Programming for Data Science

Lab01-Getting started

Instructor: Jakramate Bootkrajang

Outline

- Python distribution
- Programming environments
- Google colaboration
  - Some exercises
- Interesting websites

About Python

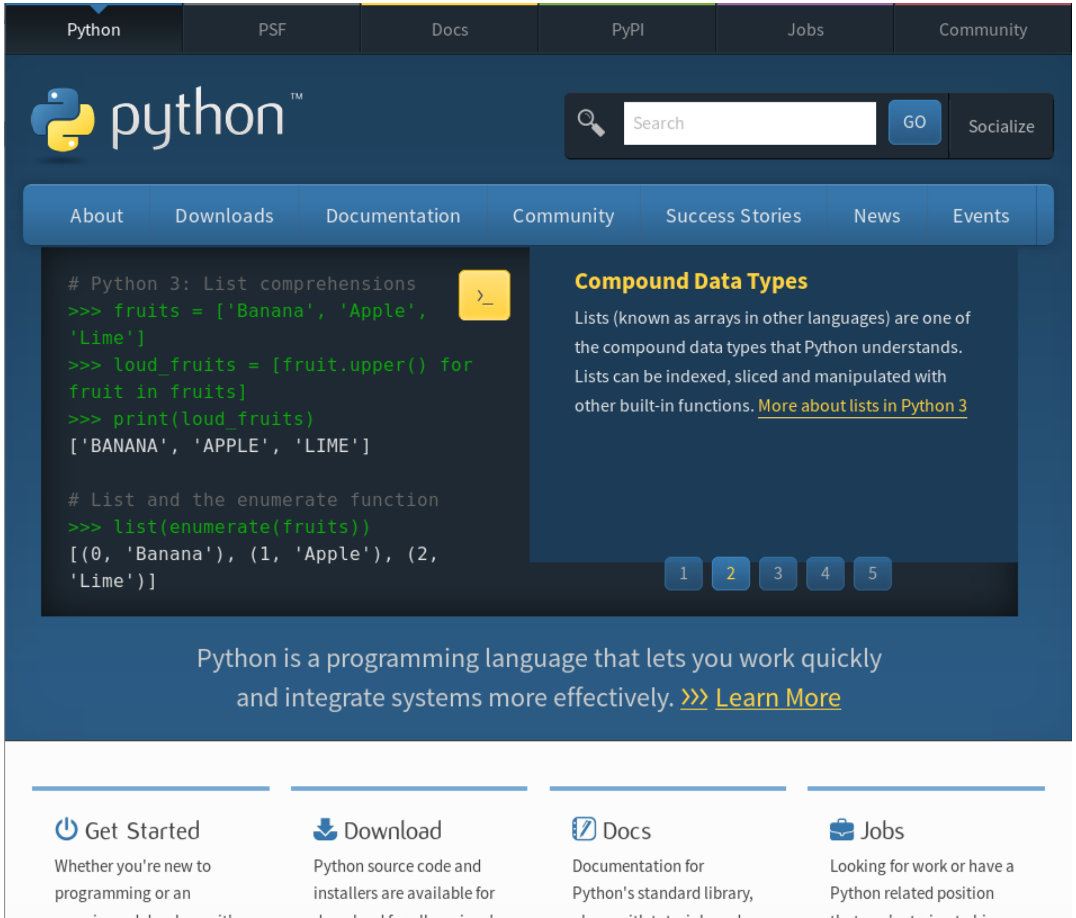
- Python 2.xx
  - Older version of Python
  - Still being used in legacy programs
- Python 3.xx
  - Newer version
  - We will use Python 3.xx in this class

Python distribution

- A distribution is
  - A set of Python interpreter + additional packages
- The most widely used ones are
  - CPython distribution (standard)
  - Anaconda distribution (CPython + packages for data science)

CPython distribution

- Standard distribution
- Can be downloaded from www.python.org



```
# Python 3: List comprehensions
>>> fruits = ['Banana', 'Apple',
'Lime']
>>> loud_fruits = [fruit.upper() for
fruit in fruits]
>>> print(loud_fruits)
['BANANA', 'APPLE', 'LIME']

# List and the enumerate function
>>> list(enumerate(fruits))
[(0, 'Banana'), (1, 'Apple'), (2,
'Lime')]
```

Compound Data Types

Lists (known as arrays in other languages) are one of the compound data types that Python understands. Lists can be indexed, sliced and manipulated with other built-in functions. More about lists in Python 3

Python is a programming language that lets you work quickly and integrate systems more effectively. >>> Learn More

Get Started

Whether you're new to programming or an

Download

Python source code and installers are available for

Docs

Documentation for Python's standard library,

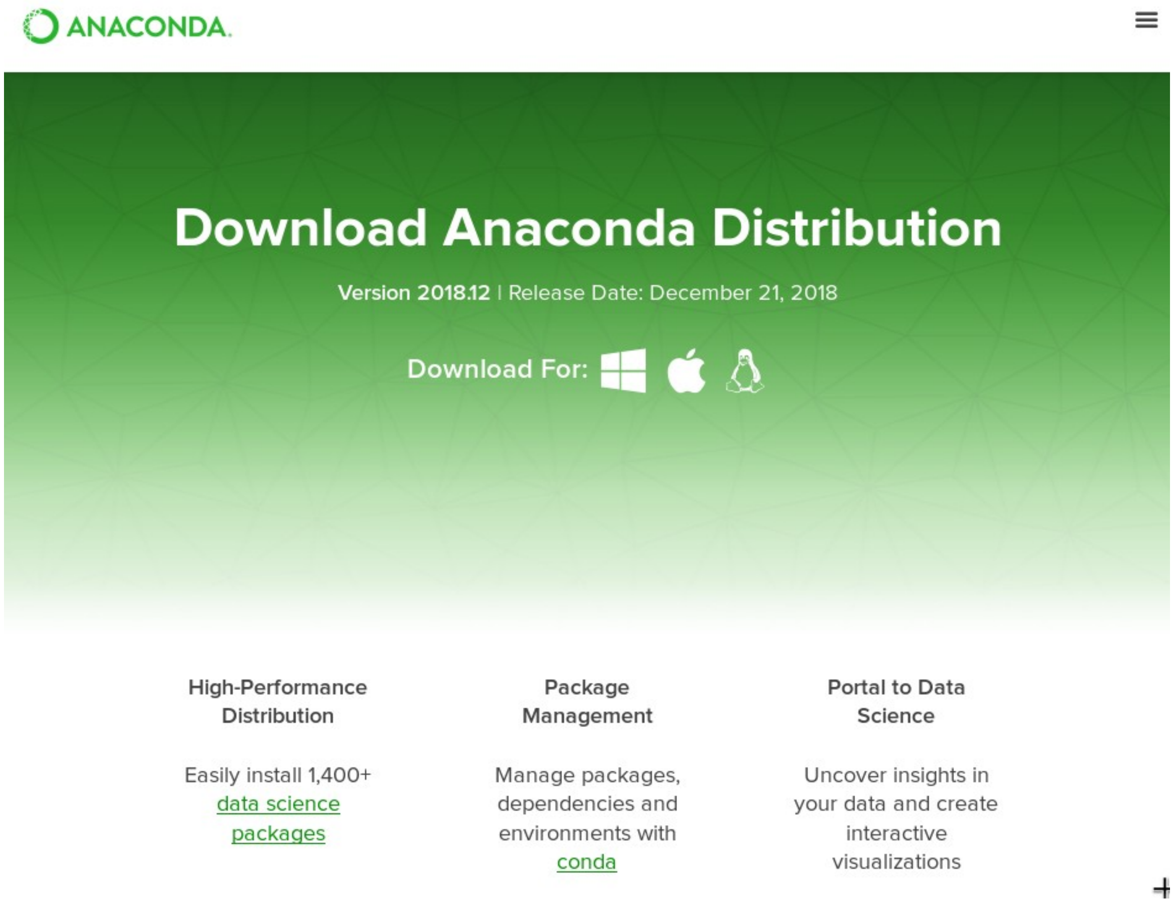
Jobs

Looking for work or have a Python related position

Anaconda distribution

- Standard distribution + appox. 1400 packages
- Those packages are essential for data science, data mining, machine learning, etc.
- There is a smaller version of Anaconda called miniconda
- Can be downloaded from
  - www.anaconda.com

Snapshot of the webpage



ANACONDA.

