



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

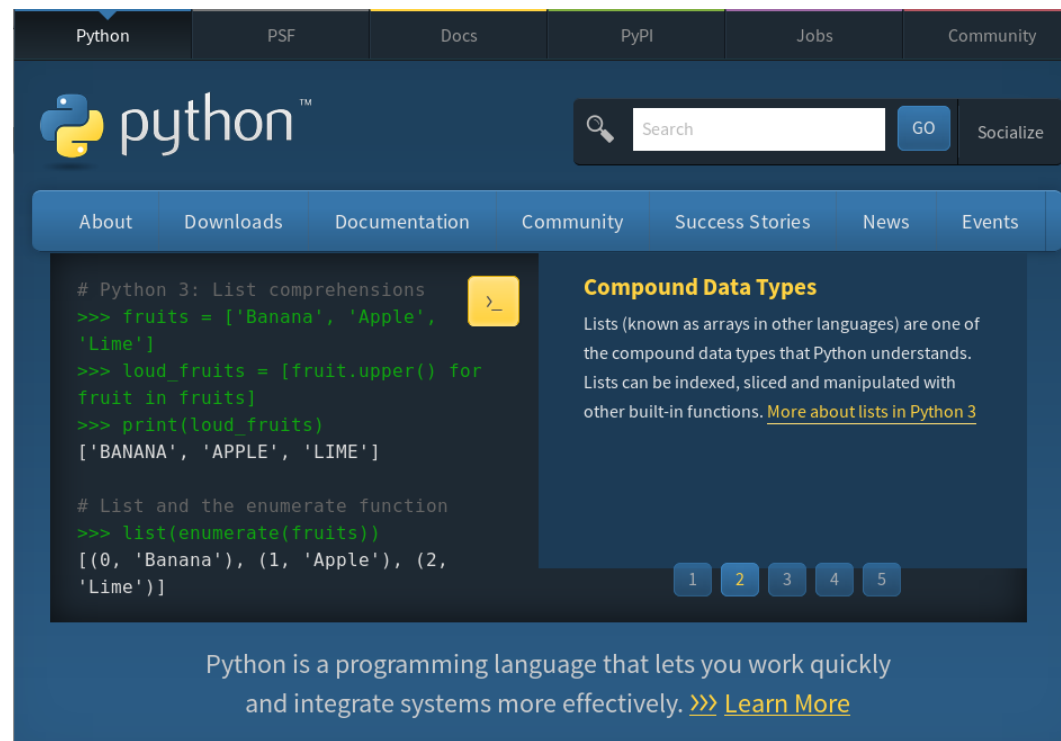


Python distribution

- 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



The screenshot shows the Python.org website. At the top, there are navigation links for Python, PSF, Docs, PyPI, Jobs, and Community. Below this is the Python logo and a search bar. A secondary navigation bar includes links for About, Downloads, Documentation, Community, Success Stories, News, and Events. The main content area features a code editor with Python 3 code examples for list comprehensions and the enumerate function. To the right of the code is a section titled "Compound Data Types" with text explaining lists and a link to "More about lists in Python 3". At the bottom of the code editor are page numbers 1 through 5. Below the code editor is a blue banner with the text: "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 experienced developer, Python is a language that's easy to learn and use.

Download

Python source code and installers are available for Windows, macOS, Linux, and BSD.

Docs

Documentation for Python's standard library, third-party libraries, and the Python interpreter.

Jobs

Looking for work or have a Python related position that needs to be filled? Visit the Python Jobs page.

Anaconda distribution

- Standard distribution + approx. 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



