Announcements

New classroom

- GHC 5222
- 6 more seats

Waitlist

Website updated with course policies

Announcements

HW1

- Due Tue 2/8

Grading infrastructure

- https://mugrade.datasciencecourse.org/
- Create new account with <andrewid>@andrew.cmu.edu
- More instructions in hw1_get_started

Plan

- Wrap up scraping slides
    - Regular expressions

- Python, jupyter notebooks, Google Colab

- Homework autograding and submission

15-388/688 - Practical Data Science: Jupyter notebook lab

J. Zico Kolter
Carnegie Mellon University

Spring 2021

Slide credits: CMU AI, Zico Kolter

Outline

Python and Jupyter Notebook

Jupyter lab

Outline

Python and Jupyter Notebook

Jupyter lab

Python

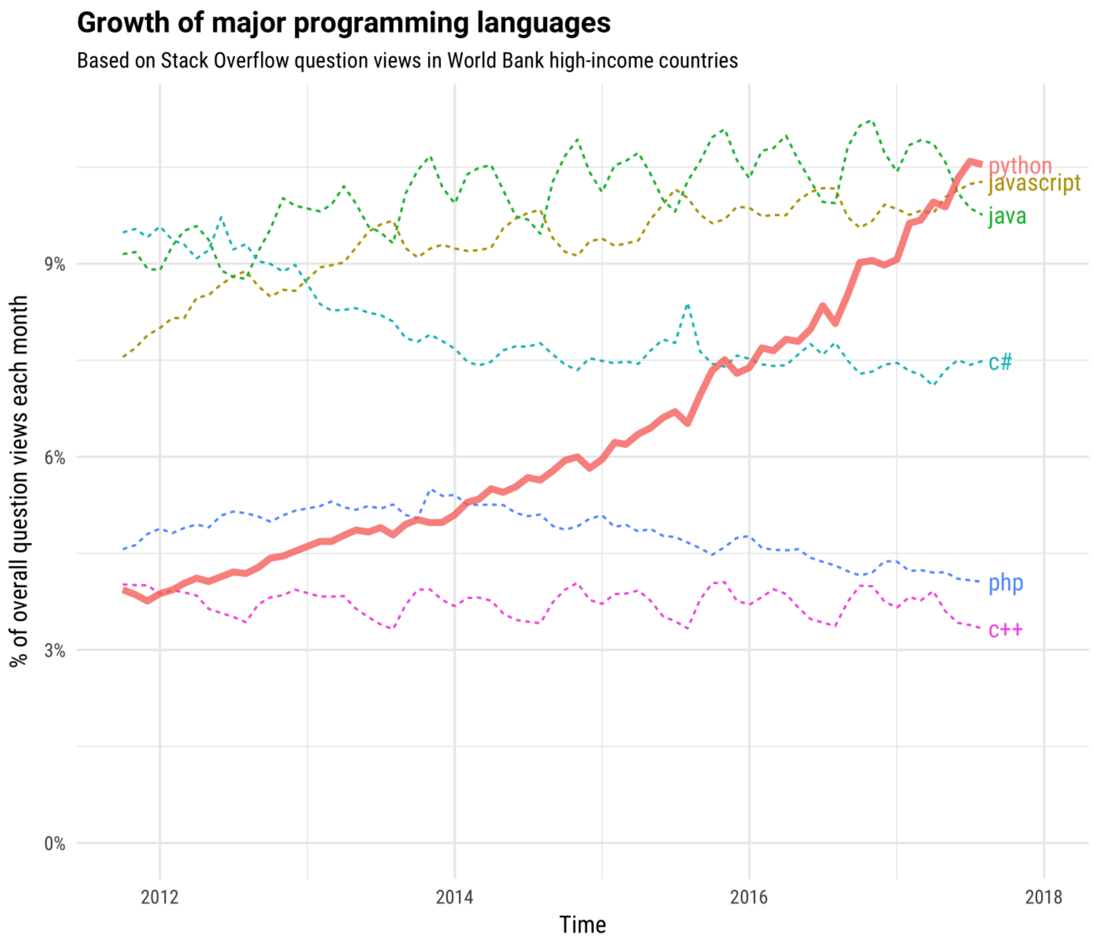
“The language of data science”