Download Anaconda Distribution

Version 2018.12 | Release Date: December 21, 2018

Download For:

High-Performance Distribution

Easily install 1,400+ data science packages

Package Management

Manage packages, dependencies and environments with conda

Portal to Data Science

Uncover insights in your data and create interactive visualizations

Benefit of using Anaconda

- Support multiple virtual environments
- Excellent package manager named conda
- Conflicting packages can be easily avoided
  - You can install two versions of Numpy on the same computer, but in different virtual environments.

Running Python

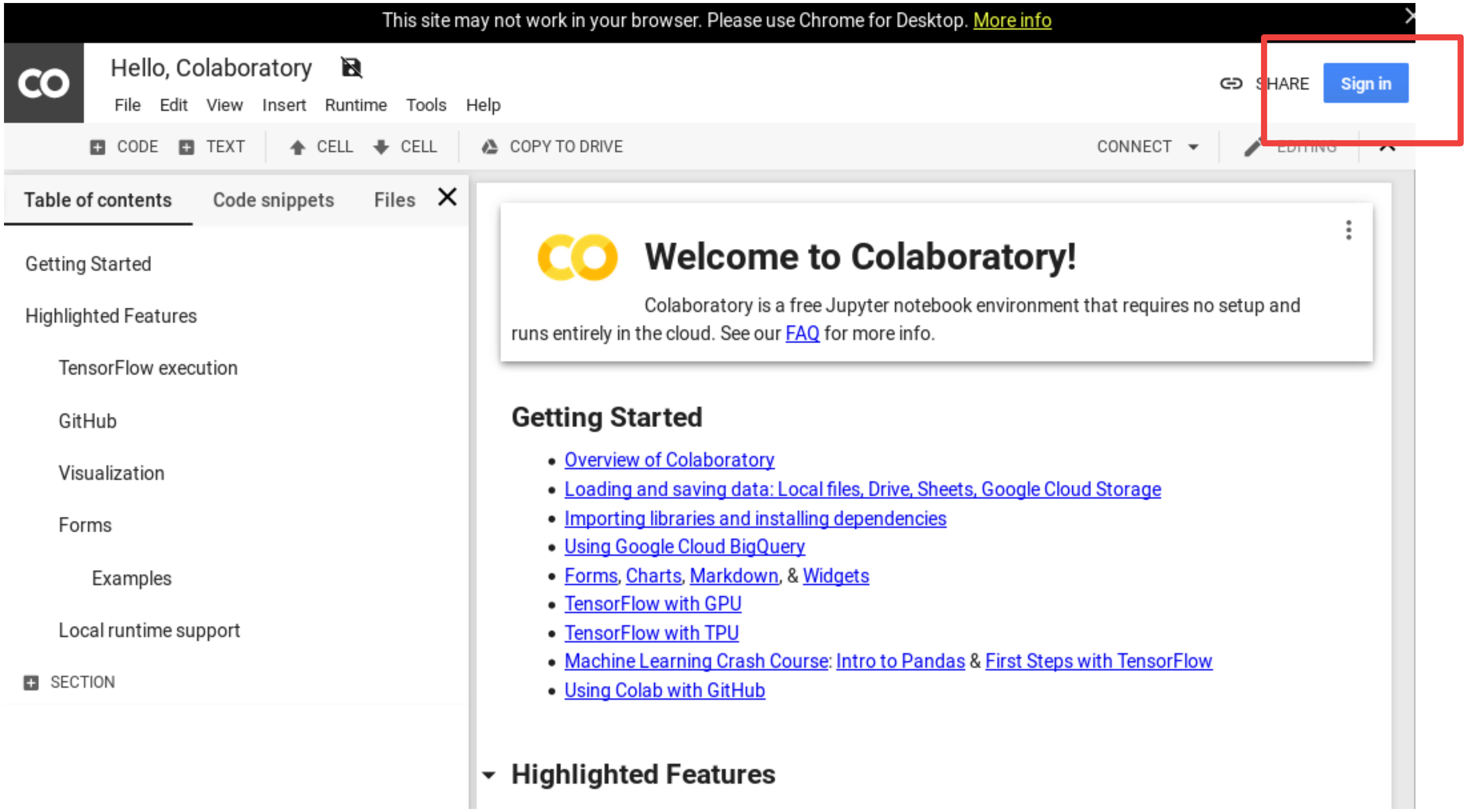
- On personal computer
  - Requires installation of python distribution
- On cloud computing service
  - Does not require software installation
  - Your codes are with you all the time
  - Usually free

Cloud computing service

- Microsoft Azure
  - Available for CMU students for free
- Google Colaboration (colab in short)
  - Free
  - Requires Google’s account (gmail)
  - Automatically links to Google drive

Google Colaboration

- Visit https://colab.research.google.com
- And sign in with your Google’s credential



This site may not work in your browser. Please use Chrome for Desktop. More info

Welcome to Colaboratory!

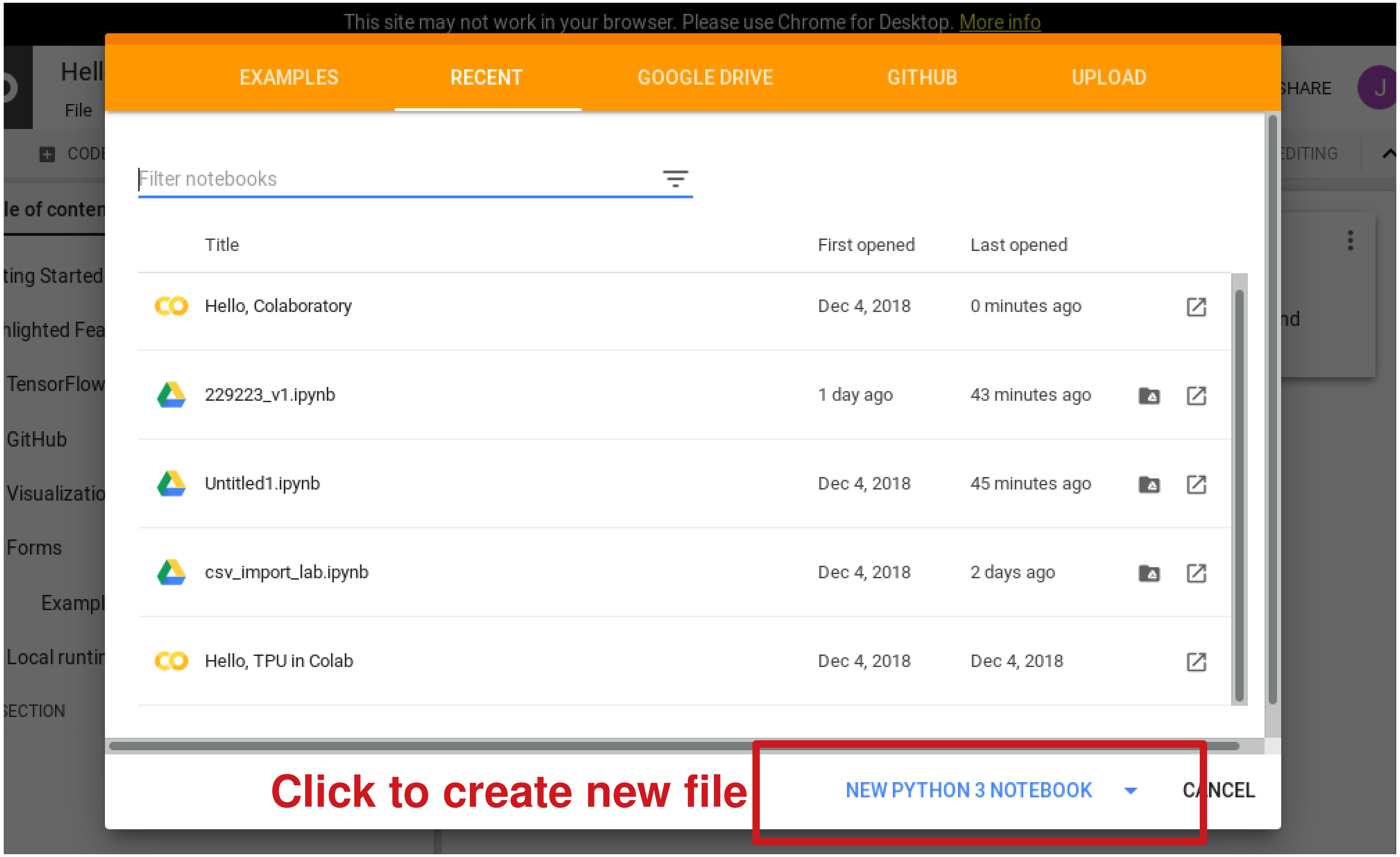
Colaboratory is a free Jupyter notebook environment that requires no setup and runs entirely in the cloud. See our FAQ for more info.