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

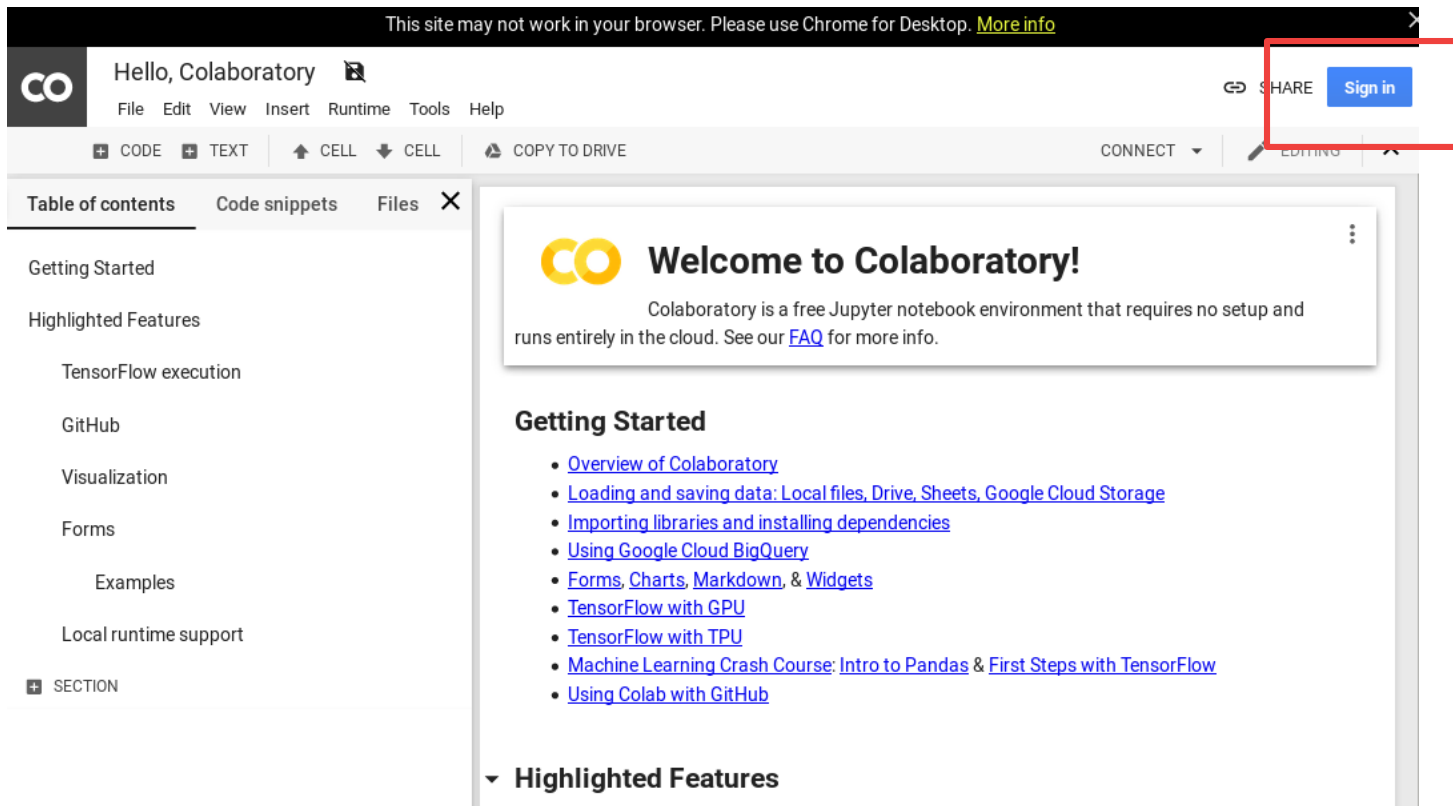


Cloud computing service

- 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













The screenshot shows the Google Colaboratory website interface. At the top, there is a warning banner: "This site may not work in your browser. Please use Chrome for Desktop. [More info](#)". Below this is the Colaboratory logo and the text "Hello, Colaboratory". A navigation menu includes "File", "Edit", "View", "Insert", "Runtime", "Tools", and "Help". On the right side, there is a "SHARE" button and a "Sign in" button, which is highlighted with a red box. Below the navigation menu, there are buttons for "+ CODE", "+ TEXT", "↑ CELL", "↓ CELL", and "COPY TO DRIVE". A "CONNECT" dropdown menu is also visible. On the left side, there is a "Table of contents" section with links to "Getting Started", "Highlighted Features", "TensorFlow execution", "GitHub", "Visualization", "Forms", "Examples", and "Local runtime support". The main content area features a "Welcome to Colaboratory!" message, stating that Colaboratory is a free Jupyter notebook environment that runs entirely in the cloud. Below this, there is a "Getting Started" section with a list of links: "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", and "Using Colab with GitHub". At the bottom, there is a "Highlighted Features" section.

After signing in

This site may not work in your browser. Please use Chrome for Desktop. [More info](#)

EXAMPLES RECENT GOOGLE DRIVE GITHUB UPLOAD

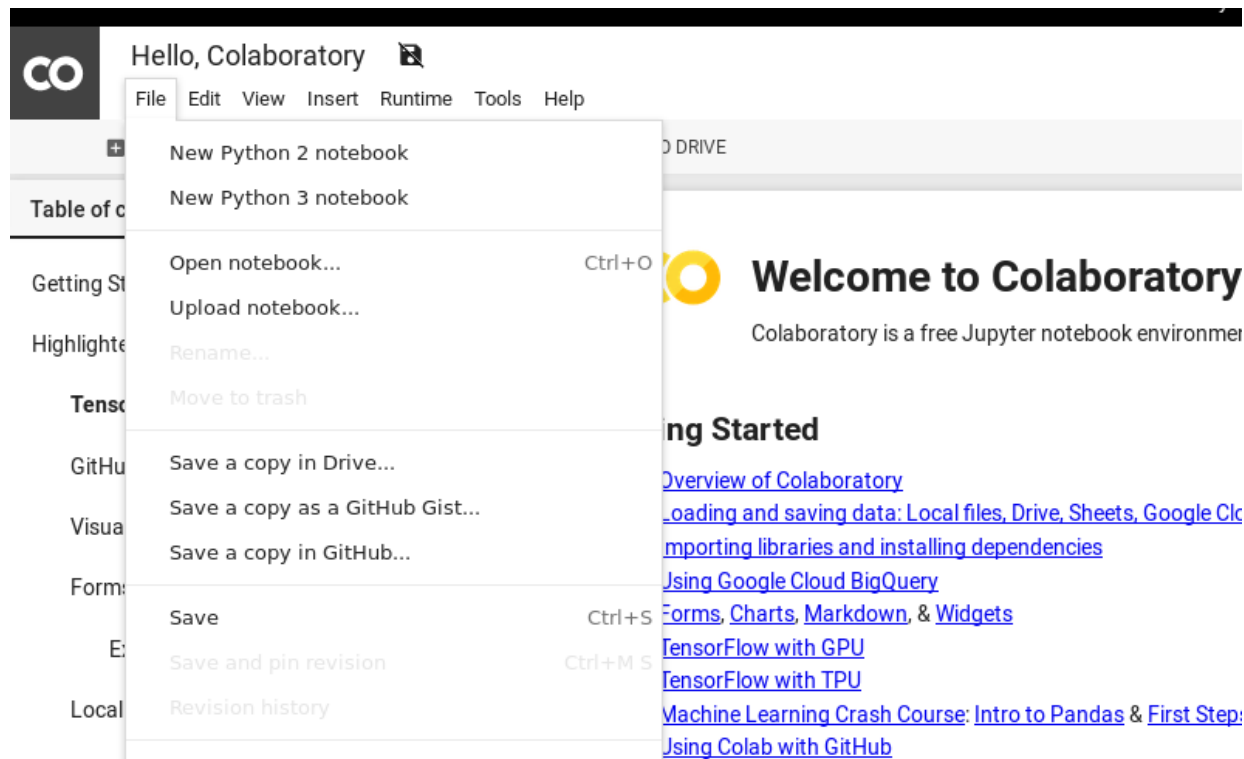
Filter notebooks

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 NEW PYTHON 3 NOTEBOOK CANCEL

