Lecture 4:
Processing Linguistic Data

LING 1340/2340: Data Science for Linguists

Na-Rae Han

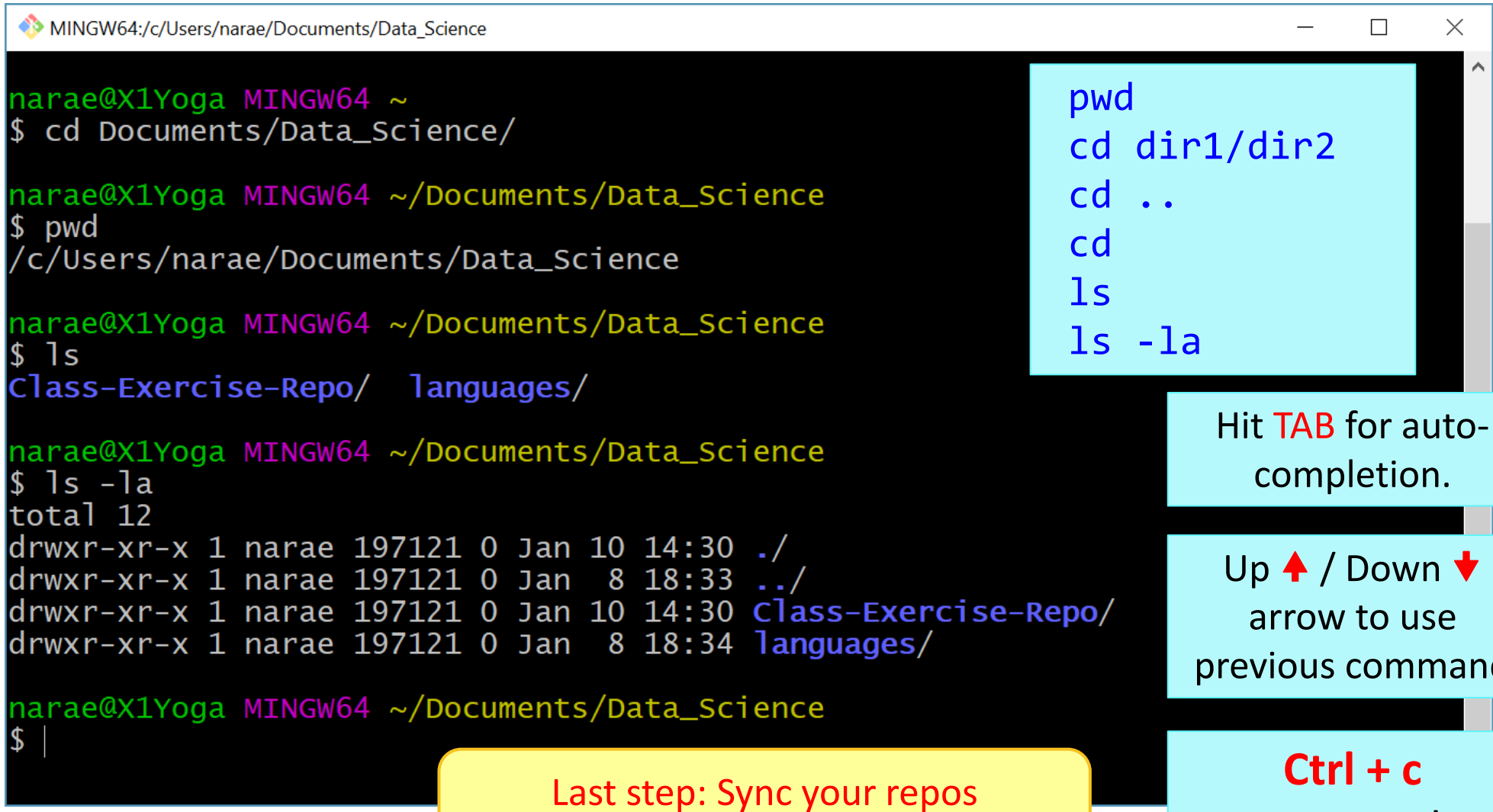
Objectives

- Homework 1: What linguistic datasets did you all process?