Getting Started

- Overview of Colaboratory
- Loading and saving data: Local files, Drive, Sheets, Google Cloud Storage
- Importing libraries and installing dependencies
- Using Google Cloud BigQuery
- Forms, Charts, Markdown, & Widgets
- TensorFlow with GPU
- TensorFlow with TPU
- Machine Learning Crash Course: Intro to Pandas & First Steps with TensorFlow
- Using Colab with GitHub

Highlighted Features

After signing in



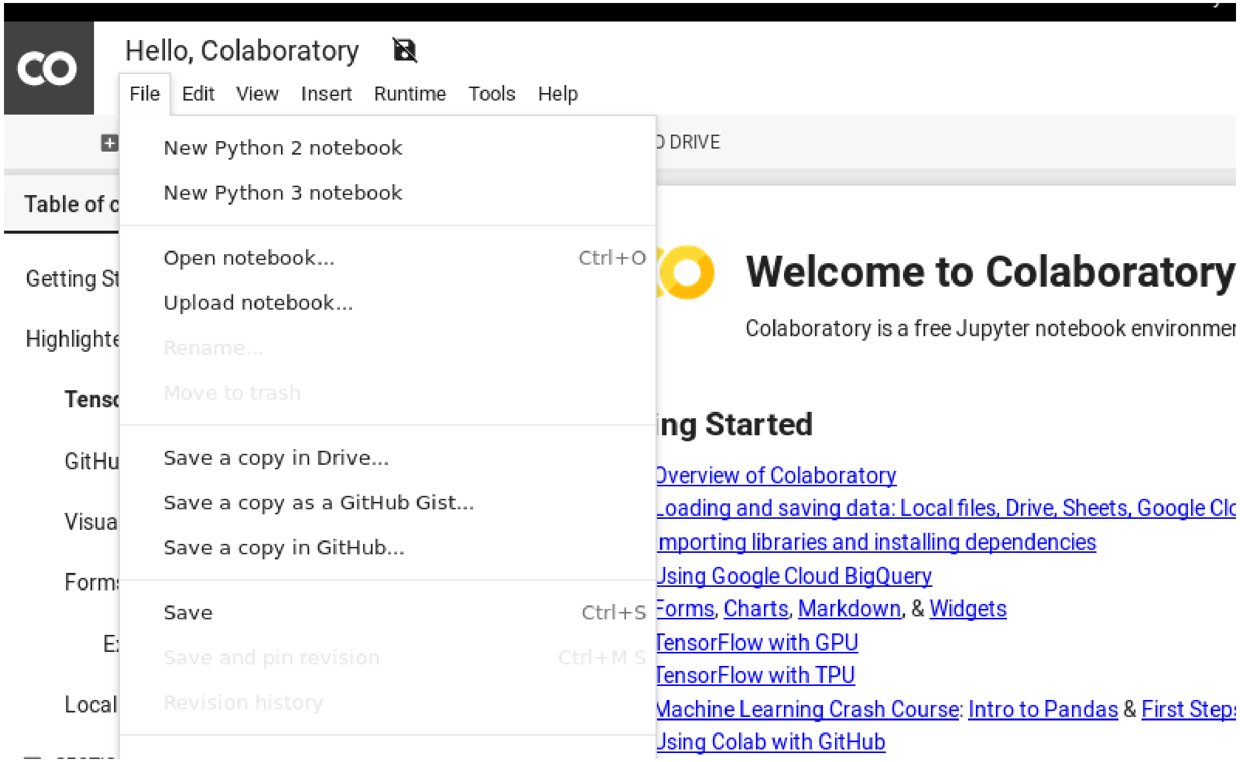
This site may not work in your browser. Please use Chrome for Desktop. More info

Title	First opened	Last opened
Hello, Colaboratory	Dec 4, 2018	0 minutes ago
229223_v1.ipynb	1 day ago	43 minutes ago
Untitled1.ipynb	Dec 4, 2018	45 minutes ago
csv_import_lab.ipynb	Dec 4, 2018	2 days ago
Hello, TPU in Colab	Dec 4, 2018	Dec 4, 2018

Click to create new file

Alternatively

- You can create new file by choosing file menu on the top-left corner
  - File → New Python 3 notebook



Hello, Colaboratory

File Edit View Insert Runtime Tools Help

New Python 2 notebook
New Python 3 notebook
Open notebook... Ctrl+O
Upload notebook...
Rename...
Move to trash
Save a copy in Drive...
Save a copy as a GitHub Gist...
Save a copy in GitHub...
Save Ctrl+S
Save and pin revision Ctrl+M S
Revision history

Welcome to Colaboratory

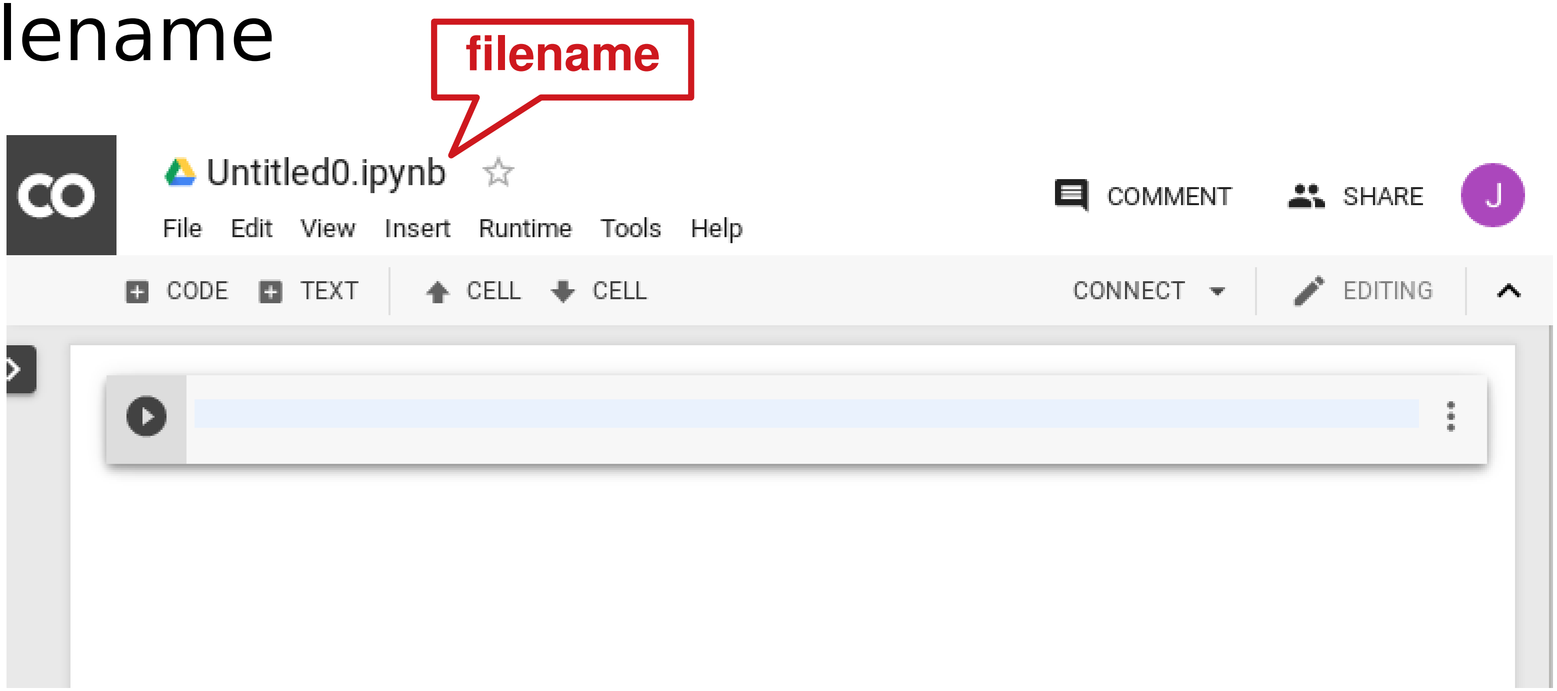
Colaboratory is a free Jupyter notebook environmer