- Especially true if the data science tasks involve lots of data processing and/or machine learning
- Less true if the tasks are more “purely statistical” (then R is more standard)

Python 2->3 debacle is behind us

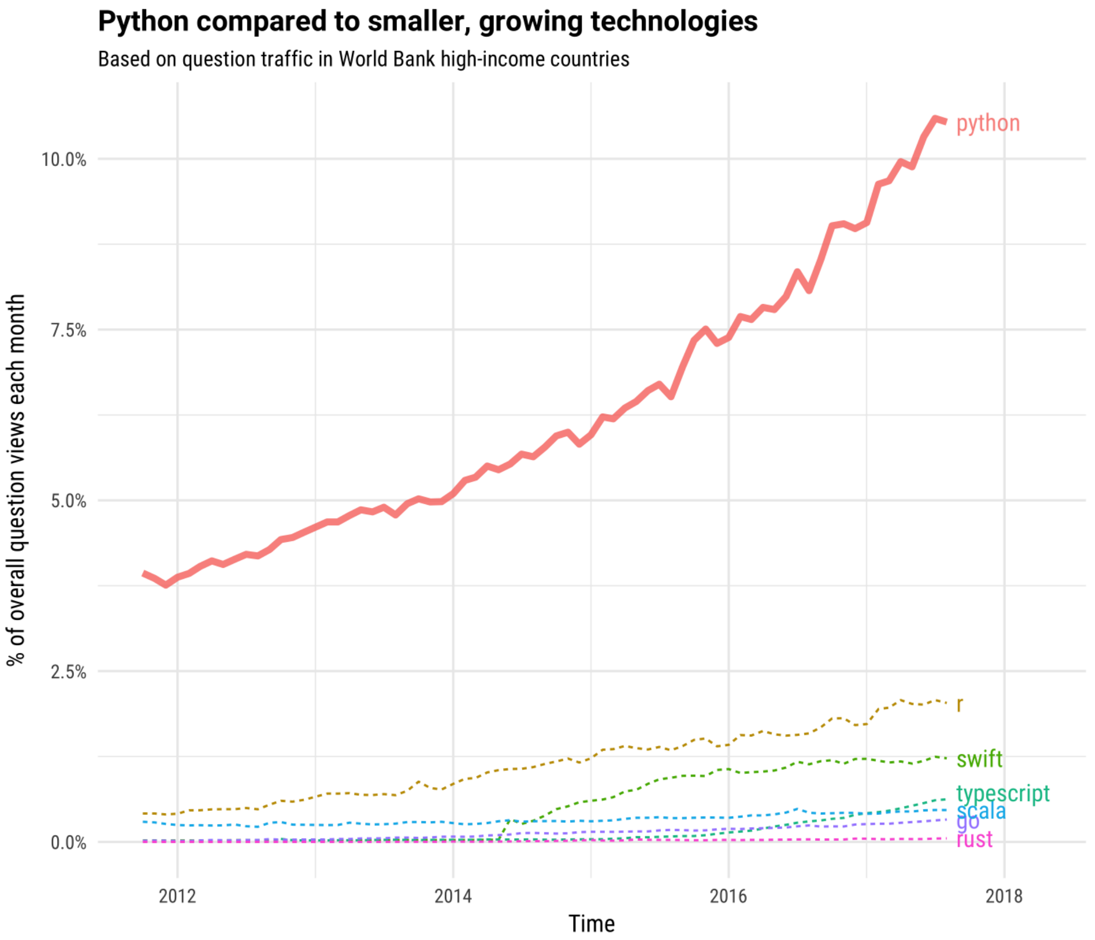
Python growth/popularity happening at the same time as the deep learning boom

Python growth



Source: https://stackoverflow.blog/2017/09/06/incredible-growth-python/

Python growth



Source: https://stackoverflow.blog/2017/09/06/incredible-growth-python/

Anaconda

For this class, we recommend you use Anaconda, a common distribution of Python, which includes several common libraries and tools including the Jupyter notebook and a package manager, available at:

https://www.anaconda.com/download/

You can verify you are using the Anaconda distribution by running Python and making sure you see something like the following:

```
$ python
Python 3.9.7 (default, Sep 16 2021, 08:50:36)
[Clang 10.0.0 ] :: Anaconda, Inc. on darwin
Type "help", "copyright", "credits" or "license" for more information.
>>>
```