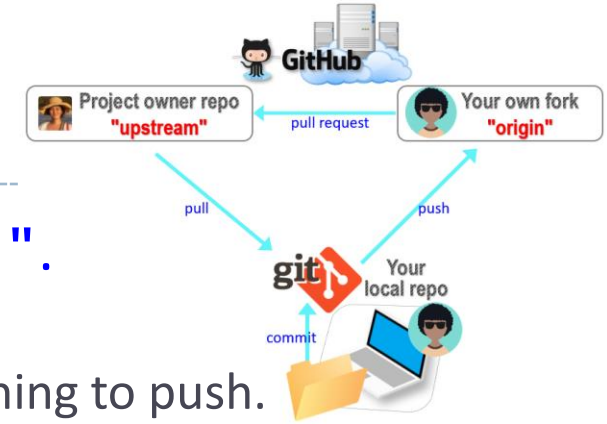
- Tools:
  - Git and GitHub
  - Jupyter Notebook
  - Using DataCamp tutorials

You should be taking NOTES!

First thing to do every class



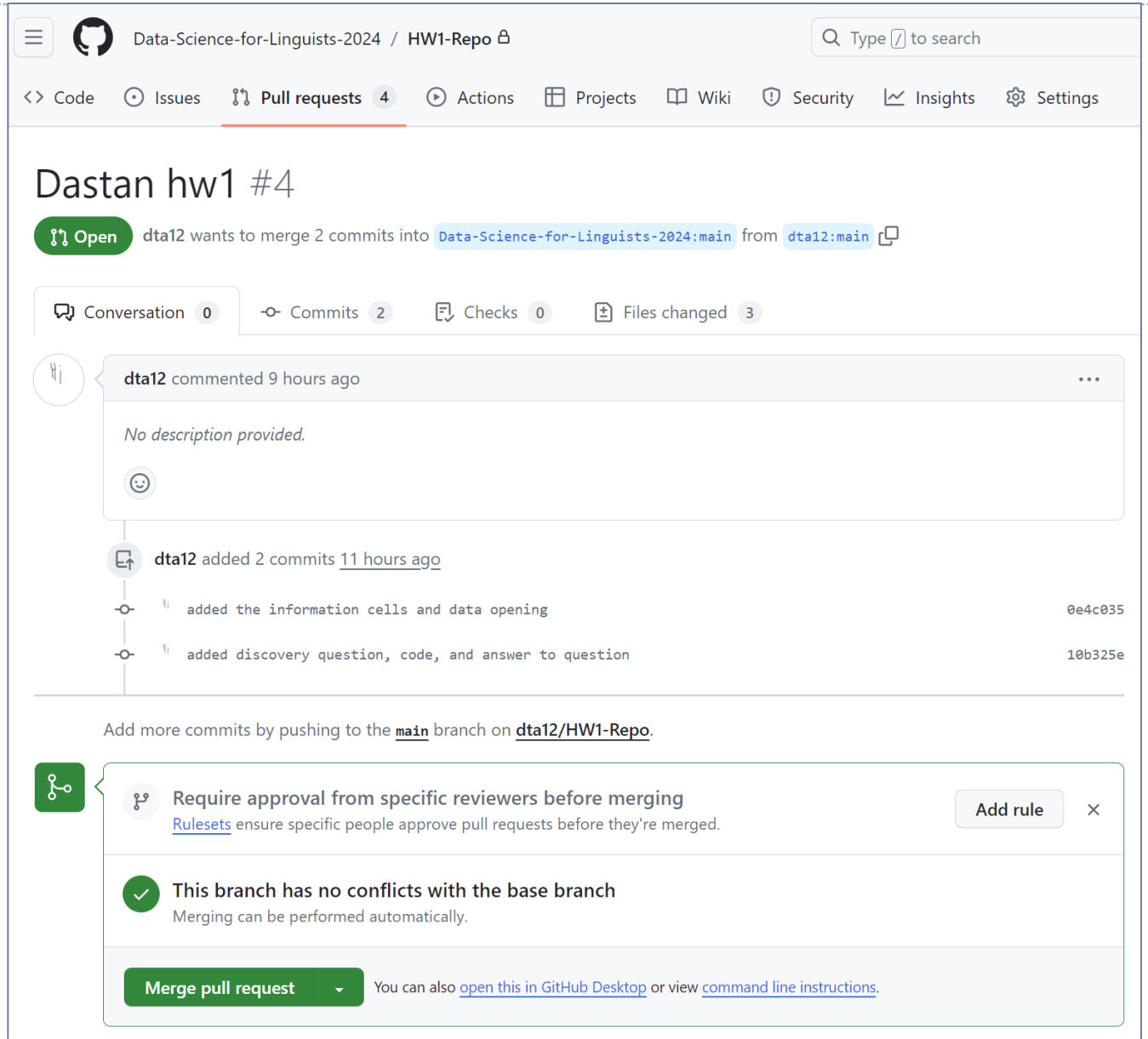
Your workflow



1. **Housekeeping**: Check YOUR WORK via `"git status"`.
   - Your local repo is clean: no unsaved/uncommitted work.
   - Your GIHUB fork already has your latest commit: there's nothing to push.
2. **Housekeeping**: Bring in updates from OTHERS.
   - On your **GitHub fork**, check what updates have accumulated in the upstream repo.
   - Through "Sync fork → Compare", make sure those updates don't have conflicts with your fork. Don't press that green "Update Branch" button!
   - Back on **command line**, pull from upstream. Now your local repo is synced with the original repo.
   - Finally, sync your GitHub fork by pushing. The universe is in order now!
3. Work on your homework, to-do, etc.
   - *Now* start your homework. Make some commits along the way.