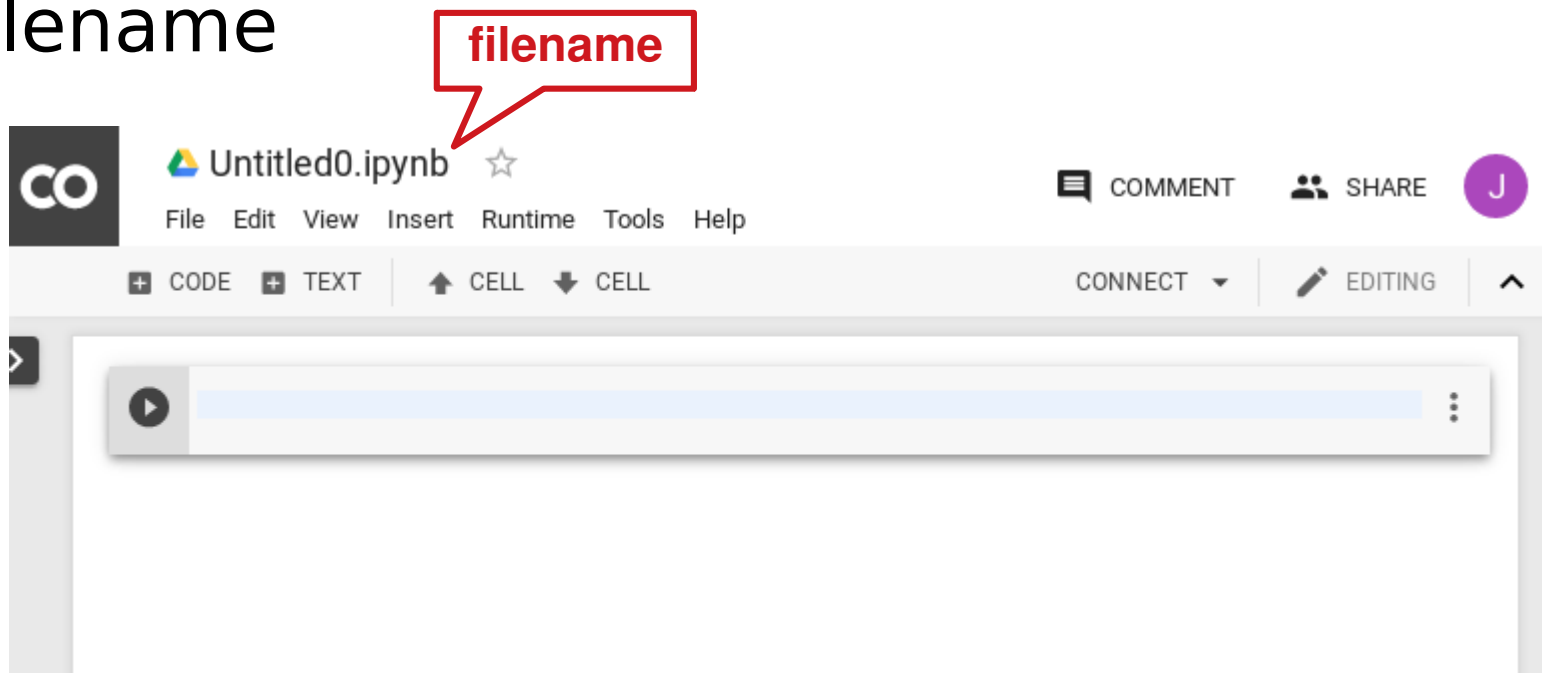
Alternatively

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

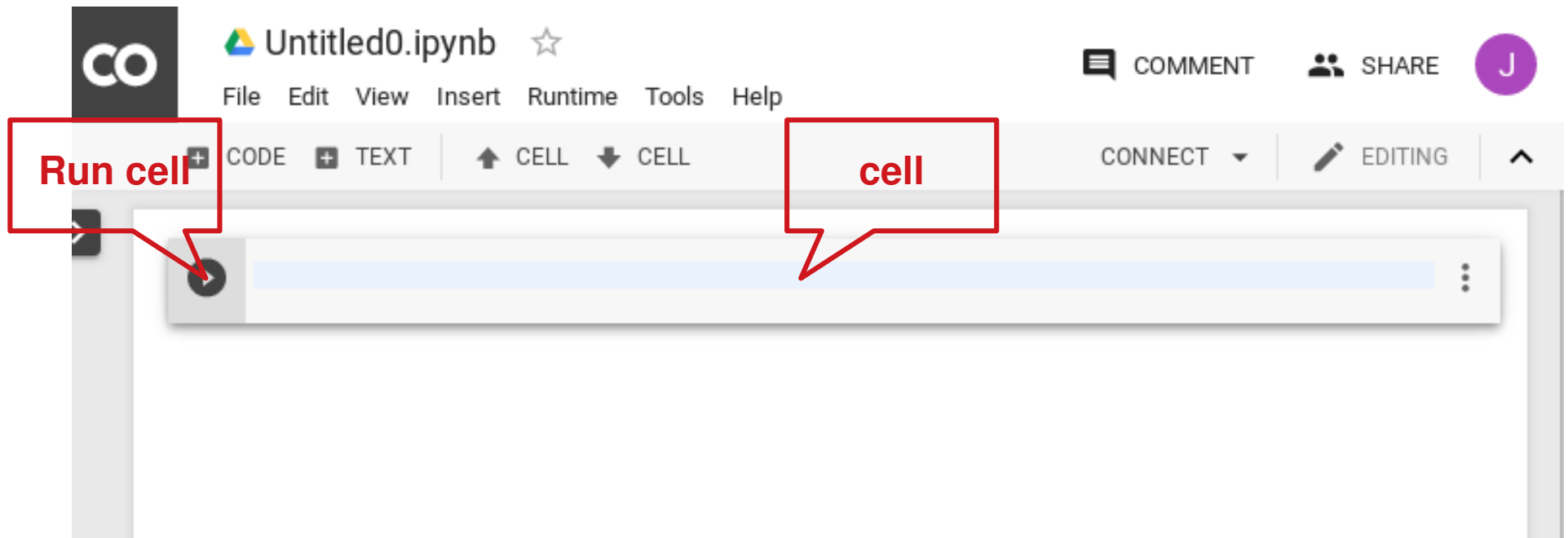


Your new Python file

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



What's in a notebook ?

