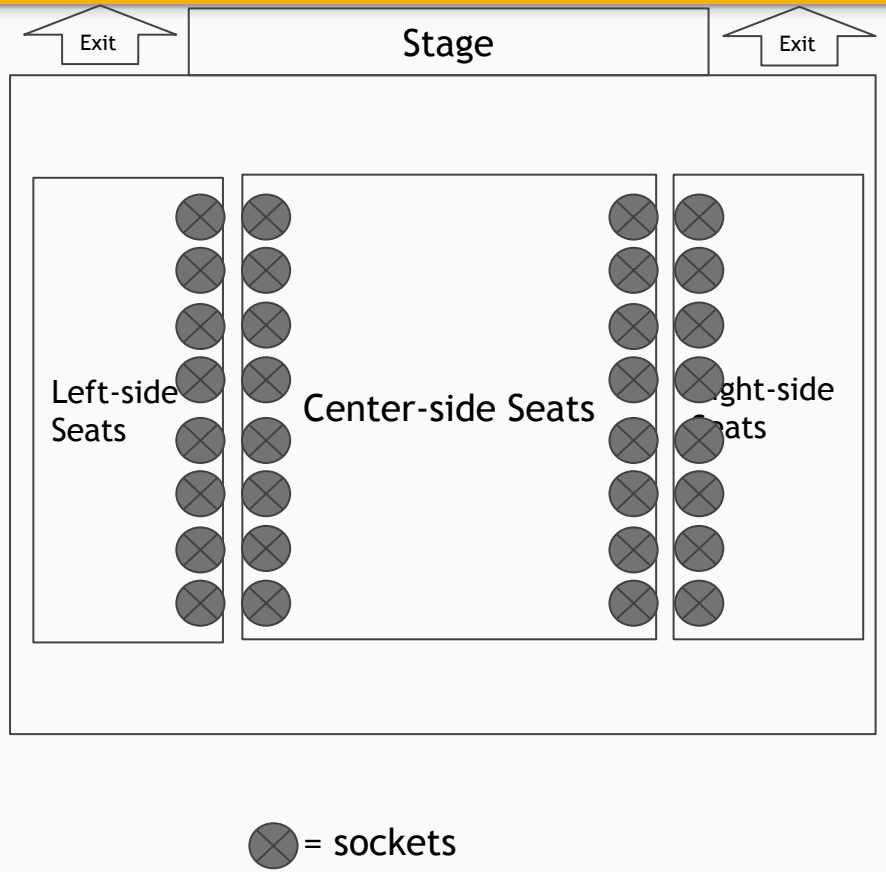
Welcome Survey

https://app.sli.do/event/irXJiigN6piRy485RmyN9x

Yasumoto International Academic Park - YIA LT6

Limited Mobile Signal. Please use on-campus wifi.
32 Sockets. But please bring your own charger.



Python Basics - Programming 101

Variables, Controls, Loops

CUHK MSc Data Science & Biz Stat. Program
STAT5106 - Programming Techniques for Data Science
Week 1 @ ~~5~~ 12 Sept 2024

Self-Introduction

- 2008 July
  BSc in Statistics CUHK (1st Hons)
  in Data Science and Bus. Stat. Stream
- 2010 July
  MPhil in Statistics CUHK
  under Supervised by Prof. Ben Chan
- 2010 Oct – 2013 Sept
  Algo Trading - CASH Dynamic Opportunities Investment
- 2013 Sept – 2014 Sept
  IT Consulting - Long Term Intelligence
- 2014 Sept – 2015 Sept
  Government - Environmental Protection Dept
- 2016 Mar – 2018 Dec
  Energy - CLP Power Limited
- 2019 Feb – 2021 Sept
  Banking - Hang Seng Bank Limited
- 2021 Oct - 2024 Oct
  Maritime - Fleet Management Limited
- 2024 Oct - ...?
  To Be Announced...

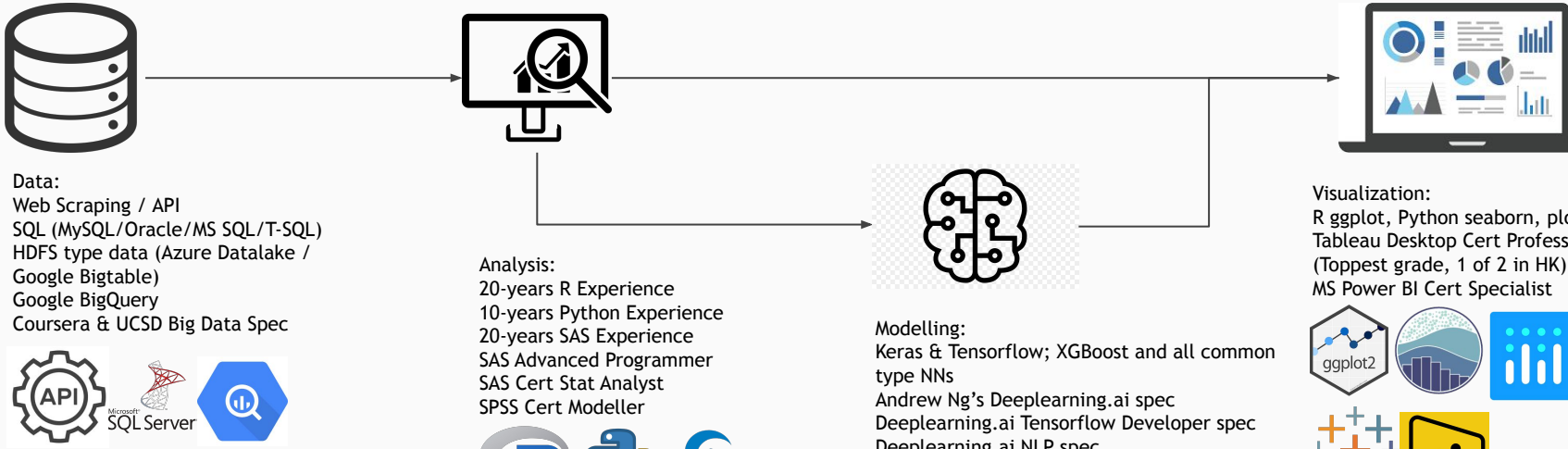
www.linkedin.com/in/kyalanlo/

kyalan.2013@gmail.com
kwokyuenlo@cuhk.edu.hk

Ability

I am a full-stack data scientist - can be end-to-end hands on, and act as team leader.