Installing additional packages

Several of the homework assignments will require that you have additional libraries

There are two typical ways to install these, via the conda package manager (part of Anaconda), and via pip:

- conda install beautifulsoup4 – install BeautifulSoup4
- conda search beautiful – search conda packages for any that includes the string “beautiful”
- pip install beautifulsoup4 – install BeautifulSoup4
- pip search beautiful – search pip packages for any that include the string “beautiful”

Rule of thumb: use conda when you can (plays nicer with Anaconda installation), but some packages can only be installed via pip

Jupyter notebook

All course assignments (and even the notes) are distributed as Jupyter notebooks

Jupyter notebooks are a browser-based environment for writing code, interspersing code and Markdown, and displaying figures, all contained in “cells”

- More info about Jupyter here: http://www.jupyter.org

Launch jupyter via the command:

- jupyter notebook
- Then it will print instructions to navigate to a specific localhost URL (a browser tab may even open automatically)

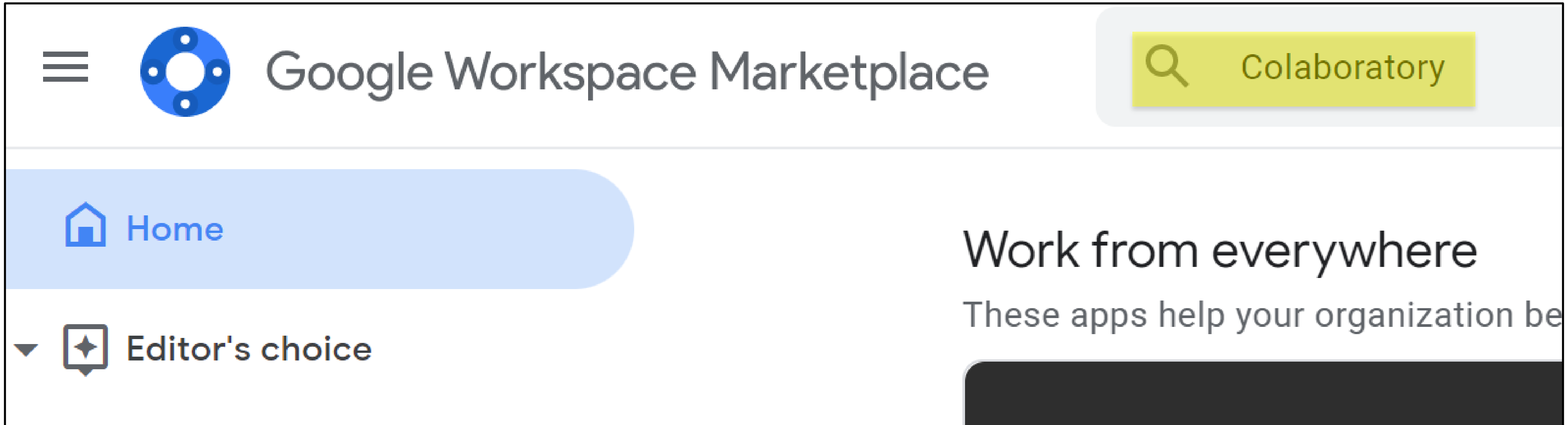
You may also use Colab notebook to complete assignments fully online

Google Colab

Google Colab (Colaboratory) is a Jupyter notebook environment that runs in the browser using Google Cloud.

Easy to open *.ipynb files from Google Drive folders.

Need to install Google Colaboratory from https://workspace.google.com/marketplace



Work from everywhere

These apps help your organization be

Tips for homework

Carefully follow problem specifications to match the output required by the autograder

Test your code locally on the provided test cases *and* additional test cases you create, to ensure it gives the expected output for all inputs you can come up with

You “should” be able to exactly know your score before you even submit, because the code passes all local tests

Outline

Python and Jupyter Notebook

Jupyter lab

Jupyter lab

(Continued in live notebook)