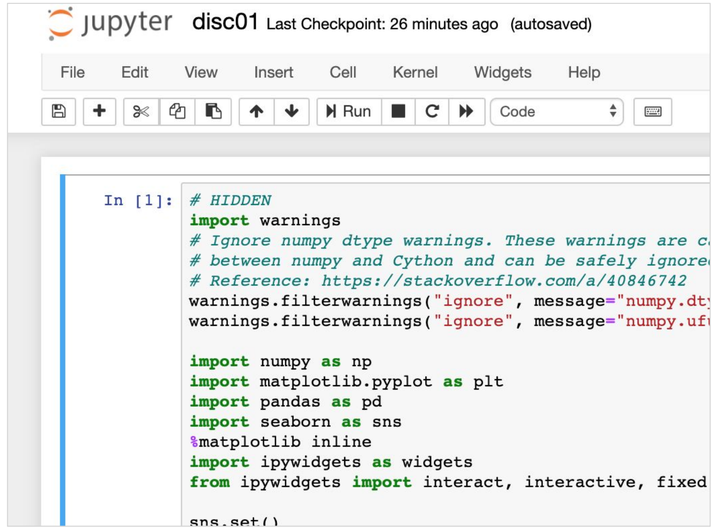


Jupyter



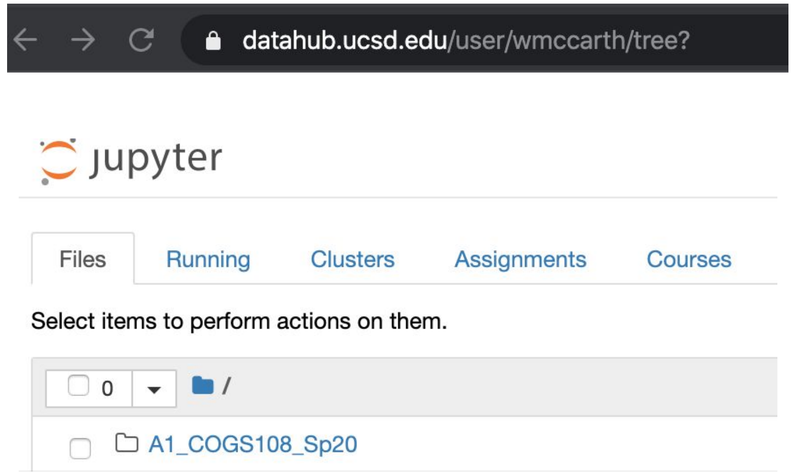
- Python code is run on a python interpreter
- Jupyter is a program that creates an interface for typing python code in a browser, that also runs that code in a python interpreter
- What does this mean?!
  - Jupyter is a way of running python programs from a browser (like chrome) (hooray!)





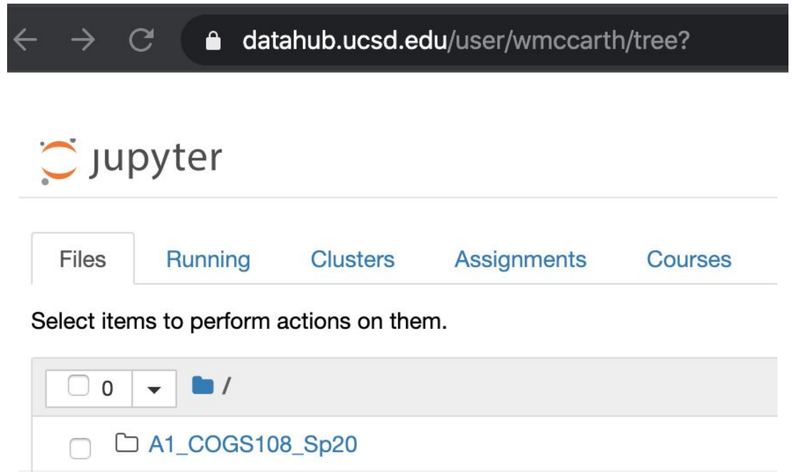
datahub.ucsd.edu

- Jupyter runs python code in a browser.
  - But Jupyter is itself just a program that's running on a computer somewhere.
- datahub lets you interact with Jupyter that's running somewhere else.



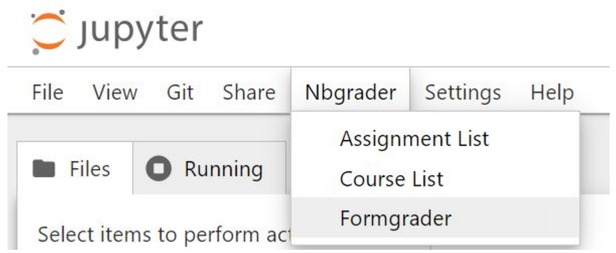
datahub.ucsd.edu

- What does this mean?!
  - You don’t need to worry about installing Jupyter
  - You can use datahub to create and run python programs (online)
  - You can use this interface to fetch and submit assignments



Working on your assignments

- Log into datahub.ucsd.edu
- Go to Assignments tab (or Nbgrader->Assignment List if you are using the new container)
- ‘fetch’ assignments you have access to -> Submit after completion
- Demo of this workflow



Your time to …

- Talk to your classmates to find potential teammates!
- Work on PracticeAssignment and D1