Cloud:



Data:
Web Scraping / API
SQL (MySQL/Oracle/MS SQL/T-SQL)
HDFS type data (Azure Datalake / Google Bigtable)
Google BigQuery
Coursera & UCSD Big Data Spec

Analysis:
20-years R Experience
10-years Python Experience
20-years SAS Experience
SAS Advanced Programmer
SAS Cert Stat Analyst
SPSS Cert Modeller

Modelling:
Keras & Tensorflow; XGBoost and all common type NNs
Andrew Ng's Deeplearning.ai spec
Deeplearning.ai Tensorflow Developer spec
Deeplearning.ai NLP spec
Michigan U Recommender System Spec

Visualization:
R ggplot, Python seaborn, plotly
Tableau Desktop Cert Professional (Toppest grade, 1 of 2 in HK)
MS Power BI Cert Specialist

The Tutors

Muhammad Danish

BSc in Data Sciences (CityU), currently pursuing Ph.D. in Statistics (CUHK)

mdanish@link.cuhk.edu.hk

Lady Shaw Building 130, 3943 7939

https://sites.google.com/view/muhdanish/home

Catherine Yu

currently pursuing Ph.D. in Statistics (CUHK)

cxryu@link.cuhk.edu.hk

Agenda

Before Mid-term = ***Programming***

After Mid-Term = ***DS Application***

Date	
Week 1 - ~~5 Sept~~ 12 Sept	
Week 2 - ~~12 Sept~~	
Week 3 - ~~19 Sept~~	
Week 4 - ~~26 Oct~~	
Week 5 - ~~3 Oct~~	
Week 6 - ~~10 Oct~~	
~~17 Oct~~	Mid-term Take-Home Exam (No class)

	ery, Sort, Merge,
	ion in Pandas &
	eaborn
	adline**
	eadline**
~~Week 12 - 28 Nov~~	ChatGPT with Large Language Model Applications
~~4 Dec~~	**Final Project deadline**
~~7 Dec~~ (Saturday)	**Final Project Presentation Day**

OPTIONS

1. **Delete the missed Lesson**, and condense contents into Lessons "Week 2" - "Week 6"
2. **Reschedule the missed lesson** into any days in remaining 5 weeks
3. Shift the **full schedule forward 1 week**.
   i.e., the following will be deferred
   a. Assignment deadlines
   b. Mid-Term schedule
   c. Final Project schedule

Agenda

Before Mid-term = ***Programming***

Date	Topic
Week 1 - ~~5 Sept~~ 12 Sept	Python Basics, Programming 101 - Variables, Controls, Loops
Week 2 - ~~12 Sept~~	Programming 102: Exception Handling, Functions, String
Week 3 - ~~19 Sept~~	Python Data Structures - List, Dictionary, Tuple Regular Expressions **Assignment 1 (for weeks 1, 2) deadline**
Week 4 - ~~26 Oct~~	Introduction to Data Science Python Packages, Numpy, Pandas basics Python vs R in Programming
Week 5 - ~~3 Oct~~	Open Data: Web Scraping **Assignment 2 (for weeks 3, 4) deadline**
Week 6 - ~~10 Oct~~	APIs, with More Example on Open Data
~~17 Oct~~	Mid-term Take-Home Exam (No class)

After Mid-Term = ***DS Application***

Date	Topic
Week 7 - ~~24 Oct~~	Data Processing in Pandas - IO, Query, Sort, Merge, Missing Value Handling
Week 8 - ~~31 Oct~~	Exploratory Data Analysis, Simulation in Pandas & Numpy
Week 9 - ~~7 Nov~~	Data Visualization in Matplotlib, Seaborn Introduction to Tableau **Assignment 3 (for weeks 7, 8) deadline**
Week 10 - ~~14 Nov~~	Bayesian Computation in Python
Week 11 - ~~21 Nov~~	Word Data Preprocessing in Python Introduction to Machine Learning Open Model: Tensorflow-hub **Assignment 4 (for weeks 9, 10) deadline**
Week 12 - ~~28 Nov~~	ChatGPT with Large Language Model Applications
~~4 Dec~~	**Final Project deadline**
~~7 Dec~~ (Saturday)	**Final Project Presentation Day**

Reference

- Python
  - Dr. Charles Russell Severance (2016) - Python for Everybody Exploring Data in Python 3, CreateSpace
  - Python for Everybody Specialization - Coursera & Michigan University https://www.coursera.org/specializations/python
  - Jake VanderPlas (2016) - Python Data Science Handbook, O'Reilly
  - Applied Data Science with Python Specialization - Coursera & Michigan University https://www.coursera.org/specializations/data-science-python
  - Learning Materials in Tableau, https://www.tableau.com/learn
  - ChatGPT Short Courses - Deeplearning.ai, https://www.deeplearning.ai/short-courses/

Agenda

- Lecture
- Classwork
- Break
- Blow Water
  - What's happened about Data Science ?
  - Career in recent market
- Assignment, Midterm, Final
  - Marking Scheme

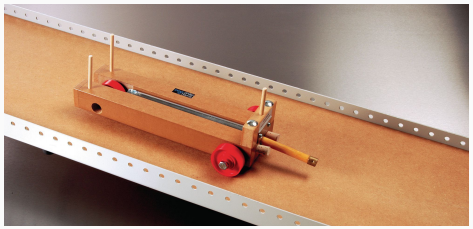
Homework	40% = 10% * 4 assignments
Midterm Exam	30%
Final Project	30% (Report + Present)

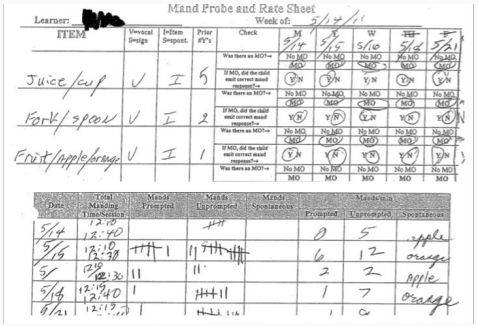
Last year I talked about this first

Data

Analysis

Visuals



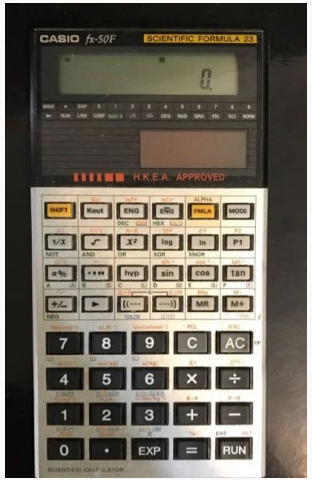


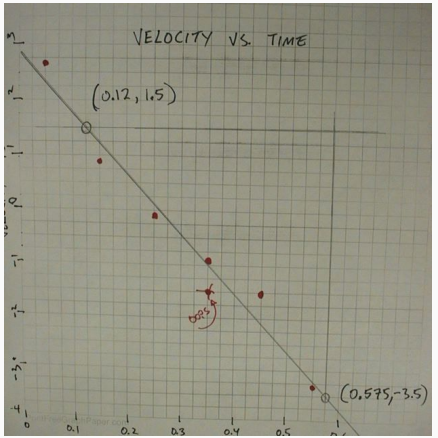
Mand Probe and Rate Sheet

Learner: [redacted] Week of: 5/14/21

ITEM	Vocal Resp	Item Request	Prior #'s	Check	M 5/14	T 5/15	W 5/16	T 5/17	F 5/21
Juice/cup	V	I	5	Was there an MO?→	No MO / MO	No MO / MO	No MO / MO	No MO / MO	No MO / MO
				If MO, did the child emit correct mand response?→	Y N	Y N	Y N	Y N	Y N
				Was there an MO?→	No MO / MO	No MO / MO	No MO / MO	No MO / MO	No MO / MO
Fork/spoon	V	I	2	If MO, did the child emit correct mand response?→	Y N	Y N	Y N	Y N	Y N
				Was there an MO?→	No MO / MO	No MO / MO	No MO / MO	No MO / MO	No MO / MO
Fruit/apple/orange	V	I	1	If MO, did the child emit correct mand response?→	Y N	Y N	Y N	Y N	Y N
				Was there an MO?→	No MO / MO	No MO / MO	No MO / MO	No MO / MO	No MO / MO