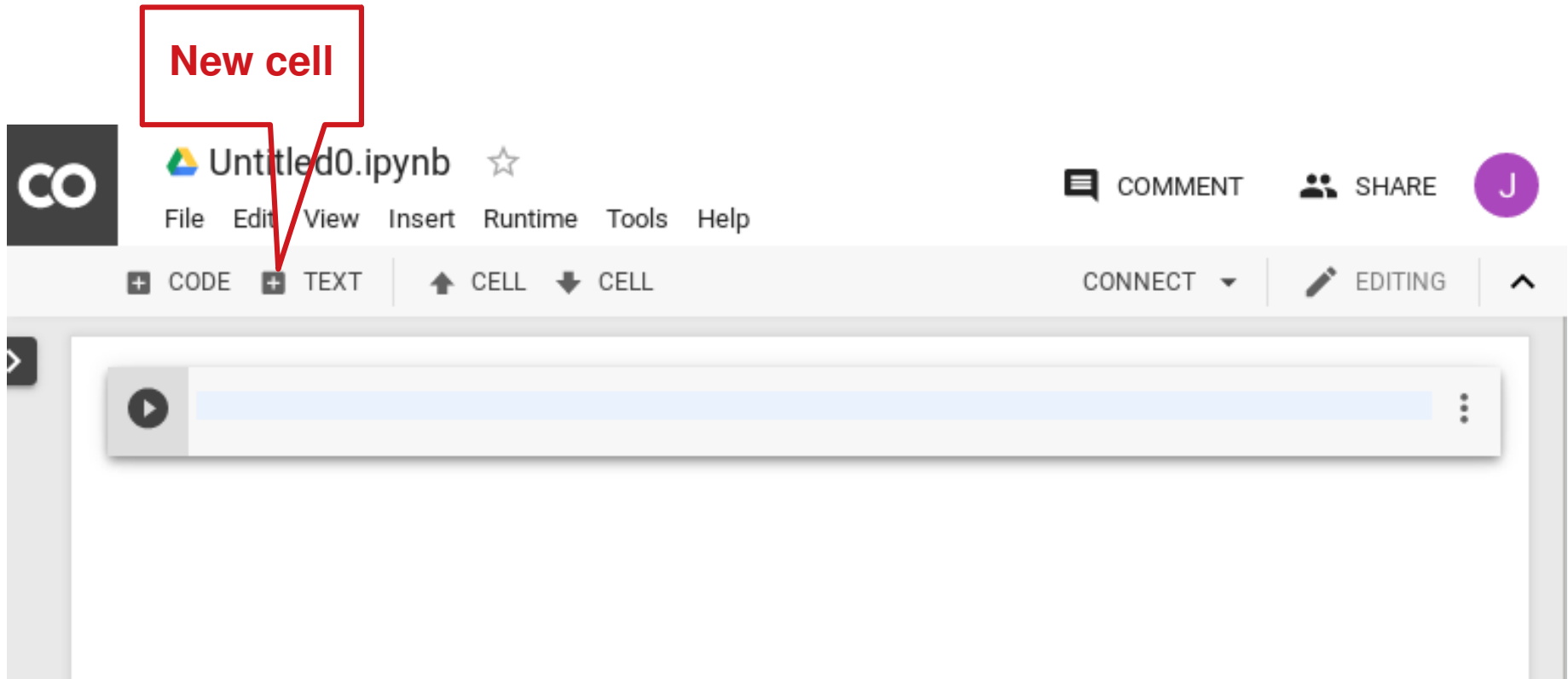


Adding new code cell

New cell

The screenshot displays the Jupyter Notebook interface. At the top left, there is a 'CO' logo. The main title is 'Untitled0.ipynb' with a star icon. Below the title is a menu bar with options: File, Edit, View, Insert, Runtime, Tools, and Help. On the right side, there are buttons for 'COMMENT', 'SHARE', and a user profile icon with the letter 'J'. Below the menu bar is a toolbar with buttons for '+ CODE', '+ TEXT', '↑ CELL', and '↓ CELL'. The '↑ CELL' button is highlighted, indicating the action of adding a new code cell. To the right of the toolbar are 'CONNECT' and 'EDITING' buttons. The main area shows a single code cell with a play button on the left and a three-dot menu on the right.

Adding new text cell



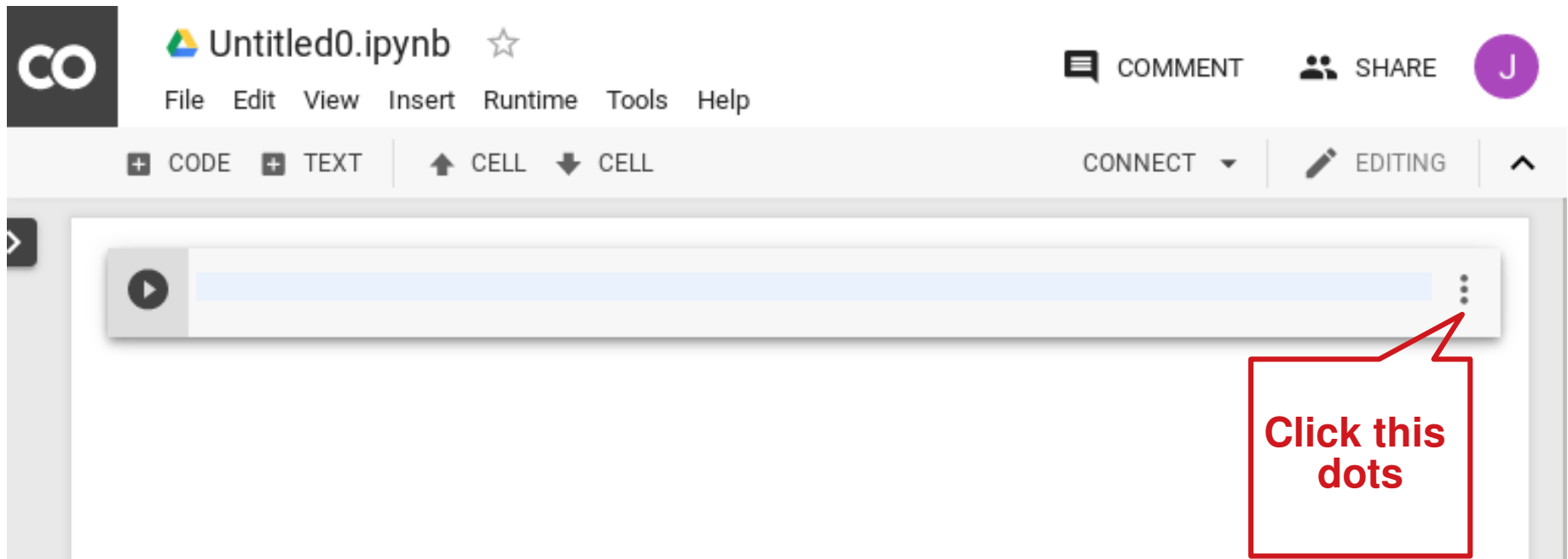
Text cell is useful for adding context to the code