ing Started

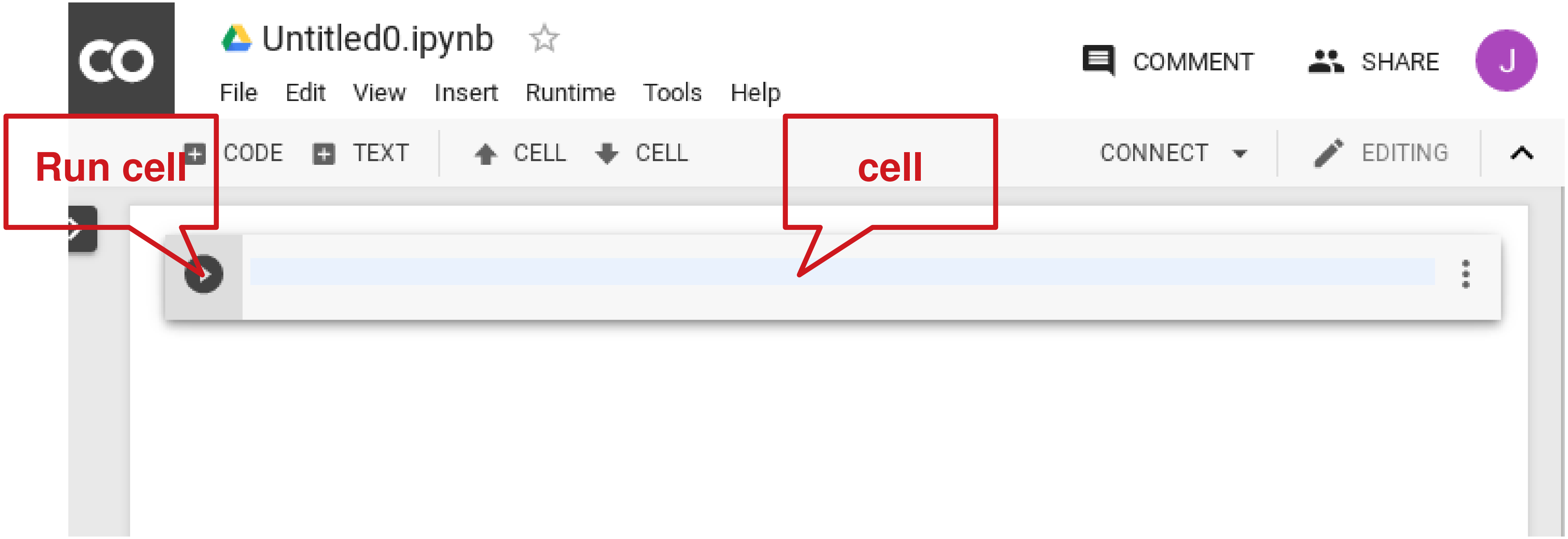
Overview of Colaboratory
Loading and saving data: Local files, Drive, Sheets, Google Clo
mporting libraries and installing dependencies
Jsing Google Cloud BigQuery
Forms, Charts, Markdown, & Widgets
TensorFlow with GPU
TensorFlow with TPU
Machine Learning Crash Course: Intro to Pandas & First Steps
Jsing Colab with GitHub

Your new Python file

- A Python sourcecode is called a notebook
- You can rename the file by double click filename

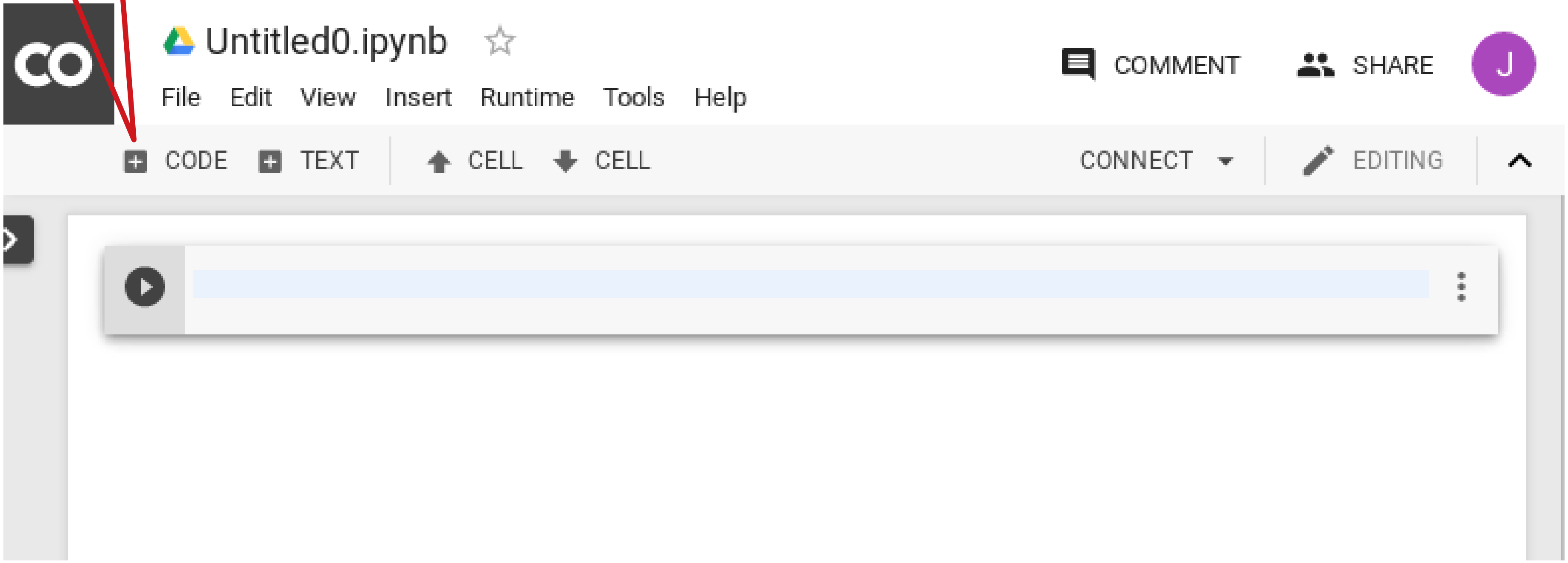


What’s in a notebook ?