   - Push to your GitHub fork for one last time.
   - Submission time: Create a **pull request**. Make sure your pull request doesn't have conflicts.

HW1: processing pull request, merging

- With everyone working on their own files/folders, merging is conflict-free:



HW1: sync your HW1-Repo

1. Configure "upstream" remote:

   ```
   git remote add upstream https://github.com/Data-Science-for-Linguists-2024/HW1-Repo.git
   ```

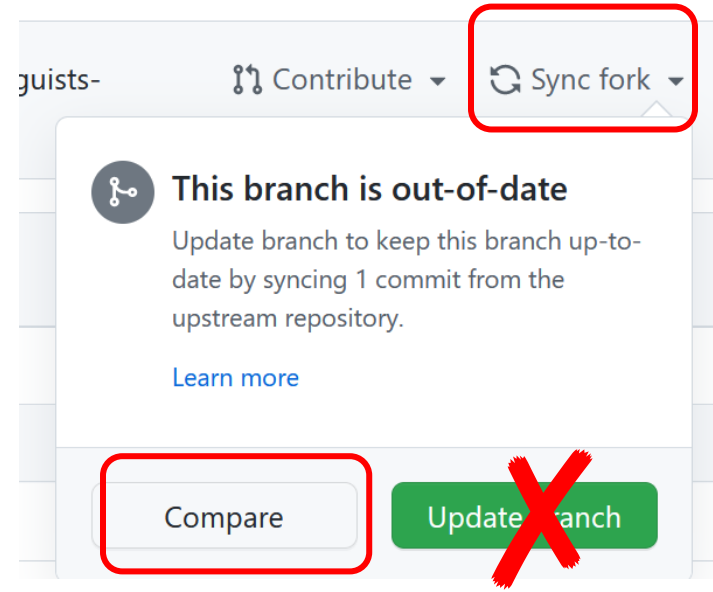
2. Check your GitHub fork, make sure there are no conflicts with upstream

3. Pull from upstream:

   ```
   git pull upstream main
   ```

4. Push to your GitHub fork:

   ```
   git push
   ```



Everyone's repos are synced.

Now, everyone has everyone's homework submission.

HW1: Review

- What did you all work on?
- You wish list: what new skills would you like to learn?
- What is the `.gitignore` file?
- Why did we exclude data files from Git?
- What is up with that "your_file_here.txt" blank file? What is `git rm`?
- Jupyter Notebook: do you like it?