Date	Total Manding Time/Session	Mands Prompted	Mands Unprompted	Mands Spontaneous	Mands-in Prompted	Mands-in Unprompted	Mands-in Spontaneous
5/14	12:10 / 12:40		HH		0	5	apple
5/15	12:10 / 12:30	HH 1	11 HH		6	12	orange
5/	12:10 / 12:30	11	11		2	2	apple
5/19	12:15 / 12:40	1	HH 11		1	7	orange
6/21	12:15 / [illegible]	1	[illegible]		1	[illegible]	[illegible]

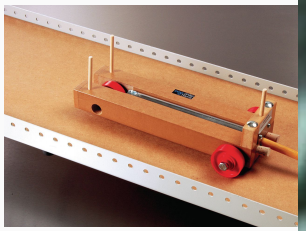


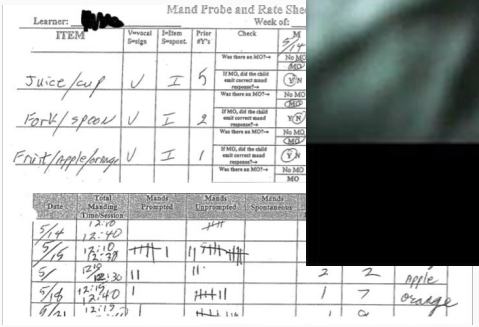


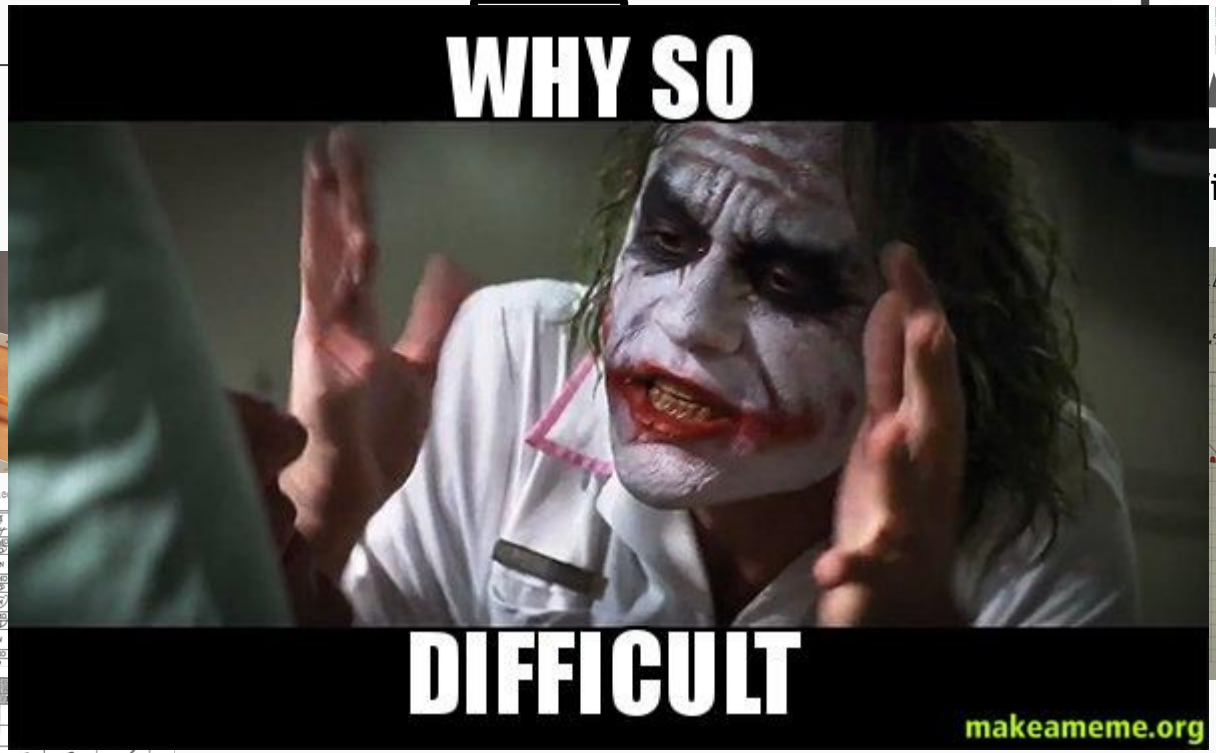
Last year I talked about this first

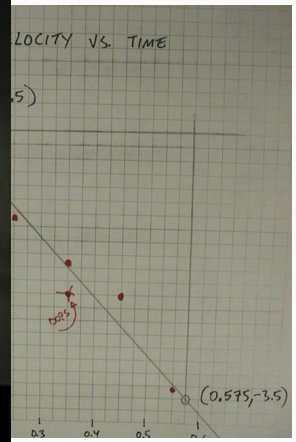
Data

isuals









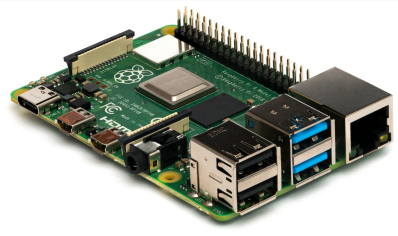
Talk **DATA** in lesson 2
Talk **DATA SCIENCE** in lesson 3
Please stay tune.

