

# **Python, Colab, and Github**

Akitaka Matsuo  
Department of Government

# **PYTHON, WHY?**

# Programming for data science

- There are two options for data science programming
  - R
  - Python
- This course will use Python. Why?



# Choice criteria

- Which your colleagues to use?
- What do you want to do?
  - There are some stuffs you can't do without Python.
    - Deep learning
    - Connection with web-application
    - Web-scraping
- Which is the standard for the field?

# Python and R

	Python	R
Purpose	General programming language	Statistical language
Community	Default language in Computer Science	Bigger than python in social sciences
Code readability	Easy to read Standardised	Sometime it's hard New standard: tidyverse
Strength	Machine learning Generally faster	CRAN visualisation

# In the end

- You have to learn both at certain level
- At the higher level, you can connect two (but mostly from R to Python)
  - tensorflow
  - spacy

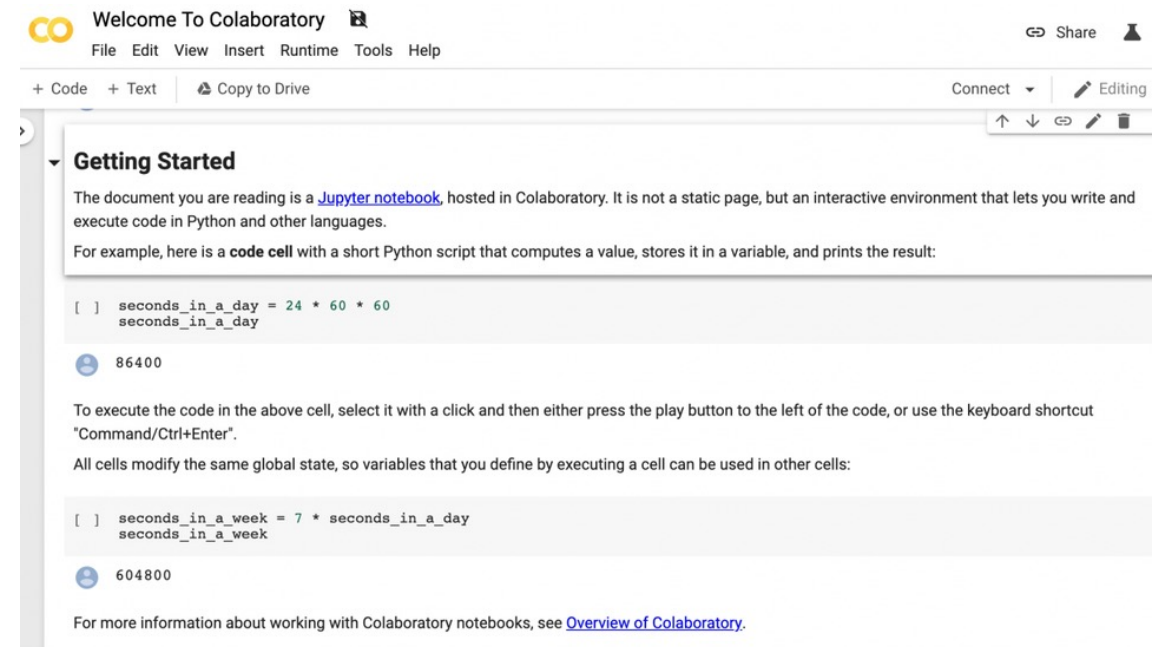
# **HOSTING PYTHON, GITHUB**

# Jupyter notebook

## Raw python

```
akitaka@DESKTOP-7VR78BU:~$ python3
Python 3.8.2 (default, Apr 27 2020, 15:53:34)
[GCC 9.3.0] on linux
Type "help", "copyright", "credits" or "license" for more information.
>>> print("hello world!")
hello world!
>>> a = 1
>>> b = 2
>>> a + b
3
>>> for i in range(10):
...     print(i)
...
0
1
2
3
4
5
6
7
8
9
>>>
```

## Jupyter notebook



The screenshot shows the Colaboratory web interface. At the top, there's a header with the Colaboratory logo, the text "Welcome To Colaboratory", and a "Share" button. Below this is a menu bar with "File", "Edit", "View", "Insert", "Runtime", "Tools", and "Help". A secondary bar contains "+ Code", "+ Text", "Copy to Drive", "Connect", and "Editing". The main content area is titled "Getting Started" and contains the following text:

The document you are reading is a [Jupyter notebook](#), hosted in Colaboratory. It is not a static page, but an interactive environment that lets you write and execute code in Python and other languages.

For example, here is a **code cell** with a short Python script that computes a value, stores it in a variable, and prints the result:

```
[ ] seconds_in_a_day = 24 * 60 * 60
    seconds_in_a_day
```

86400

To execute the code in the above cell, select it with a click and then either press the play button to the left of the code, or use the keyboard shortcut "Command/Ctrl+Enter".

All cells modify the same global state, so variables that you define by executing a cell can be used in other cells:

```
[ ] seconds_in_a_week = 7 * seconds_in_a_day
    seconds_in_a_week
```

604800

For more information about working with Colaboratory notebooks, see [Overview of Colaboratory](#).



# Hosting python

- We will use python exclusively in this course
- There are many options to run python codes
  - Cloud
    - We will use Google Colab
    - Later we will see how to use python on other cloud environment
  - Local
    - If you want to work in your local environment
      - Install Anaconda
      - [https://www.youtube.com/watch?v=-sNX\\_ZMVpQM](https://www.youtube.com/watch?v=-sNX_ZMVpQM) (or search “install anaconda and jupyter”)

# Google Colab

- Colab (Colaboratory)
- A free Jupyter notebook environment
  - No setup
  - No fee
- Restrictions
  - Running time
  - Deletion of working directory
- Additional advantage
  - Free GPU/TPU (but very limited these days)

# Google Colab

- Google drive integration:
  - Notebooks are saved in your Google drive
  - You can access files in Google drive
- To use Colab, you need a gmail account.
  - If you don't have one, please make one before the class this week
  - Also you need to create a git account

# Git/Github

- **Git** = a **version control** system
  - System keeps the record of file changes
  - Files are hosted online server
    - Easy to collaborate with other people
  - Version control
    - Tracking file changes
    - Revert to the previous versions
    - Create branches
- **Github**: A popular hosting service of git
  - Plus: gui, private/public hosting, etc...

# Git/Github, basics

- **Repository:** A collection of codes for specific purpose
  - Research project
  - Package development (Python, R, etc)
- **Clone:** Create a local copy of repository
- **Commit:** Create a snapshot of updates
- **Push:** Upload the update on the server
- **Fork:** Get a copy of someone's repository into your account
- These works are usually done in command line, but using Github and Colab, we will do some work by just clicking...