Exercise 1

- Add one code cell and one text cell

Deleting cell



The image shows a Jupyter Notebook interface. At the top left is the 'CO' logo. The main title is 'Untitled0.ipynb' with a star icon. Below the title is a menu bar with 'File', 'Edit', 'View', 'Insert', 'Runtime', 'Tools', and 'Help'. On the right side of the top bar are 'COMMENT' and 'SHARE' buttons, and a purple circular profile icon with the letter 'J'. Below the menu bar is a toolbar with '+ CODE', '+ TEXT', '↑ CELL', and '↓ CELL' buttons. To the right of these are 'CONNECT' and 'EDITING' buttons. The main area shows a single code cell with a light blue background. On the left side of the cell is a play button icon. On the right side of the cell is a three-dot menu icon. A red callout box with a red border and a red arrow pointing to the three-dot menu icon contains the text '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

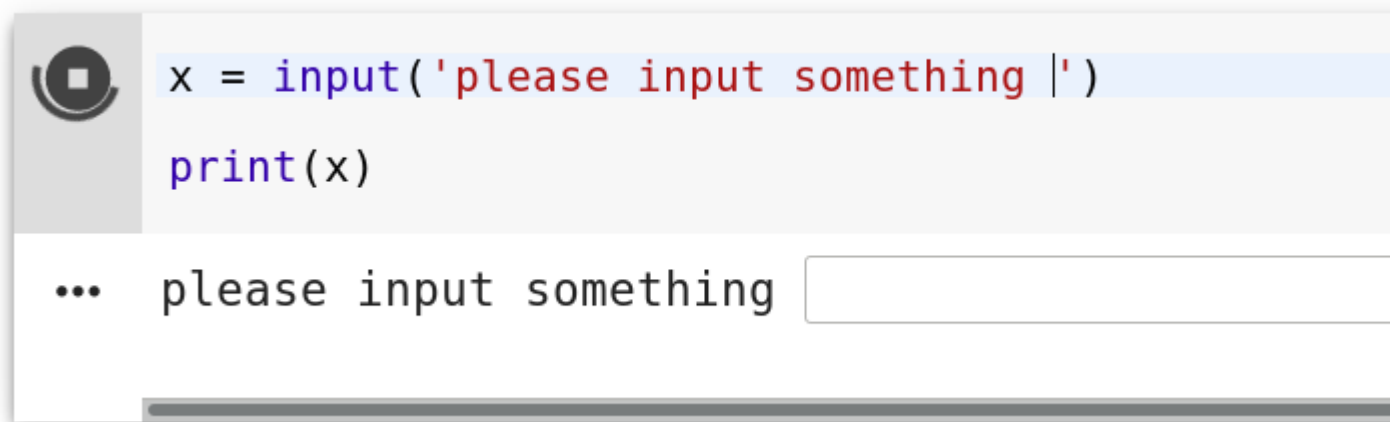
```
[1] x = 2  
    y = 5  
  
    print(x+y)
```

↪ 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

The screenshot shows a Python IDLE shell window. The top part contains two lines of Python code: `x = input('please input something |')` and `print(x)`. The first line is highlighted in blue. Below the code, the shell shows the prompt `... please input something` followed by an empty text input field, indicating that the program is waiting for user input.

- 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
```

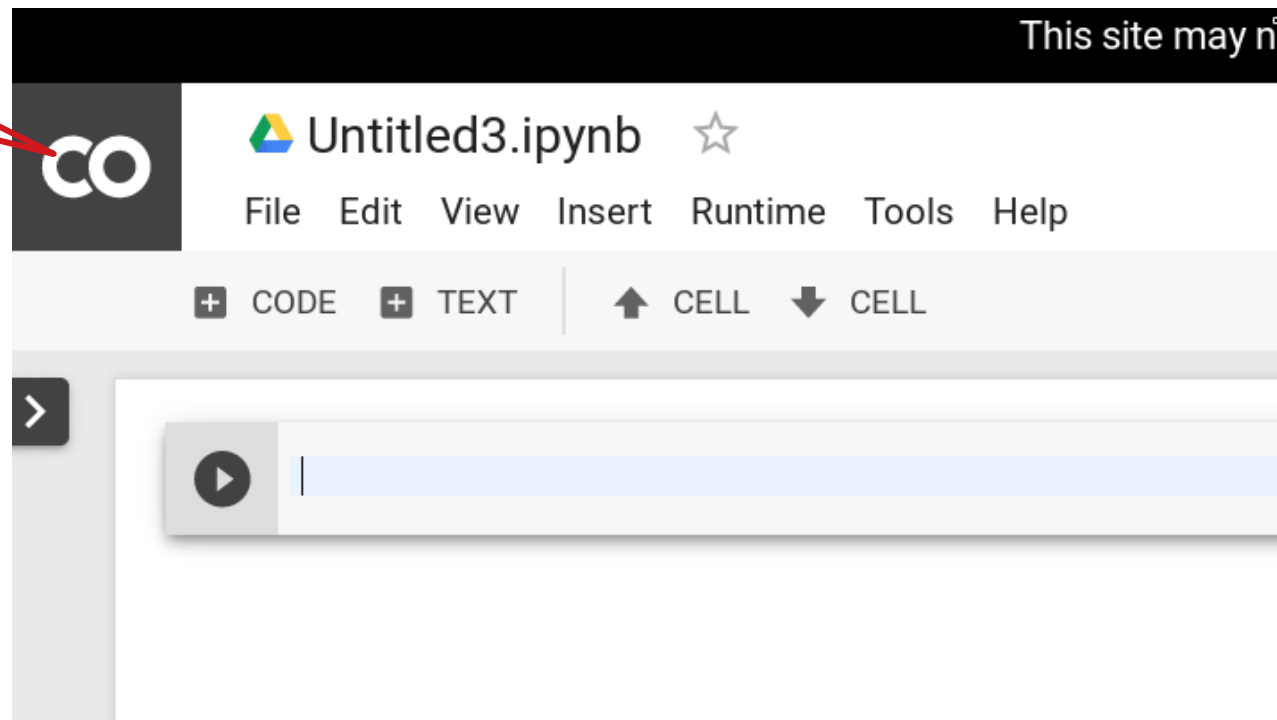
```
▶ print(x+y)
```