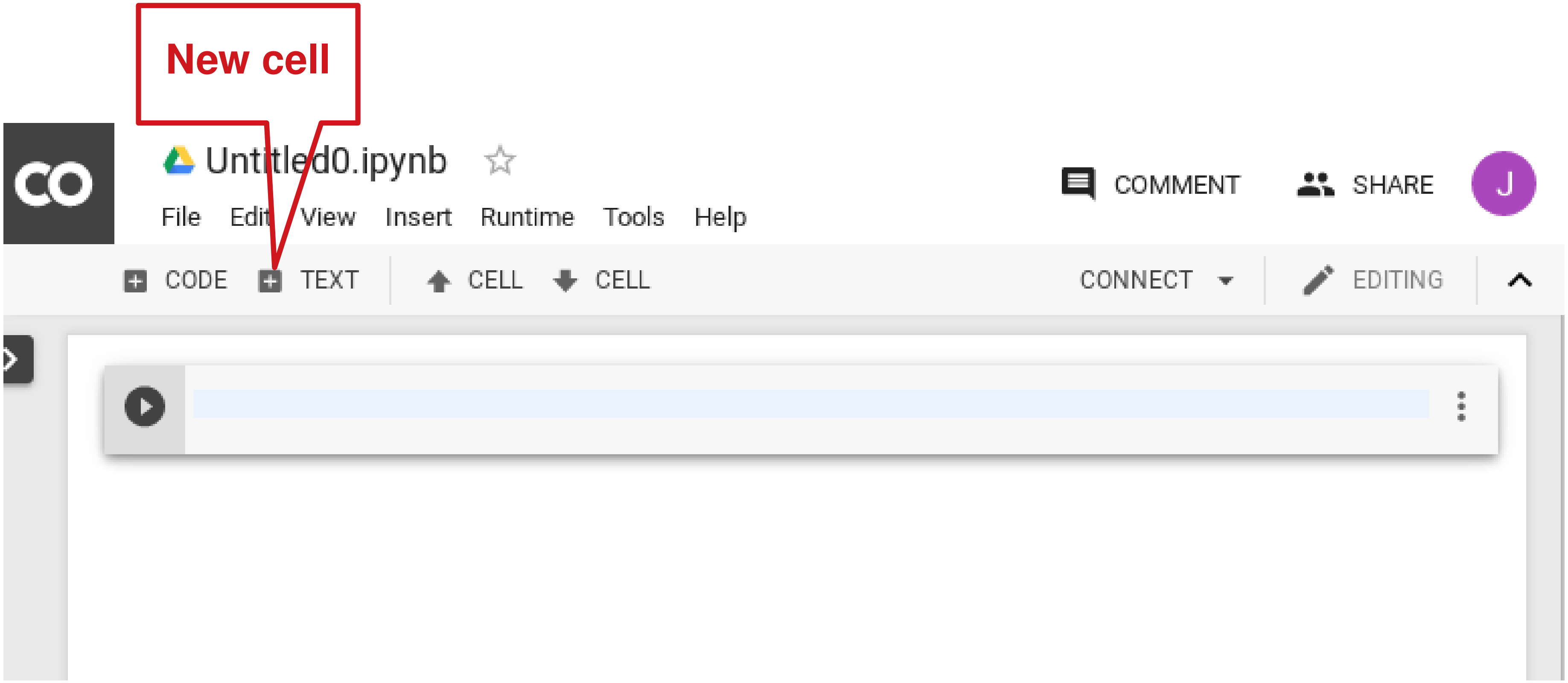


Adding new code cell

New cell



Adding new text cell

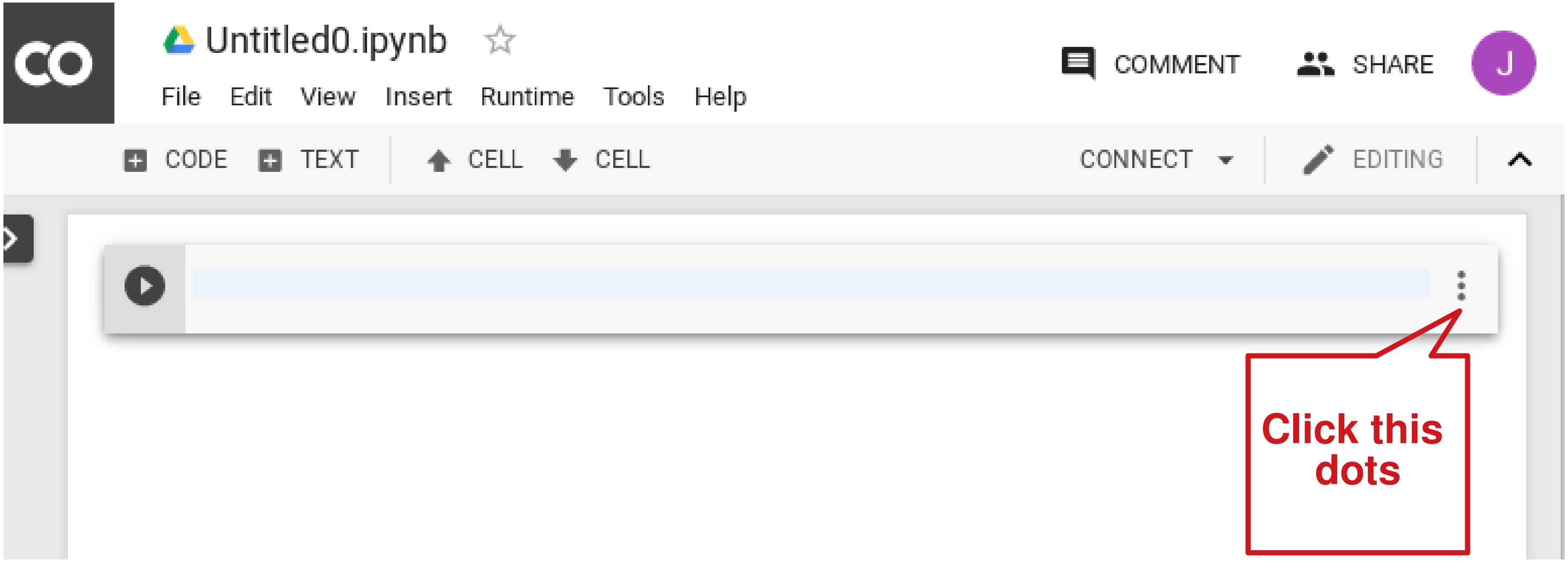


Text cell is useful for adding context to the code

Exercise 1

- Add one code cell and one text cell

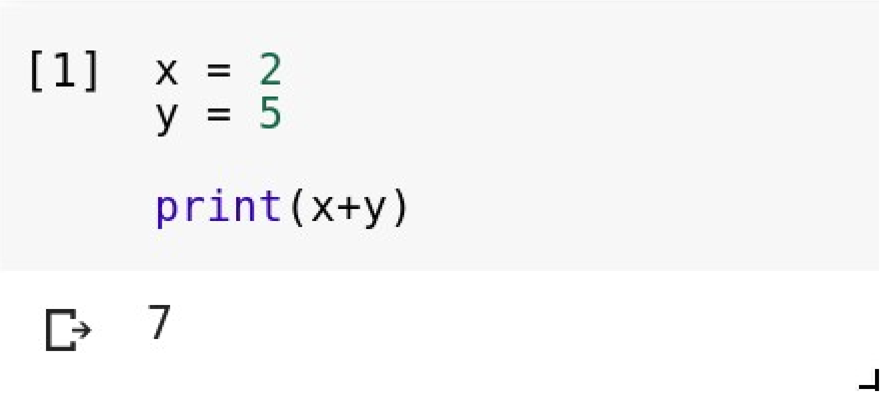
Deleting cell



Click this dots

Running a cell

- Click the `play’ button in front of cell
- Or press `CTRL-ENTER’
- Let’s try running the following code

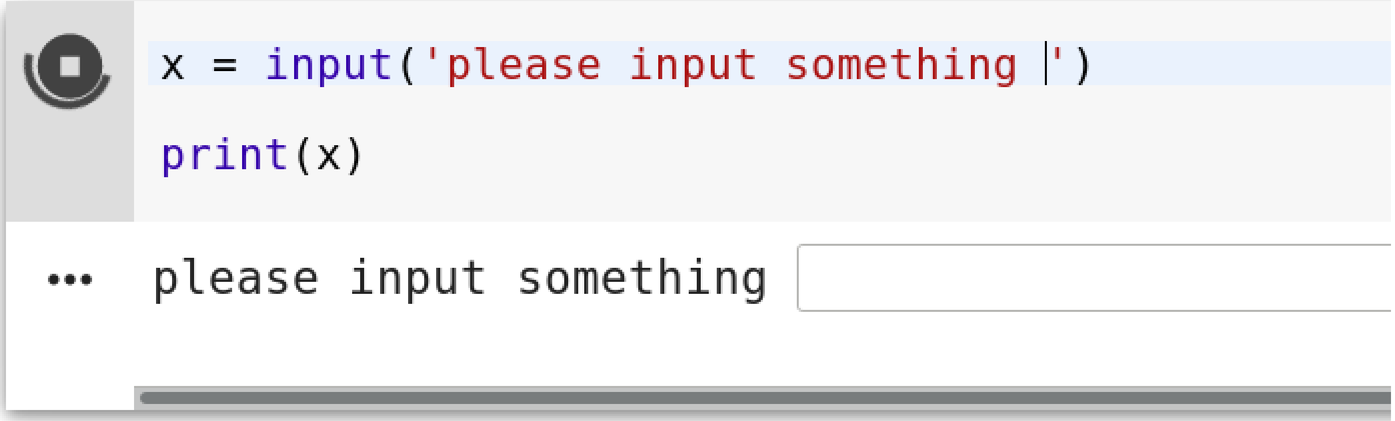


```
[1]  x = 2
     y = 5

     print(x+y)
```

```
7
```

Getting input from user

- You can get input from user in the same way you did with IDLE



```
x = input('please input something ')
print(x)
```