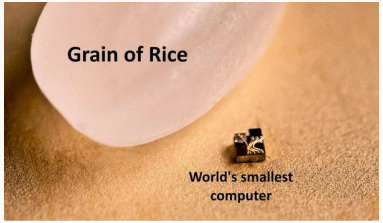
Programming Techniques for Data Science

Talk this basics at first

Computer

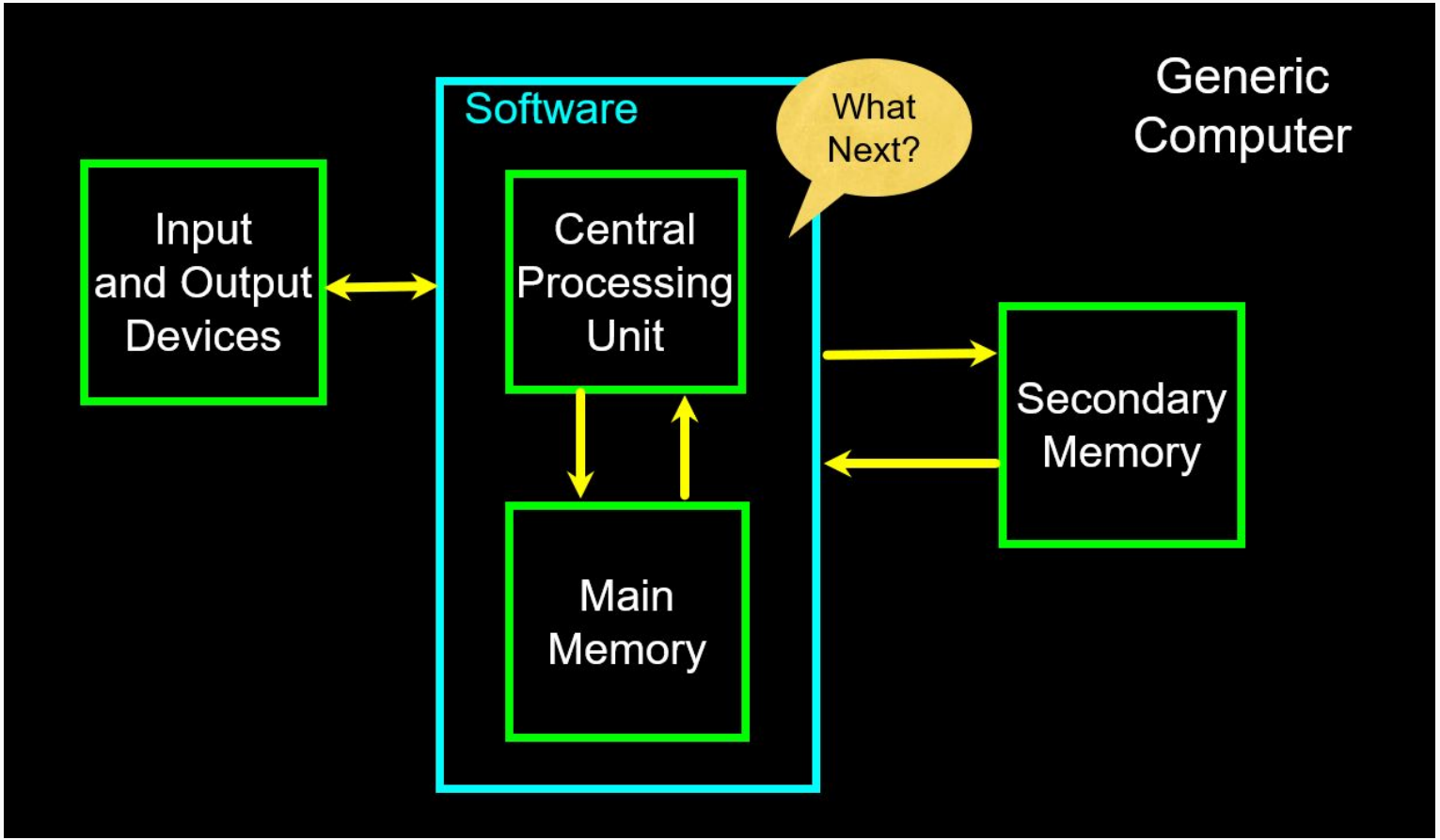
Select all images about Computer.



