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

J. Zico Kolter
Carnegie Mellon University

Fall 2019

Announcements

Recitation tomorrow (Thursday, **9/5**) from **6-8pm** in Doherty Hall 2210 (this room)

Waitlist will cleared continually throughout the day and tomorrow (there is room in the course)

Homework advice: make sure your code passes included local tests (and ideally, write more tests, and look up the relevant ones on Diderot) **before** submitting to Diderot

Outline

Python and Jupyter Notebook

Jupyter lab

Outline

Python and Jupyter Notebook

Jupyter lab

Python

“The language of data science”

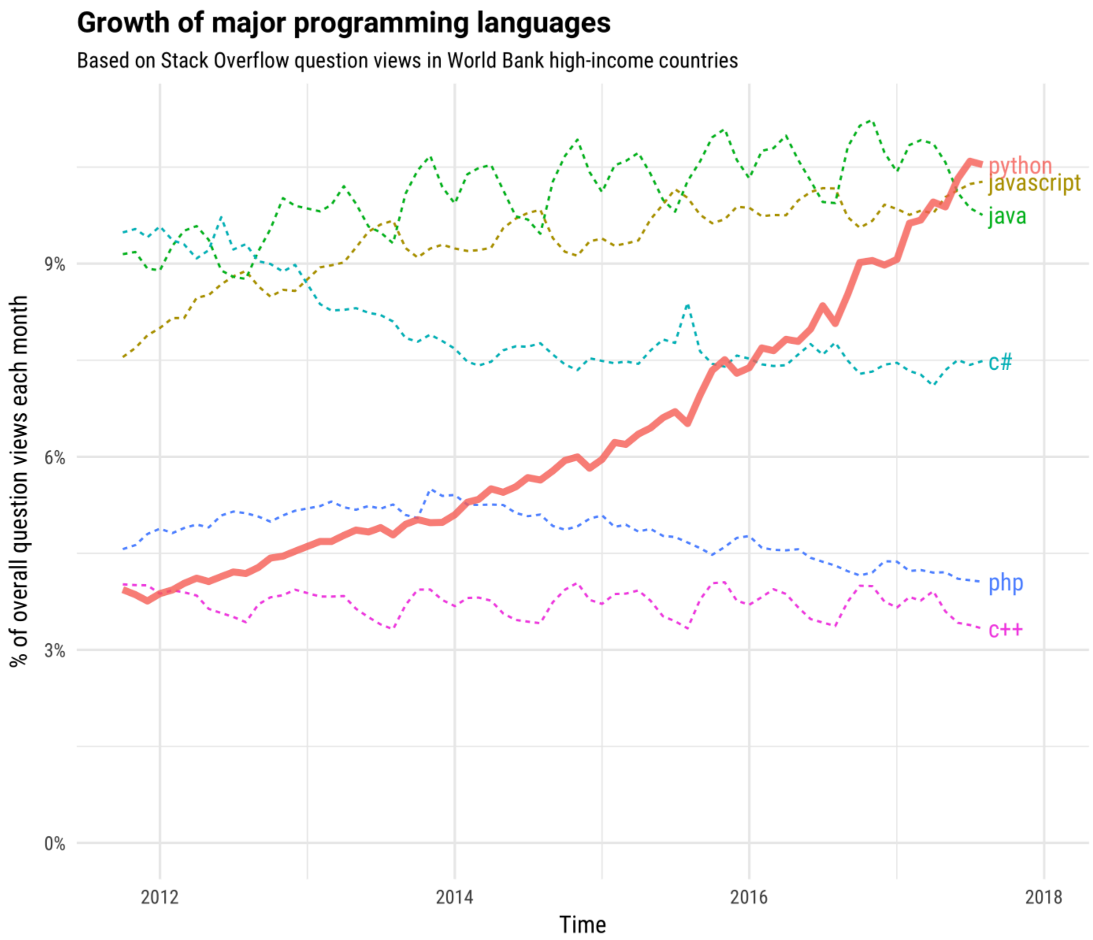
- Especially true if the data science tasks involve lots of data processing and/or machine learning
- Less true if the tasks are more “purely statistical” (then R is more standard)

Python 2->3 debacle

The most visible changes to the language in Python 3 (honestly) are:

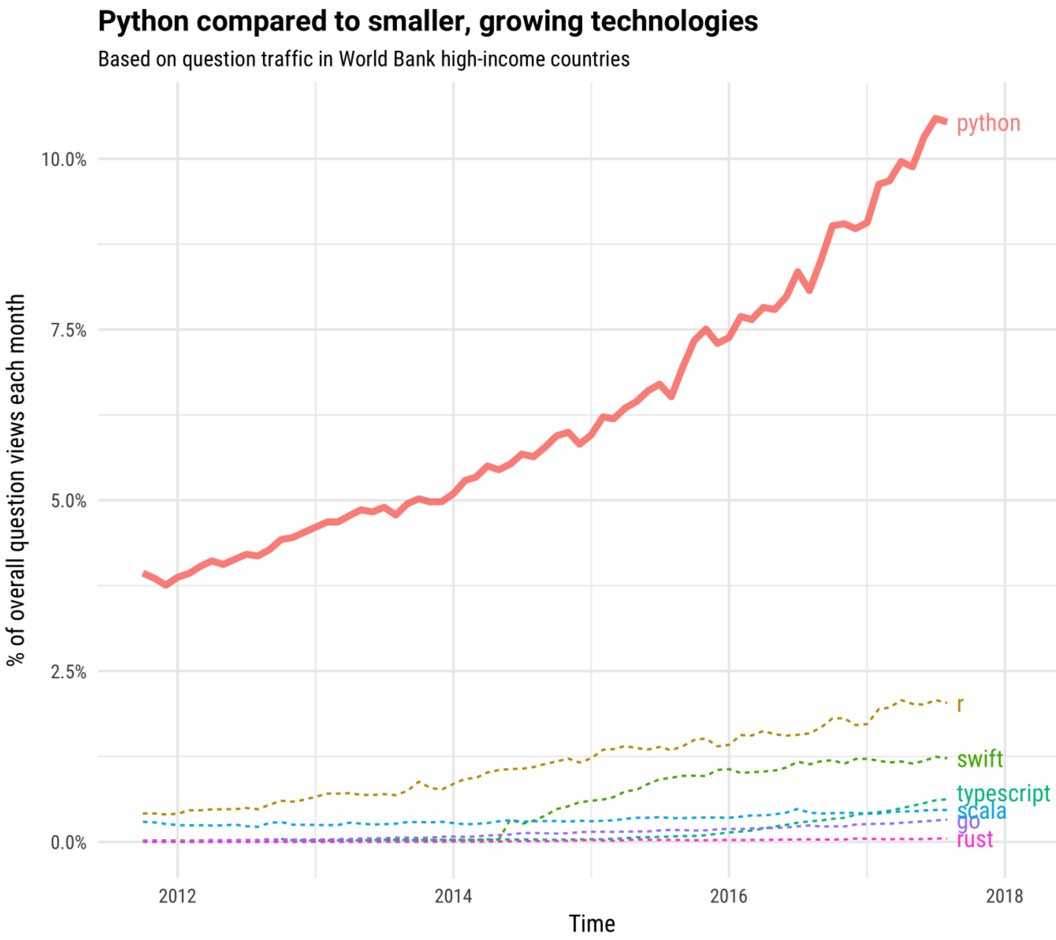
1. `print` is a command, not a statement (so you need parentheses)
2. `1/2` returns 0.5 (floating point), not 0 (integer); to get 0, you use the operation `1//2`

Python growth



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

Python growth



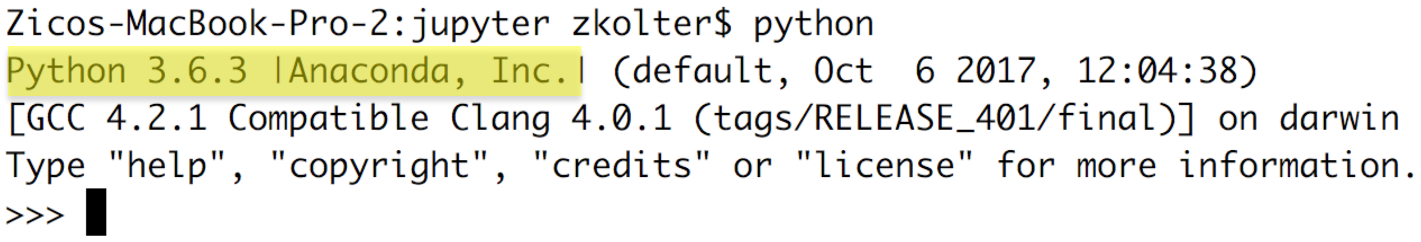
Source: https://stackoverflow.blog/2017/09/06/incredible-growth-python/