```
☞ 5
```

Colab and Google drive

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

Go to
Google Drive

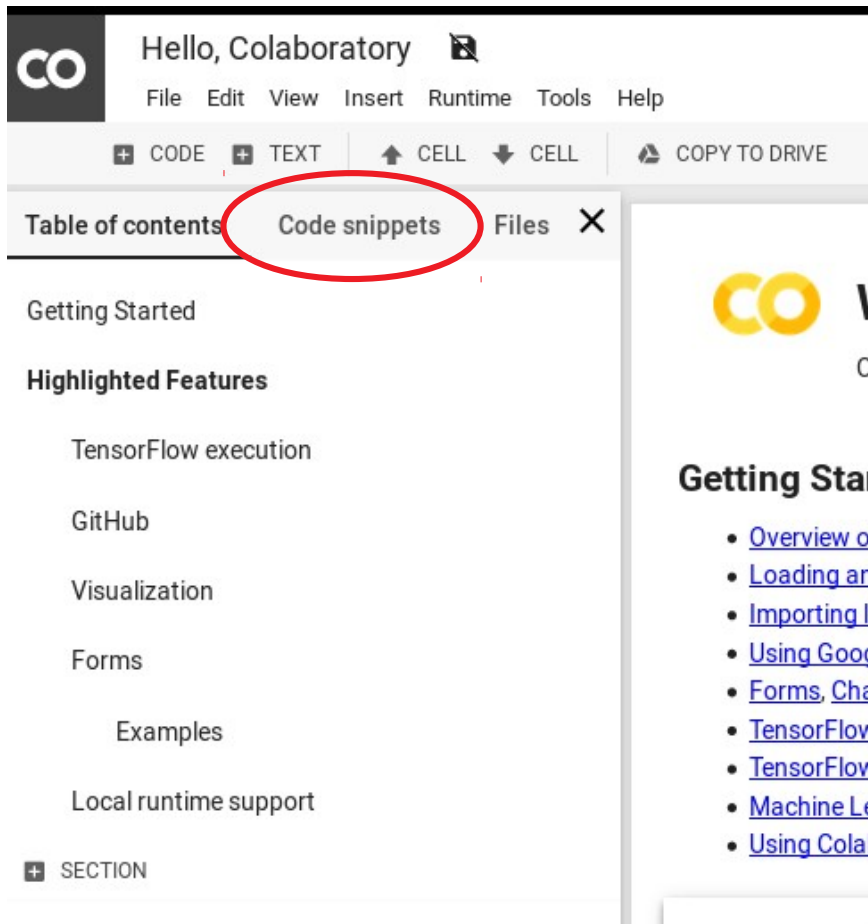




Code snippets

- 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



CO Hello, Colaboratory

File Edit View Insert Runtime Tools Help

+ CODE + TEXT ↑ CELL ↓ CELL COPY TO DRIVE

Table of contents **Code snippets** Files X

Getting Started

Highlighted Features

- TensorFlow execution
- GitHub
- Visualization
- Forms
- Examples
- Local runtime support

SECTION

CO

Getting Star

- [Overview of](#)
- [Loading an](#)
- [Importing li](#)
- [Using Goo](#)
- [Forms, Cha](#)
- [TensorFlow](#)
- [TensorFlow](#)
- [Machine Le](#)
- [Using Colat](#)

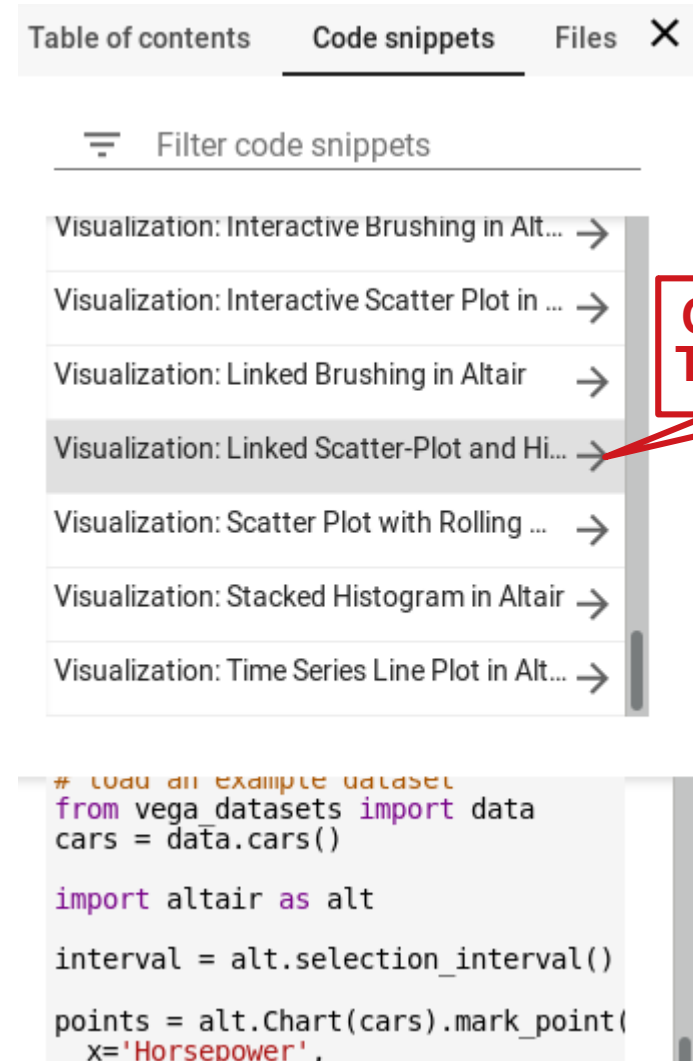


Table of contents **Code snippets** Files X

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

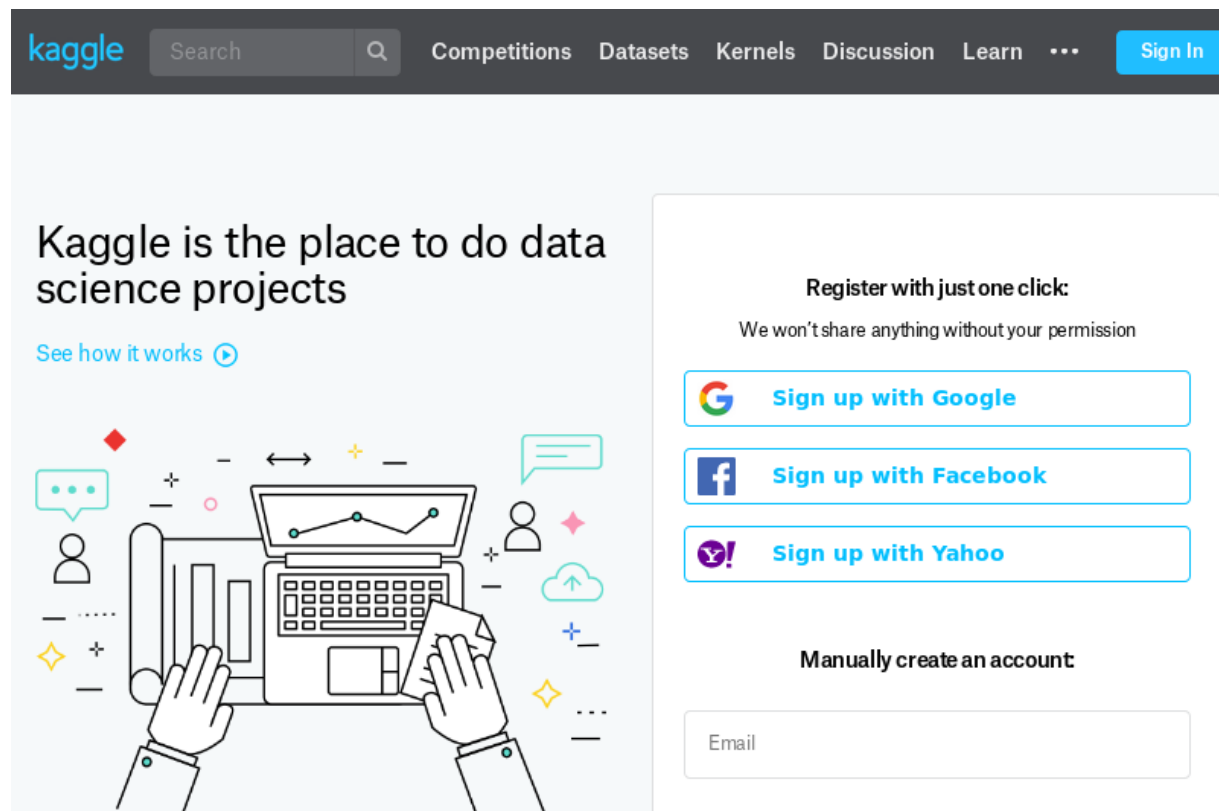
interval = alt.selection_interval()

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/>



The image shows a screenshot of the Kaggle website's homepage. At the top, there is a dark navigation bar with the Kaggle logo on the left, a search bar, and links for 'Competitions', 'Datasets', 'Kernels', 'Discussion', and 'Learn'. A 'Sign In' button is on the right. Below the navigation bar, the main content area features the text 'Kaggle is the place to do data science projects' and a link 'See how it works'. To the right, there is a registration section titled 'Register with just one click' with the subtext 'We won't share anything without your permission'. This section contains three buttons for 'Sign up with Google', 'Sign up with Facebook', and 'Sign up with Yahoo!'