Term Project overview

- Go over the term project guidelines:
  - https://naraehan.github.io/Data-Science-for-Linguists-2024/project

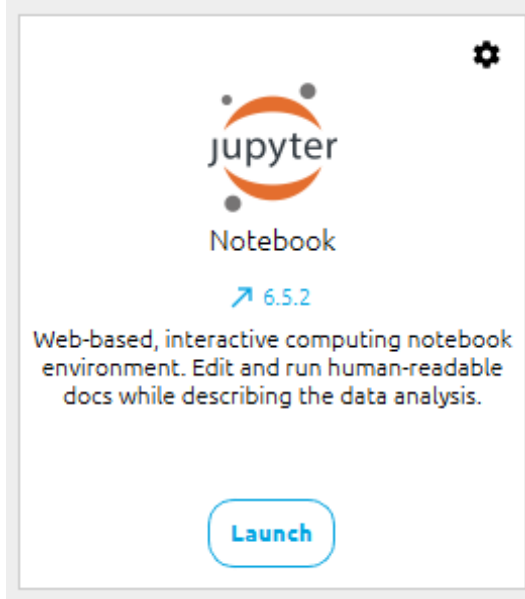
- Also talk about what makes a good/successful project, ways to come up with project ideas

Jupyter Notebook

- Allows you to create and share documents that contain live code cells, output, equations, visualizations and explanatory text.
- Learn how to use it. Your Python code should be in the Jupyter Notebook format:
  - xxxx.ipynb
- You can launch it from the command line.
  - Move into the desired directory, and then execute