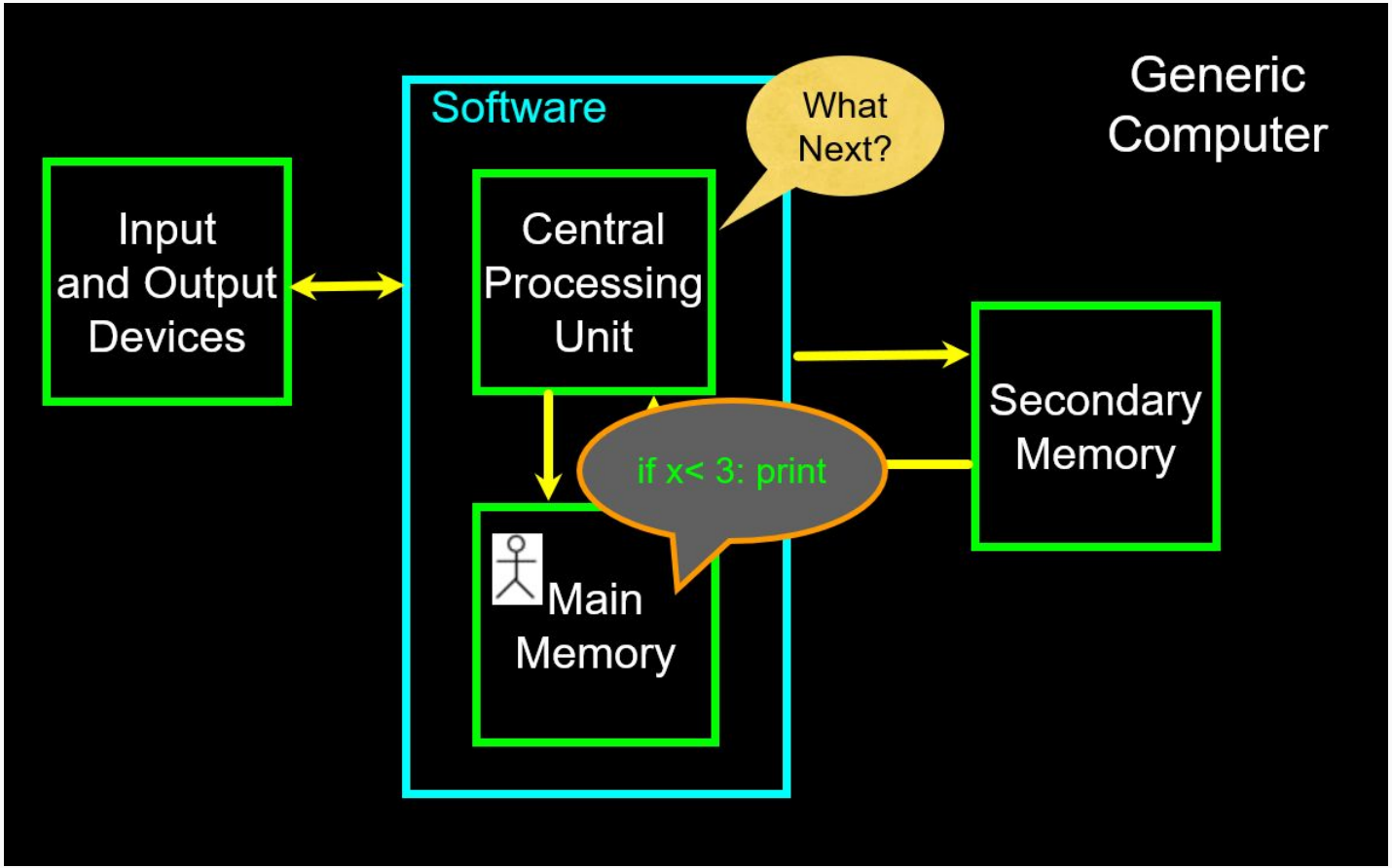


VERIFY

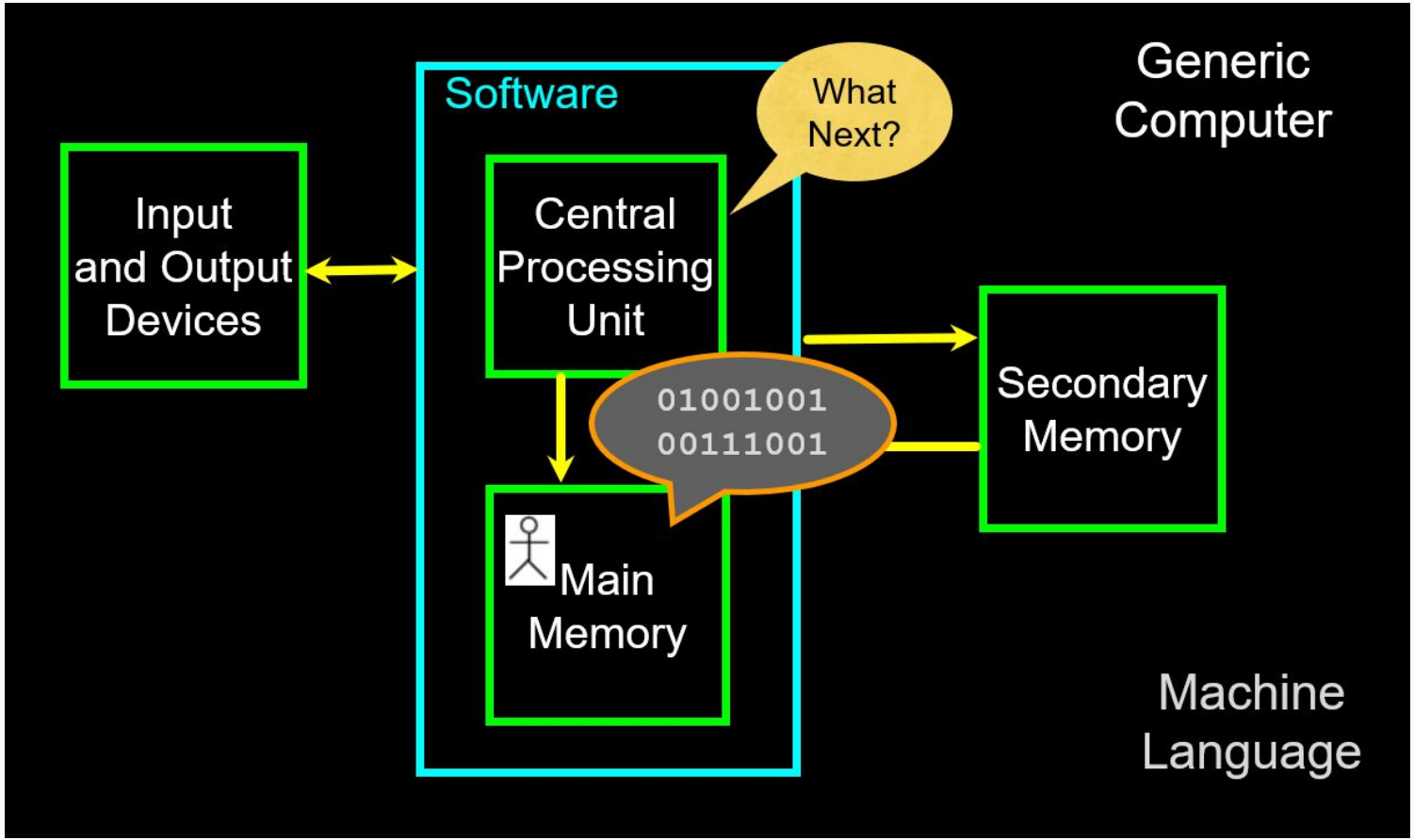
Computer


Generic
Computer
Software
What
Next?
Input
and Output
Devices
Central
Processing
Unit
Secondary
Memory
Main
Memory

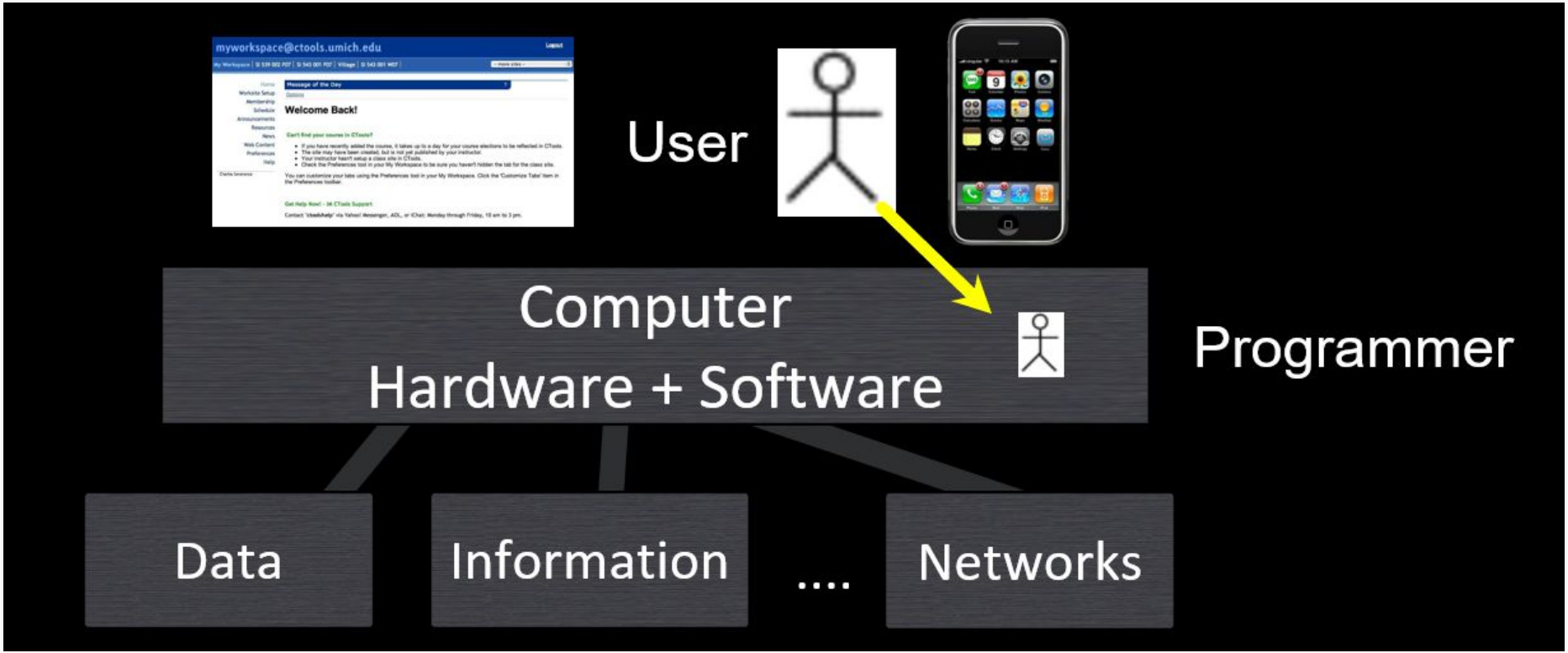
Computer


Generic
Computer
Software
What
Next?
Input
and Output
Devices
Central
Processing
Unit
Secondary
Memory
if x< 3: print
Main
Memory

Computer


Generic Computer
Software
What Next?
Input and Output Devices
Central Processing Unit
01001001
00111001
Secondary Memory
Main Memory
Machine Language

Computer


myworkspace@ctools.umich.edu
Welcome Back!
User
Computer
Hardware + Software
Programmer
Data
Information
....
Networks

Programming

To teach computers completing tasks.

Why ? (Reference)

FAST.