. Below these is a section for 'Manually create an account' with an 'Email' input field. The background of the main content area has a light blue and white illustration of a person's hands typing on a laptop, surrounded by various data science icons like a line graph, a bar chart, a person icon, and mathematical symbols.

Useful websites [2]

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



The screenshot shows the header of the UCI Machine Learning Repository website. It features the UCI logo (University of California, Irvine) and a drawing of a kangaroo. The text reads "Machine Learning Repository" and "Center for Machine Learning and Intelligent Systems". There are navigation links for "About", "Citation Policy", "Donate a Data Set", and "Contact". A search bar is present with a "Search" button. Below the search bar, there are radio buttons for "Repository" and "Web", and a "Google" logo. A link "View ALL Data Sets" is also visible.







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.



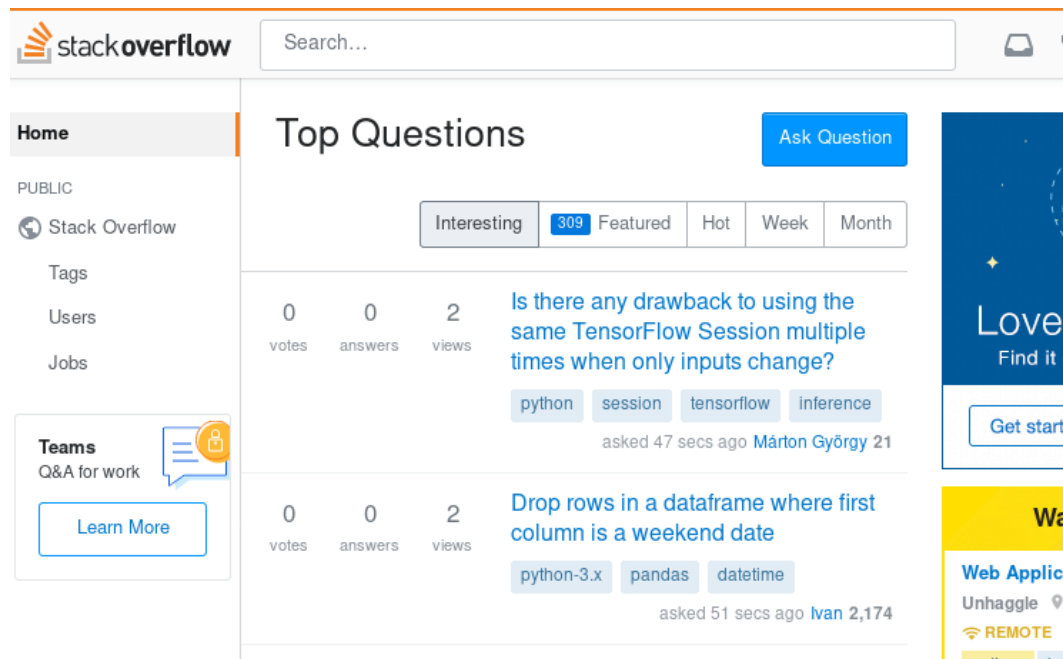
In Collaboration With:



Latest News:	Newest Data Sets:	Most Popular Data Sets (hits since 2007):
<p>09-24-2018: Welcome to the new Repository admins Dheeru Dua and Efi Karra Taniskidou!</p> <p>04-04-2013: Welcome to the new Repository admins Kevin Bache and Moshe Lichman!</p> <p>03-01-2010: Note from donor regarding Netflix data</p> <p>10-16-2009: Two new data sets have been added.</p>	<p>11-30-2018:  2.4 GHZ Indoor Channel Measurements</p> <p>11-16-2018:  Electrical Grid Stability Simulated Data</p> <p>11-09-2018:  BAUM-2</p>	<p>2336473:  Iris</p> <p>1365382:  Adult</p> <p>1047803:  Wine</p>