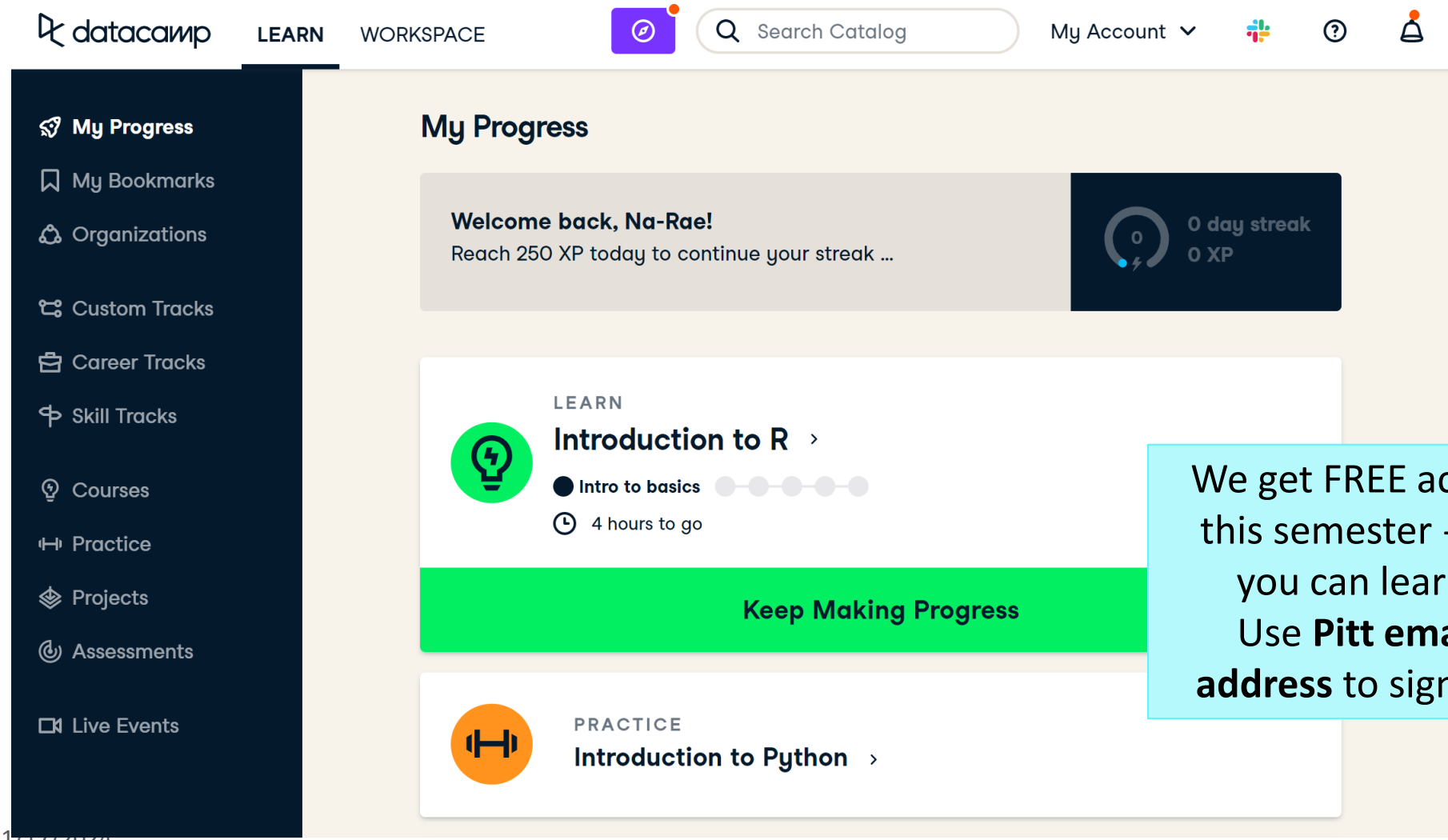
    `jupyter notebook &`

    ←'&' is not necessary, but it lets you keep using the terminal
  - If it doesn't work, then edit your system's path variable or just use a shortcut provided by your OS.



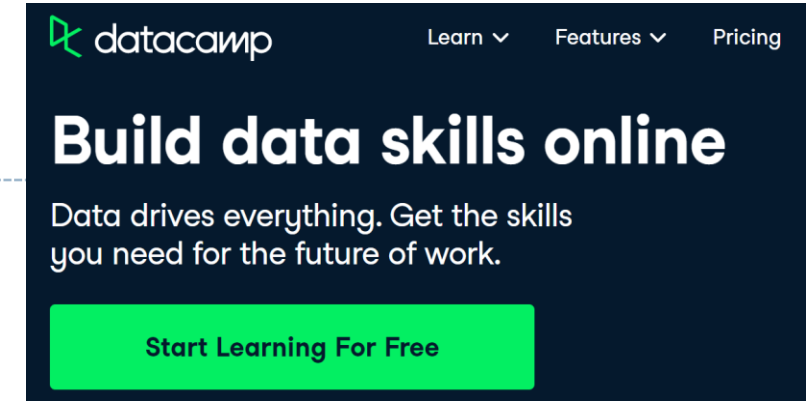
Course Group on DataCamp

- Video-based, interactive tutorials



We get FREE access this semester -- all you can learn! Use **Pitt email address** to sign up.