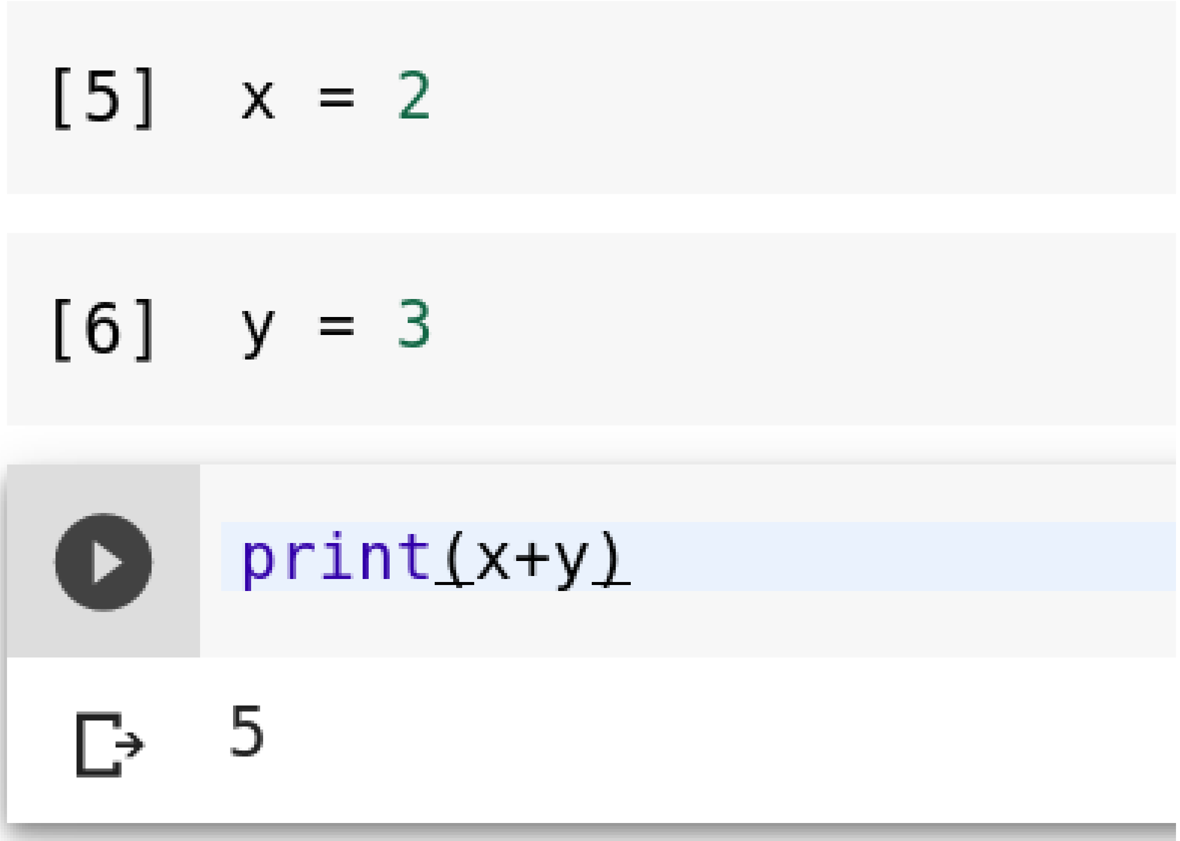
```
please input something
```

- It waits for an input and will continue after you pressed ENTER

Notes

- Once you’ve defined variables in a cell (and run it), the variables can be referenced in the subsequent cells.
- You can split your BIG code into multiple cells
- It will be easier to debug your code

Example



```
[5]  x = 2
```

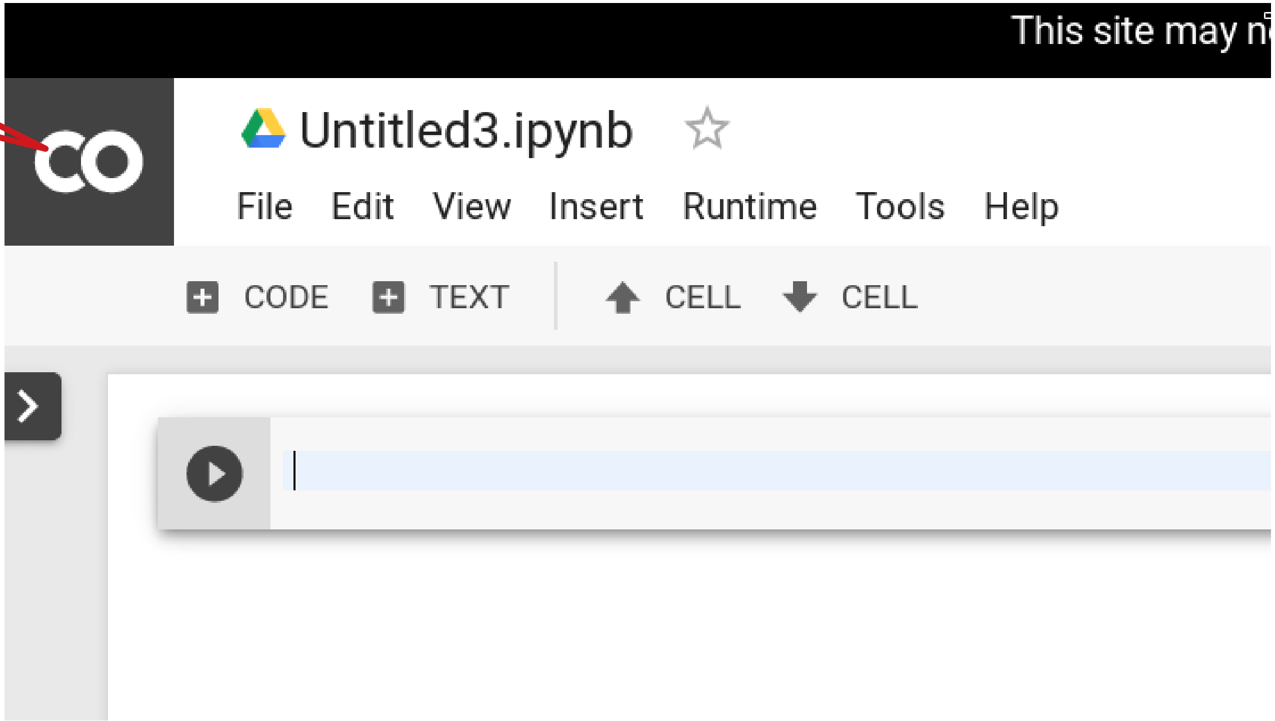
```
[6]  y = 3
```

```
print(x+y)
```

```
5
```

Colab and Google drive

- Files will be saved on your Google drive
- Click Colab icon to open the drive

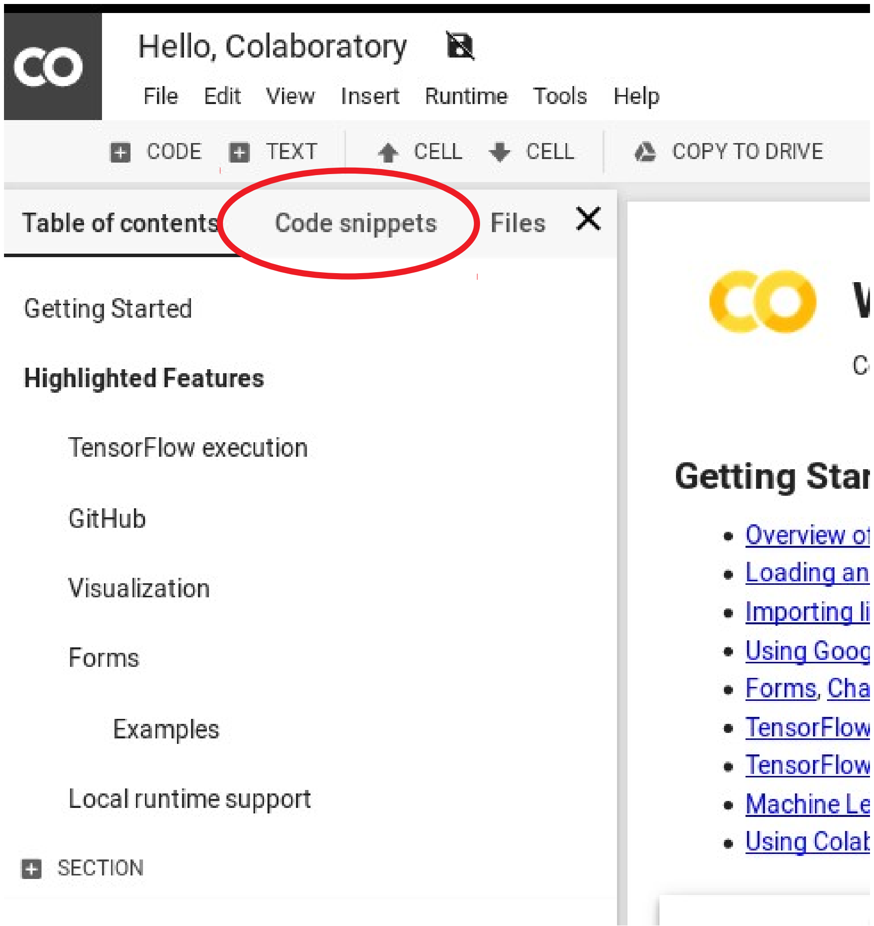


Go to Google Drive

Code snippets

- A snippet is a short code for doing some specific task
- Colab provides many useful snippets as examples
- They can be reused. (Problem solving using analogy and reduction)