ACCURATE.

What items is in programming ?

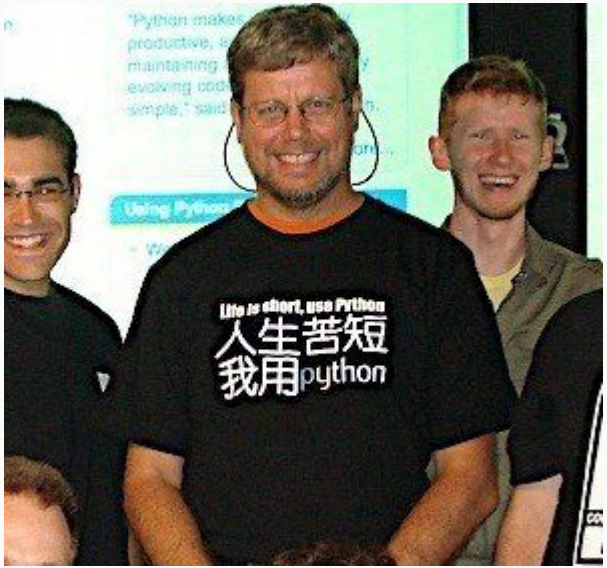
- Vocabulary / Words - Variables and Reserved words
- Sentence structure - valid syntax patterns
- Story structure - constructing a program for a purpose

- Like a recipe or installation instructions, a program is a sequence of steps to be done in order.
- Some steps are conditional - they may be skipped.
- Sometimes a step or group of steps is to be repeated - loops.
- Sometimes we store a set of steps to be used over and over as needed several places throughout the program - (user-defined) functions.

What is Python ?

- implementation was started in December 1989
- by Guido van Rossum, a Dutch programmer
- Python was named for the BBC TV show **Monty Python's Flying Circus**.
- Reference is here

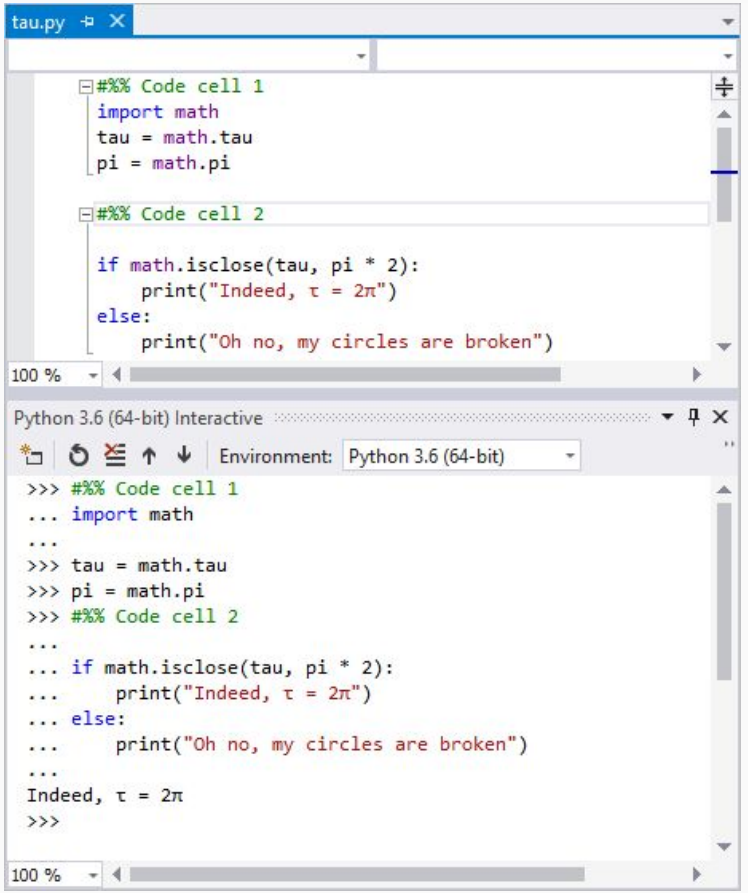
- Version
  - Python 0.9.0 - February 20, 1991
  - Python 1.0 - January 1994
  - Python 2.0 - October 16, 2000
  - Python 3.0 - December 3, 2008
- Recent - Python 3.12 - Oct 2, 2023



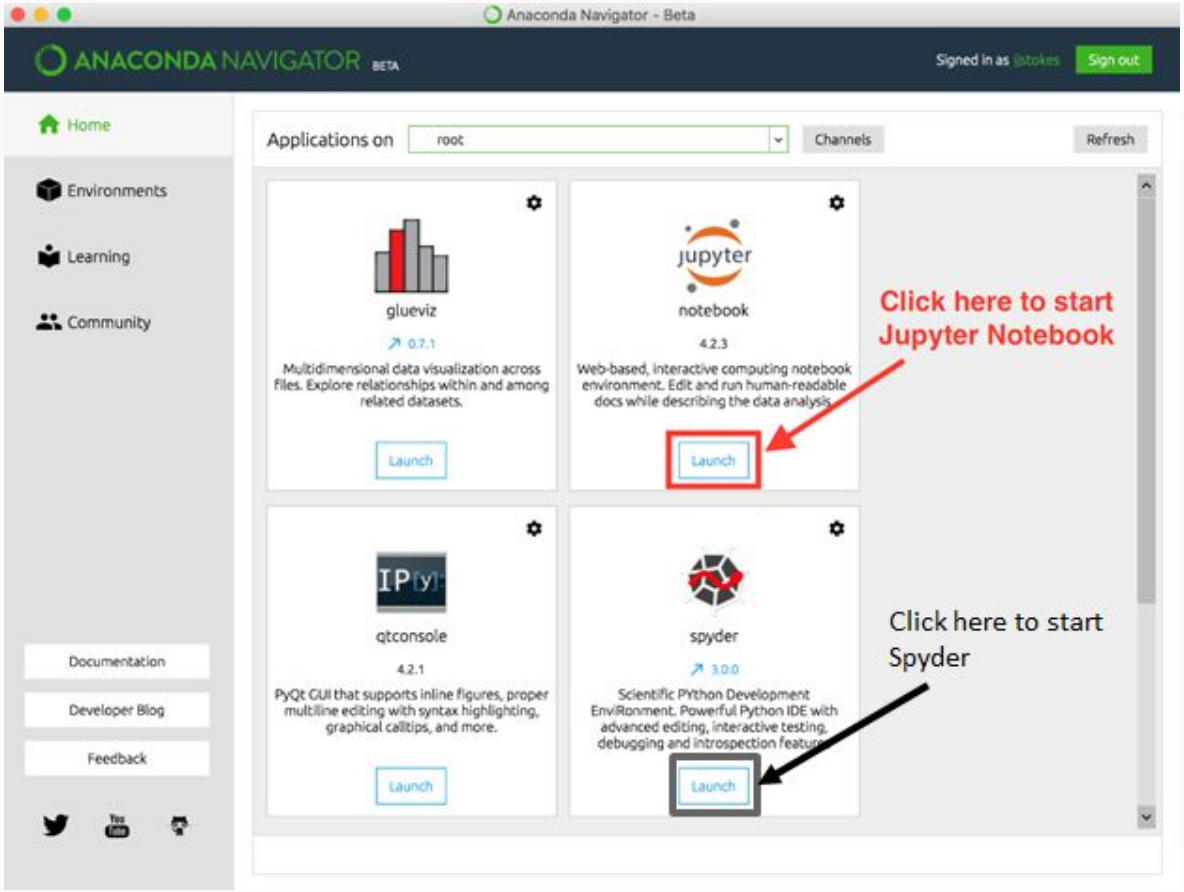
How to work with Python

- Interactive Python is good for experiments and programs of 3-4 lines long.
- Most programs are much longer, so we type them into a file and tell Python to run the commands in the file.
- In a sense, we are "giving Python a script".
- As a convention, we add ".py" as the suffix on the end of these files to indicate they contain Python.