How to use DataCamp



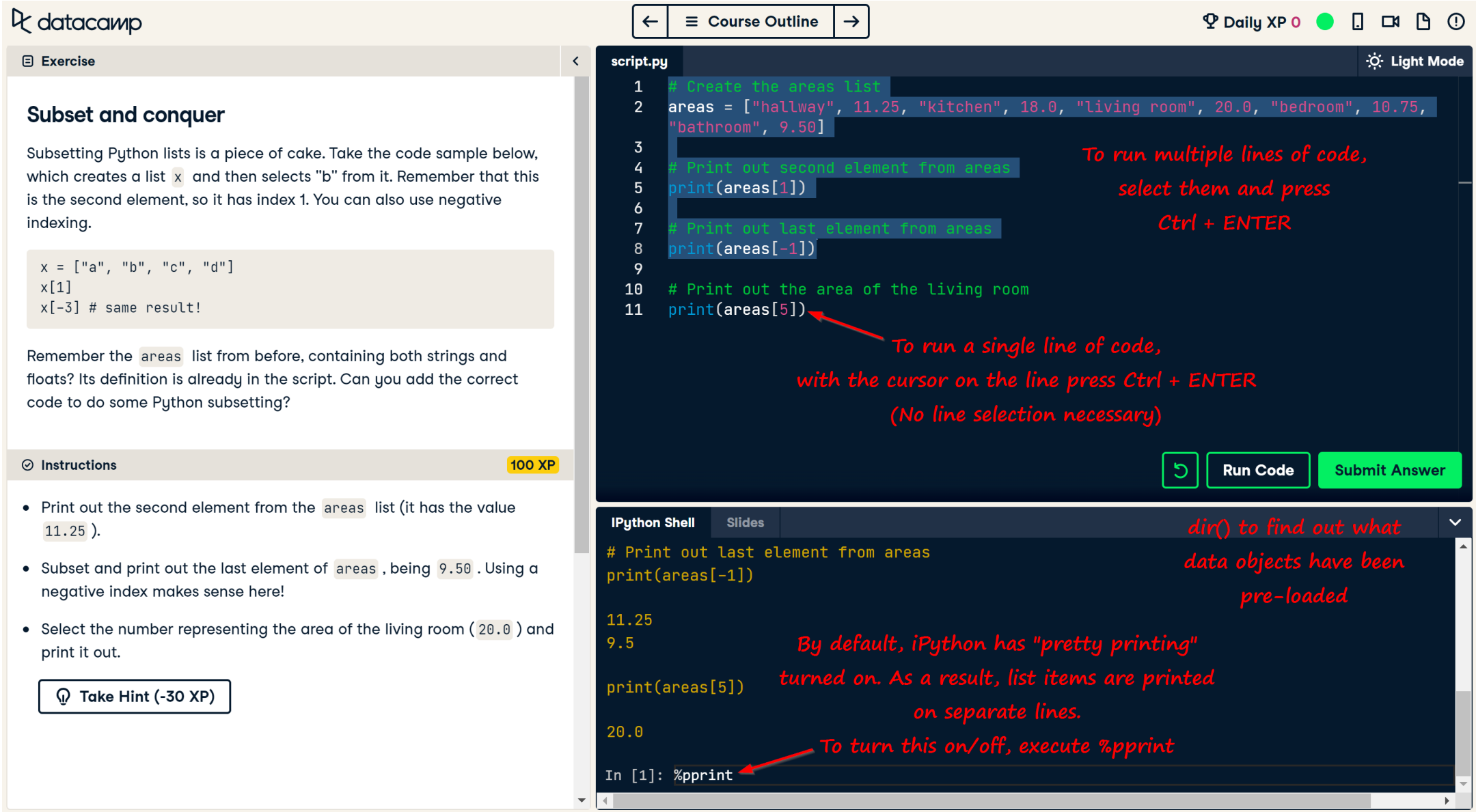
- Topics for the next couple of weeks:
  - numpy library
  - pandas library
  - visualization libraries such as `matplotlib`
- Which video tutorials? Find them on our Resources page:

- (DataCamp) Career Track: Data Scientist with Python (includes all courses below and more) [track]
  - (DataCamp) Introduction to Python for Data Science, Ch.4 NumPy [course] [Ch.4 Numpy]
  - (DataCamp) Intermediate Python for Data Science. Focus on Matplotlib, Numpy & Pandas. [course]
  - (DataCamp) Data Manipulation with pandas [course]
  - (DataCamp) Joining Data with pandas [course]

- Great learning resource, but **not mandatory**. They *complement* the textbook nicely.
- Online exercise interface needs some getting used to.

  ➔ next slide

https://campus.datacamp.com/courses/intro-to-python-for-data-science/chapter-2-python-lists?ex=7



Wrapping up

- To-do #3 out
  - Study numpy
- Learn:
  - numpy and pandas. DataCamp has good tutorials.
- Office hours
  - Na-Rae: Tue 2-3:30pm, Wed 1-2:30
  - Ashley: Mon 2:10-4pm, Tue 5:30-7pm, Th 1:10-3:50pm
- CompLing hire job talk this Friday 3pm @G8 CL
- PyLing next Wed 6pm @ 2818 CL and Zoom