Let’s try some snippet



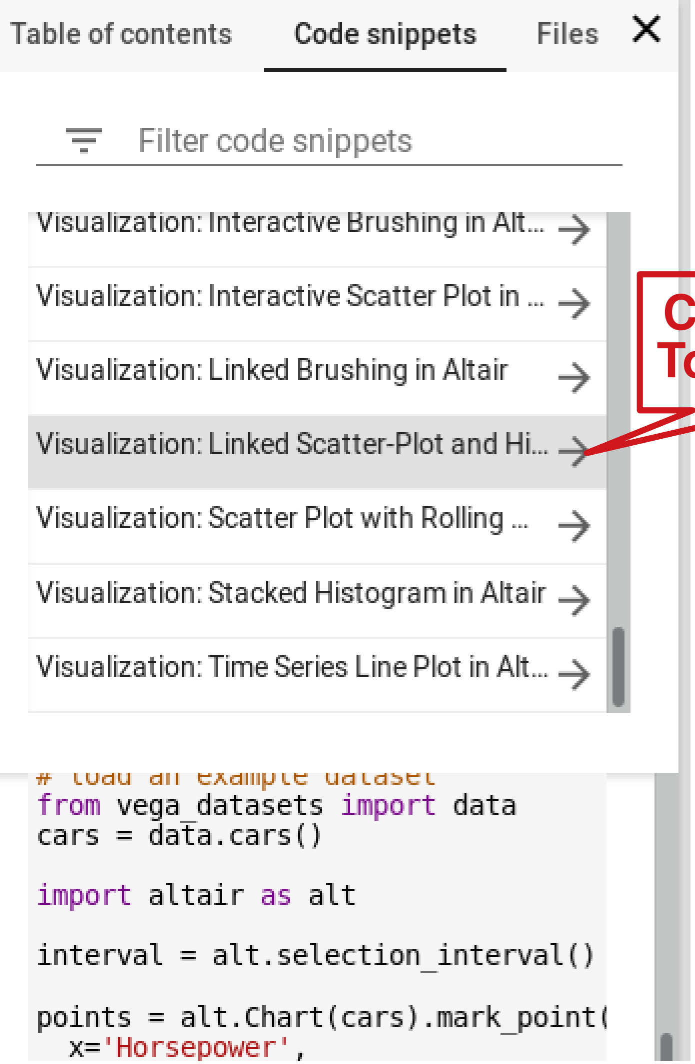


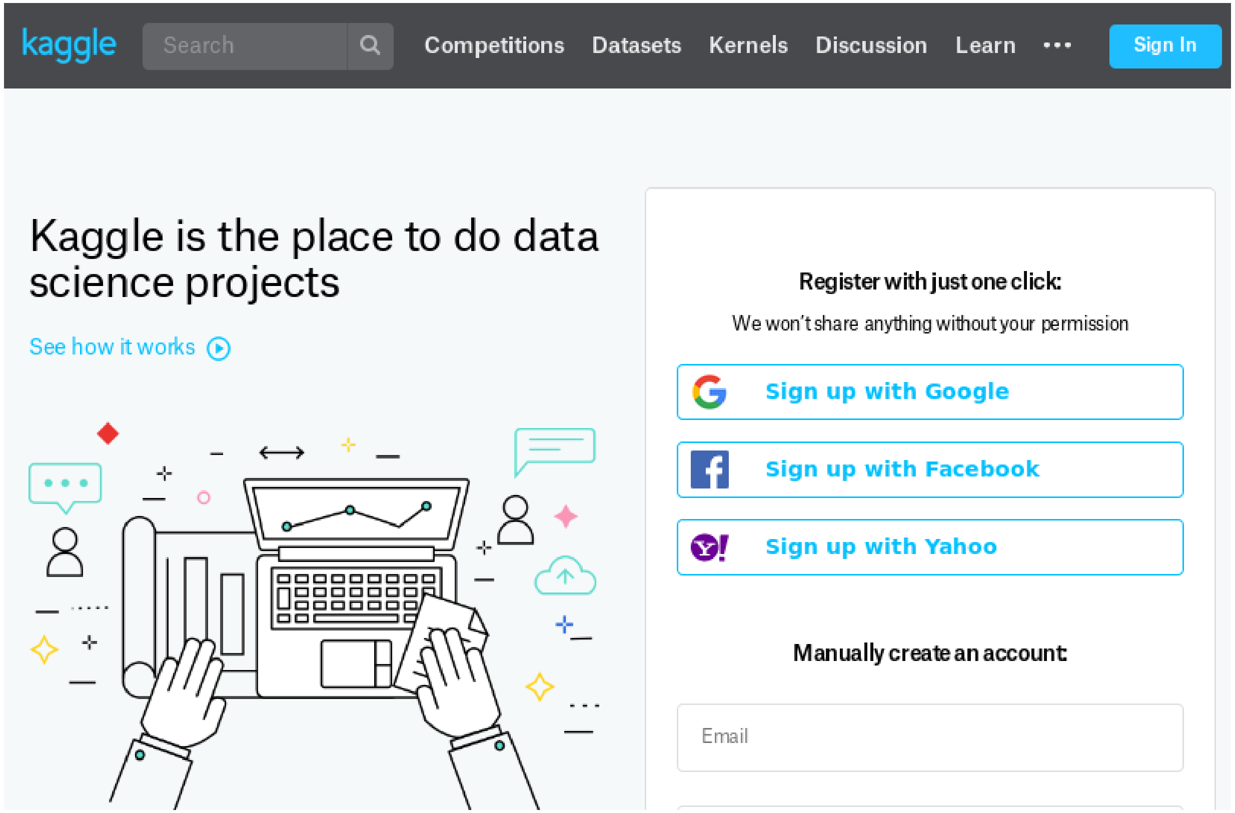
Table of contents
Code snippets
Files
Filter code snippets
Visualization: Interactive Brushing in Alt...
Visualization: Interactive Scatter Plot in ...
Visualization: Linked Brushing in Altair
Visualization: Linked Scatter-Plot and Hi...
Visualization: Scatter Plot with Rolling ...
Visualization: Stacked Histogram in Altair
Visualization: Time Series Line Plot in Alt...

```
# load an example dataset
from vega_datasets import data
cars = data.cars()

import altair as alt

interval = alt.selection_interval()

points = alt.Chart(cars).mark_point(
  x='Horsepower',
```

Click to add
To notebook

Useful websites [1]

- Data science competition / learning hub
- https://www.kaggle.com/



Kaggle is the place to do data science projects

See how it works

Register with just one click:

We won't share anything without your permission

Sign up with Google

Sign up with Facebook

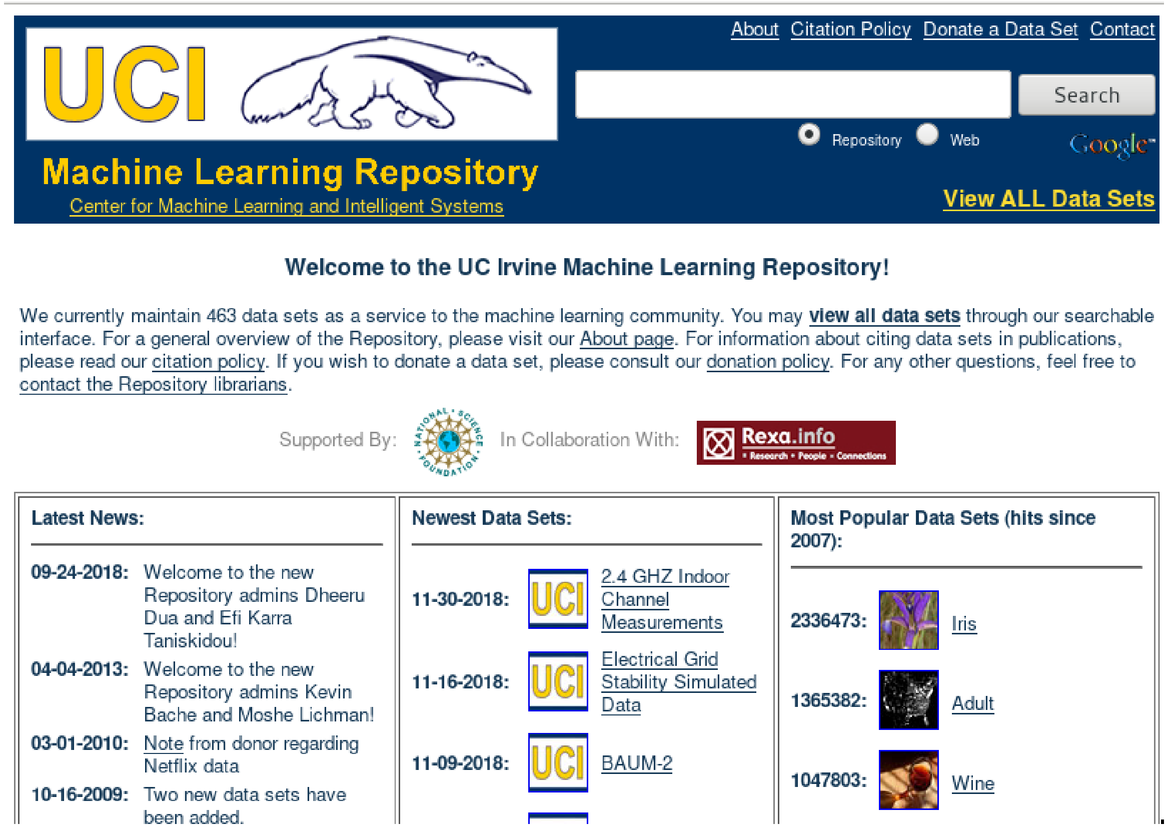
Sign up with Yahoo

Manually create an account:

Email

Useful websites [2]

- Data sharing webset
- https://archive.ics.uci.edu/ml/index.php



Welcome to the UC Irvine Machine Learning Repository!