Useful Website [3]

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



The screenshot displays the Stack Overflow homepage. At the top, there is a search bar and a navigation menu. The main content area is titled 'Top Questions' and features a list of questions. The first question is 'Is there any drawback to using the same TensorFlow Session multiple times when only inputs change?' with 0 votes, 0 answers, and 2 views. The second question is 'Drop rows in a dataframe where first column is a weekend date' with 0 votes, 0 answers, and 2 views. The left sidebar contains navigation links for Home, PUBLIC, Stack Overflow, Tags, Users, Jobs, and Teams. The right sidebar has a 'Love' section and a 'Web Applica' section.

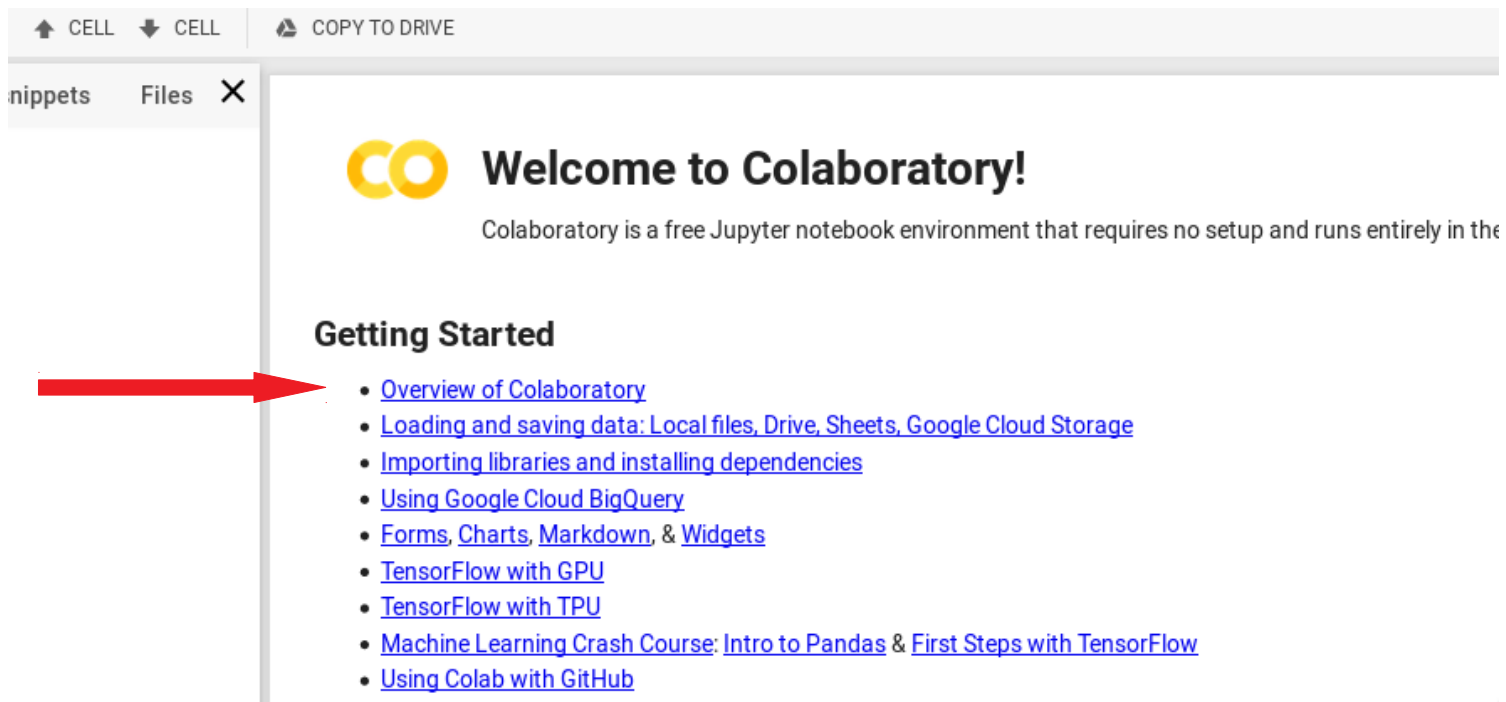


The End

- Thank you and have fun !!

Homework 0

- Read Overview of Colaboration



The screenshot shows the Colaboratory web interface. At the top, there are navigation buttons: 'CELL' (up arrow), 'CELL' (down arrow), and 'COPY TO DRIVE'. Below this is a sidebar with 'nippets' and 'Files' tabs. The main content area features the Colaboratory logo (two overlapping yellow circles) and the text 'Welcome to Colaboratory!'. Below this, it states 'Colaboratory is a free Jupyter notebook environment that requires no setup and runs entirely in the'. A section titled 'Getting Started' contains a list of links. A red arrow points to the first link, 'Overview of Colaboratory'.

↑ CELL ↓ CELL COPY TO DRIVE

nippets Files X

CO 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](#)