Anaconda

For this class, we strongly 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:



```
Zicos-MacBook-Pro-2:jupyter zkolter$ python
Python 3.6.3 |Anaconda, Inc.| (default, Oct  6 2017, 12:04:38)
[GCC 4.2.1 Compatible Clang 4.0.1 (tags/RELEASE_401/final)] 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 navigate to http://localhost:8888 (or possibly a later port number, if you have multiple notebooks open)

Tips for homework

Carefully follow problem specifications to match the output required by Diderot

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 Diderot score before you even submit, because the code passes all local tests (or at least most of the tests)

Outline

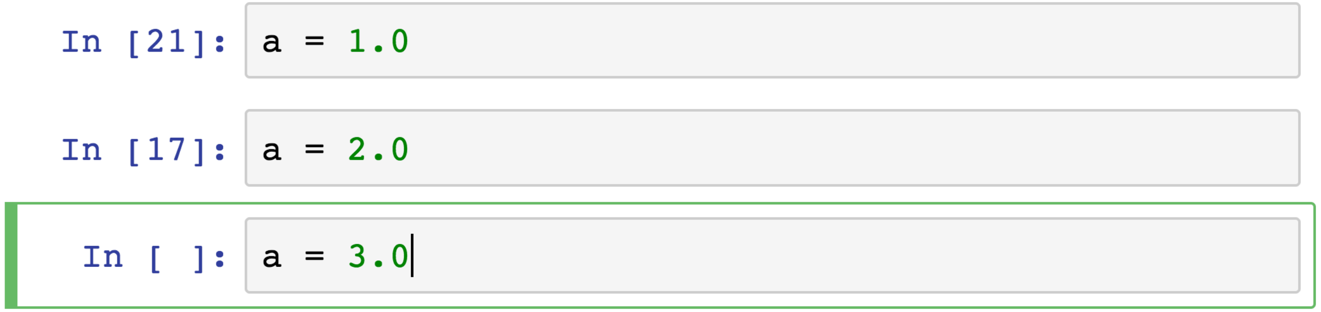
Python and Jupyter Notebook

Jupyter lab

Jupyter lab

(Continued in live notebook)

Poll: Jupyter notebook



```
In [21]: a = 1.0

In [17]: a = 2.0

In [ ]: a = 3.0
```

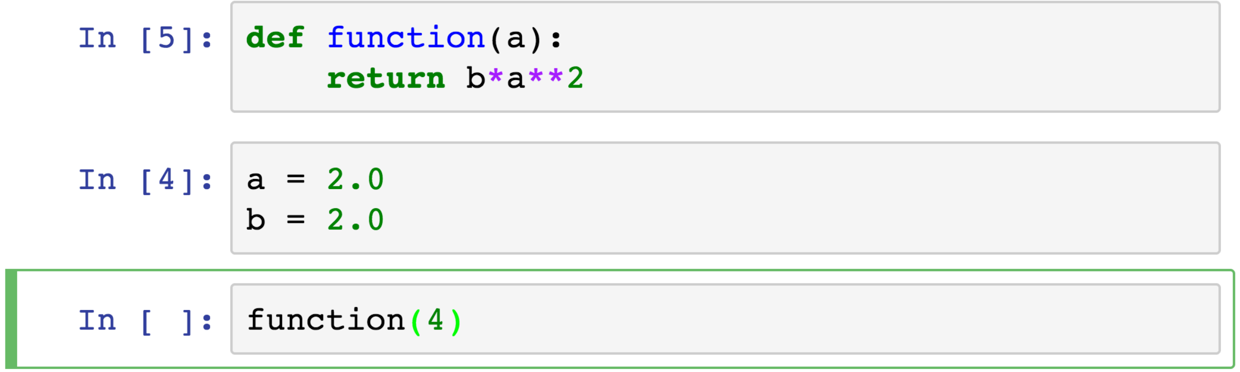
What is the current value of the variable `a` in this notebook (assuming that no other cells exist)?

1. 1.0

2. 2.0

3. 3.0

Poll: Jupyter notebook



```
In [5]: def function(a):
            return b*a**2

In [4]: a = 2.0
        b = 2.0

In [ ]: function(4)
```

What will be the output of the selected cell?

1. 8.0

2. 16.0

3. 32.0

4. Error: “b” is undefined