We currently maintain 463 data sets as a service to the machine learning community. You may **view all data sets** through our searchable interface. For a general overview of the Repository, please visit our About page. For information about citing data sets in publications, please read our citation policy. If you wish to donate a data set, please consult our donation policy. For any other questions, feel free to contact the Repository librarians.

Supported By: In Collaboration With:

Latest News:

09-24-2018: Welcome to the new Repository admins Dheeru Dua and Efi Karra Taniskidou!

04-04-2013: Welcome to the new Repository admins Kevin Bache and Moshe Lichman!

03-01-2010: Note from donor regarding Netflix data

10-16-2009: Two new data sets have been added.

Newest Data Sets:

11-30-2018: 2.4 GHZ Indoor Channel Measurements

11-16-2018: Electrical Grid Stability Simulated Data

11-09-2018: BAUM-2

Most Popular Data Sets (hits since 2007):

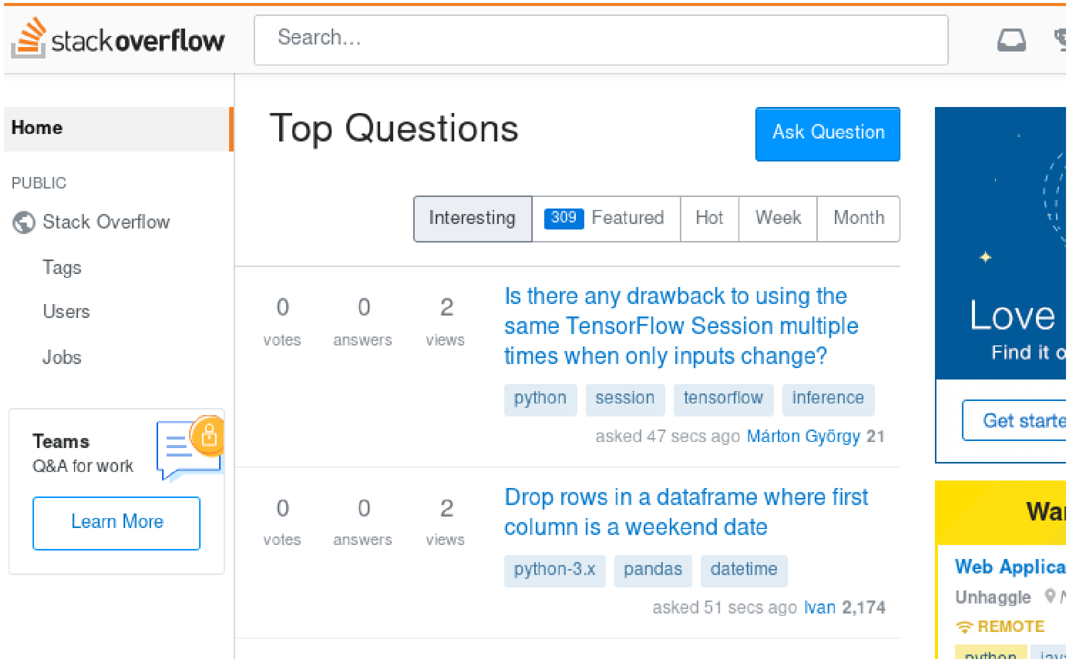
2336473: Iris

1365382: Adult

1047803: Wine

Useful Website [3]

- For asking programming related questions
- Or looking for solutions to problems similar to yours
- https://stackoverflow.com/



0 votes 0 answers 2 views

Is there any drawback to using the same TensorFlow Session multiple times when only inputs change?

python session tensorflow inference

asked 47 secs ago Márton György 21

0 votes 0 answers 2 views

Drop rows in a dataframe where first column is a weekend date

python-3.x pandas datetime

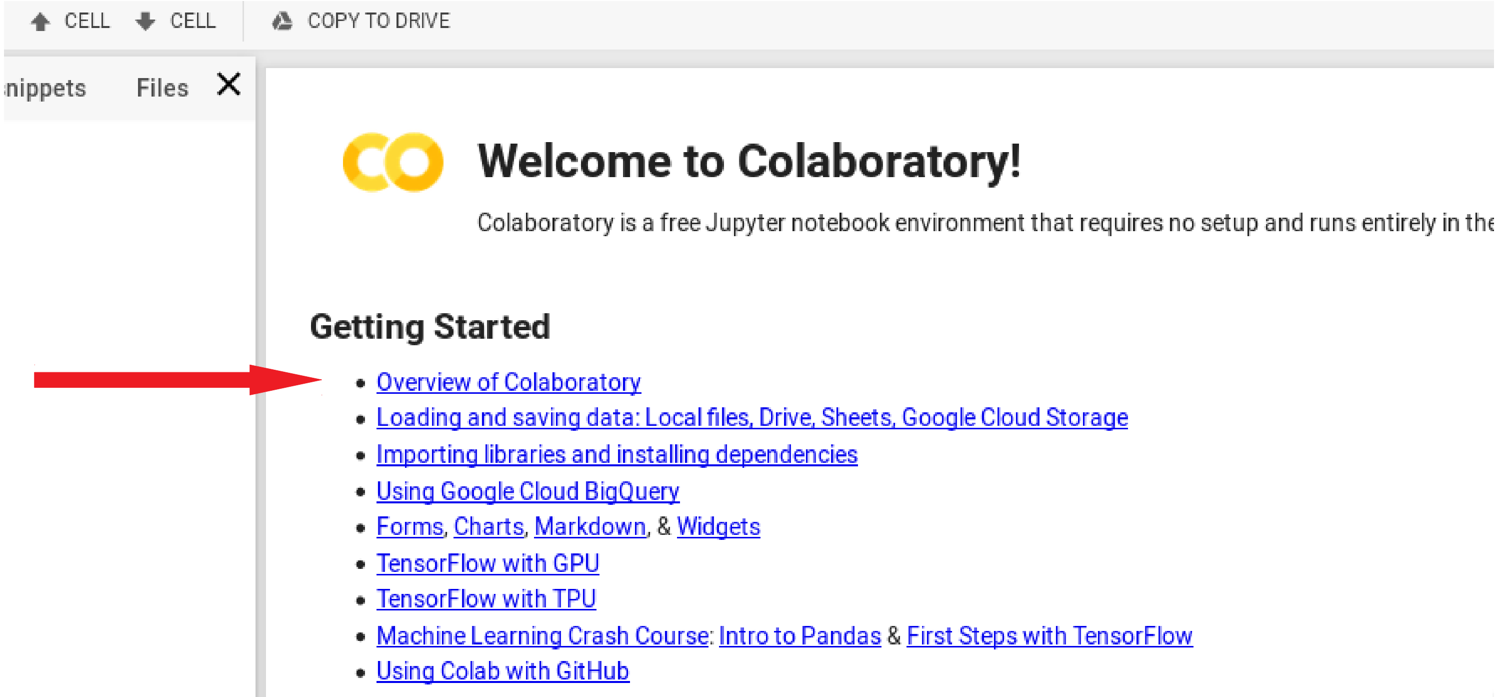
asked 51 secs ago Ivan 2,174

The End

- Thank you and have fun !!

Homework 0

- Read Overview of Colaboration



Welcome to Colaboratory!

Colaboratory is a free Jupyter notebook environment that requires no setup and runs entirely in the

Getting Started

- Overview of Colaboratory
- Loading and saving data: Local files, Drive, Sheets, Google Cloud Storage
- Importing libraries and installing dependencies
- Using Google Cloud BigQuery
- Forms, Charts, Markdown, & Widgets
- TensorFlow with GPU
- TensorFlow with TPU
- Machine Learning Crash Course: Intro to Pandas & First Steps with TensorFlow
- Using Colab with GitHub