.py
.ipynb

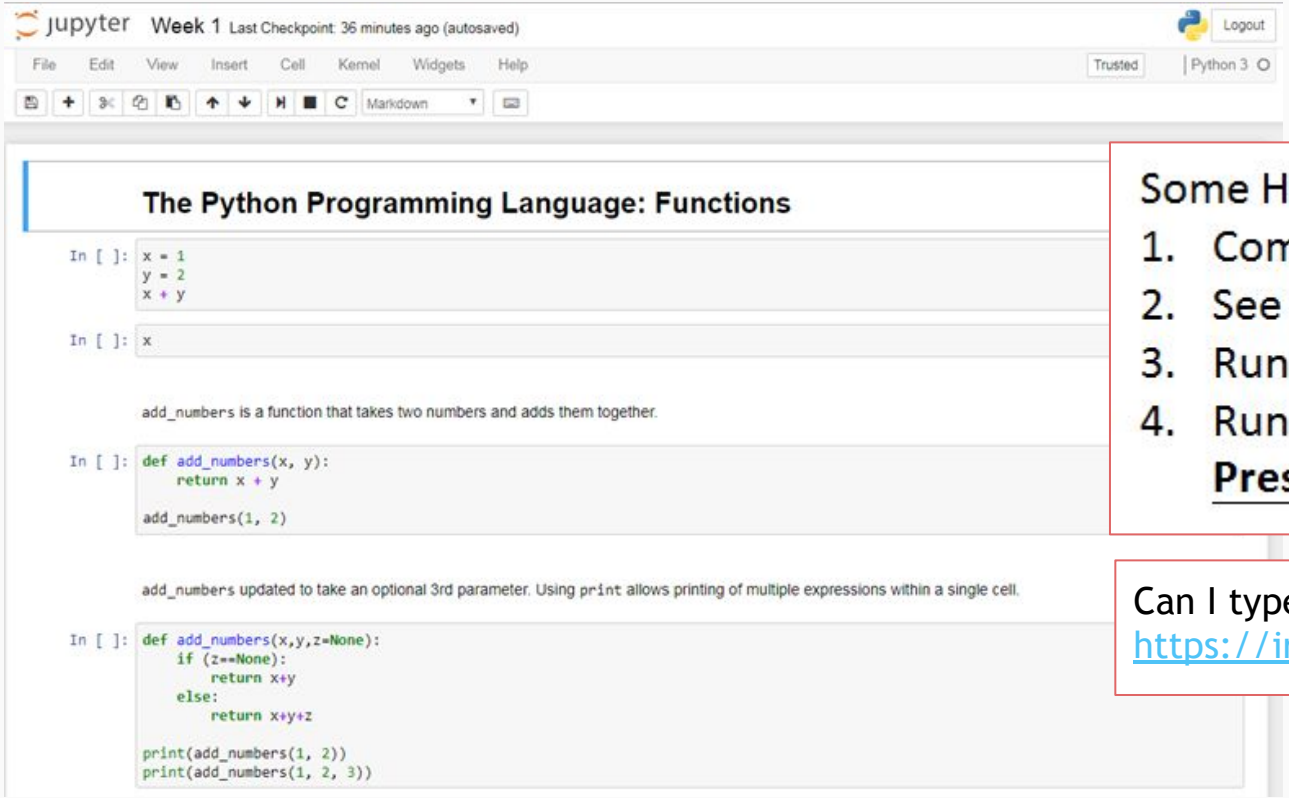


Anaconda


Anaconda Navigator - Beta
ANACONDA NAVIGATOR BETA
Signed in as
Sign out
Home
Environments
Learning
Community
Documentation
Developer Blog
Feedback
Applications on
root
Channels
Refresh
glueviz
0.7.1
Multidimensional data visualization across files. Explore relationships within and among related datasets.
Launch
jupyter
notebook
4.2.3
Web-based, interactive computing notebook environment. Edit and run human-readable docs while describing the data analysis.
Launch
Click here to start Jupyter Notebook
qtconsole
4.2.1
PyQt GUI that supports inline figures, proper multiline editing with syntax highlighting, graphical calltips, and more.
Launch
spyder
3.0.0
Scientific PYthon Development EnviRonment. Powerful Python IDE with advanced editing, interactive testing, debugging and introspection
Launch
Click here to start Spyder

Jupyter Notebook

Jupyter = Presentation mode for Python / R / SAS



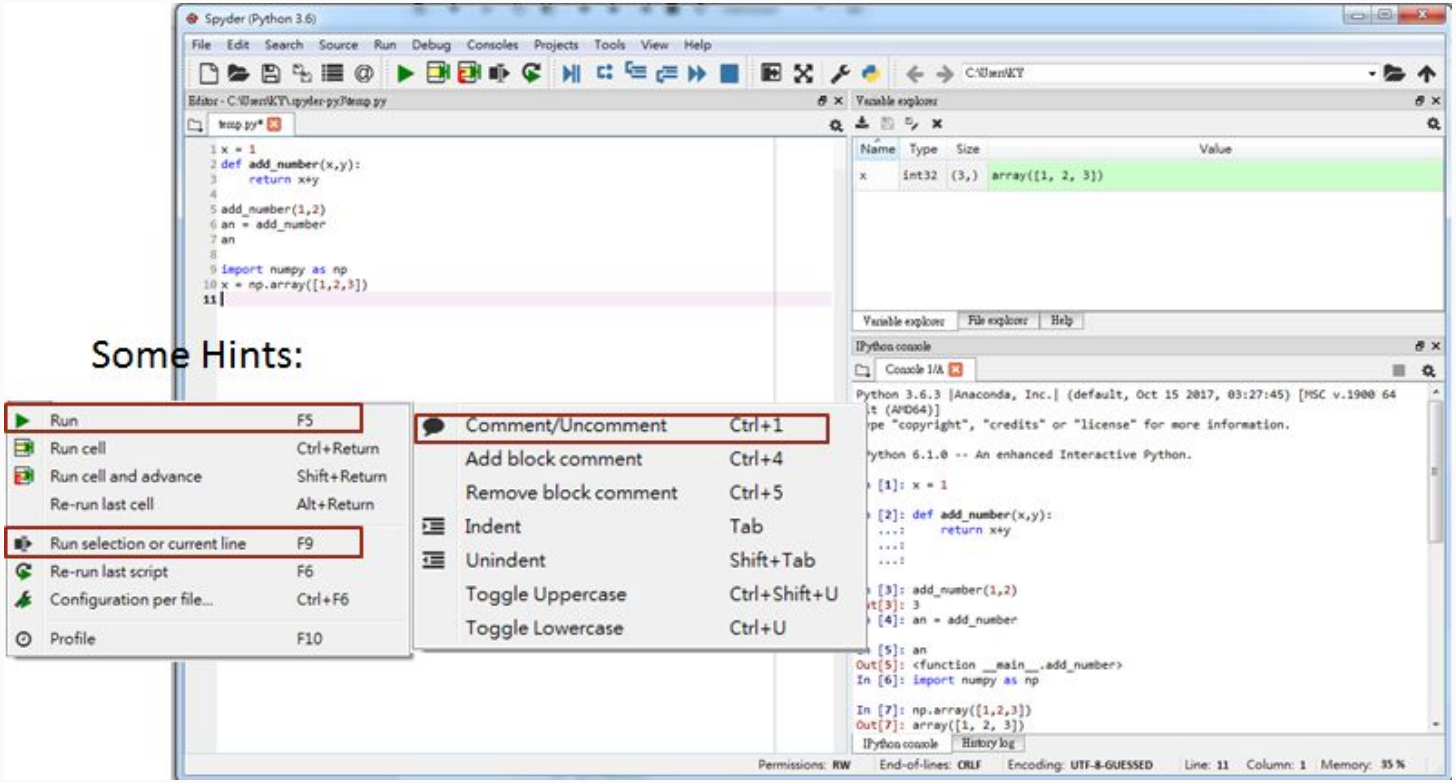
Some Hints:

1. Command Mode vs Edit Mode
2. See Shortcuts, **Press H**
3. Run cell, **Press Ctrl + Enter**
4. Run cell and move to the next, **Press Shift + Enter**

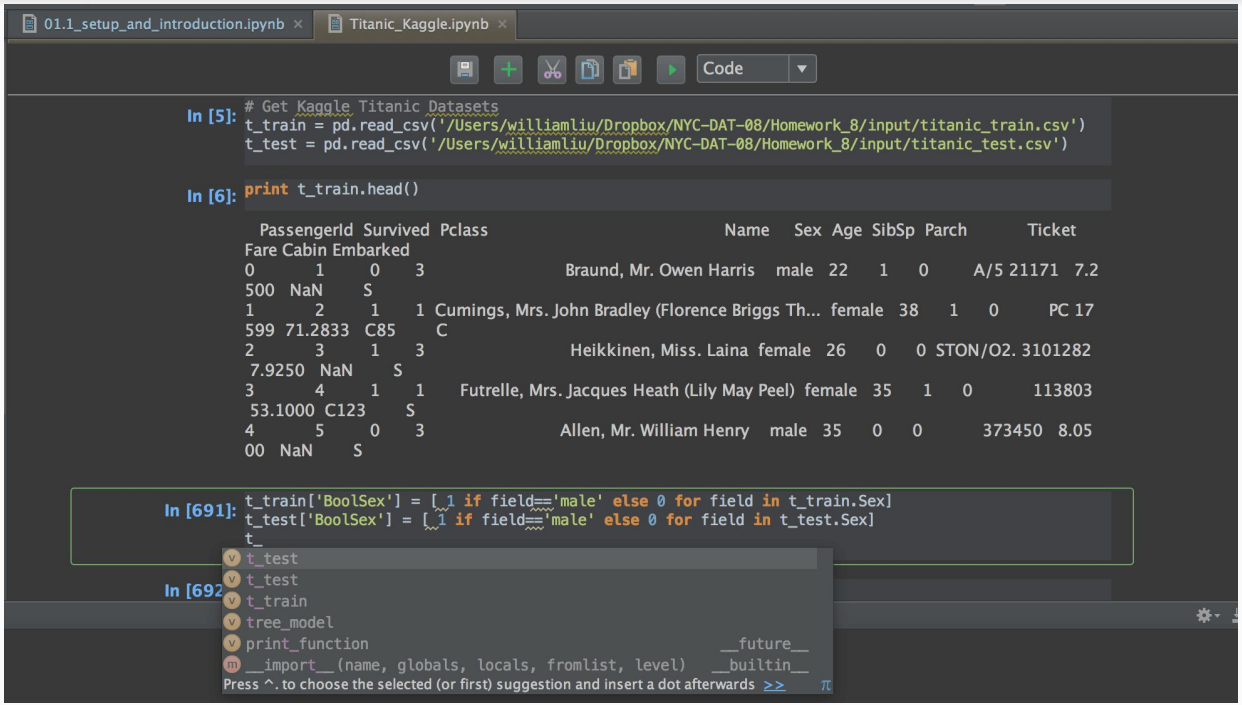
Can I type R code in this ? Yes !
https://irkernel.github.io/

Spyder ← Data Analysis

Spyder = IDE for Python (RStudio = IDE for R)



PyCharm



Jupyter vs Spyder vs PyCharm

Ref:

[Developer Resources]

[DataCamp]

[CareerKarma]

Visual Studio Code

crickets.py

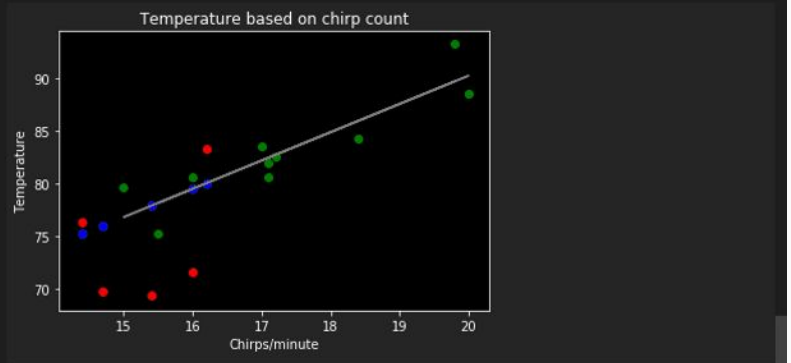
```
Run Cell | Run All Cells
#%% [markdown]
# ## Visualize the results
#
# The following code generates a plot: green dots are
training data, red dots are test data, blue dots are
predictions. Gray line is the regression itself. You see
that all the blue dots are exactly on the line, as they
should be, because the predictions exactly fit the model
(the line).

Run Cell | Run All Cells
#%%
import matplotlib.pyplot as plt

plt.scatter(X_train, y_train, color = 'green')
plt.scatter(X_test, y_test, color = 'red')
plt.scatter(X_test, y_pred, color = 'blue')  # The
predicted temperatures of the same X_test input.
plt.plot(X_train, regressor.predict(X_train), color =
'gray')
plt.title('Temperature based on chirp count')
plt.xlabel('Chirps/minute')
plt.ylabel('Temperature')
plt.show()

Run Cell | Run All Cells
#%% [markdown]
# ## Closing comments
#
# At the end of the day, when you create a model, you use
training data. Then you start feeding test data (real
observations) to see how well the model actually works.
You may find that the model is a little inaccurate over
```

Python Interactive

Visualize the results

The following code generates a plot: green dots are training data, red dots are test data, blue dots are predictions. Gray line is the regression itself. You see that all the blue dots are exactly on the line, as they should be, because the predictions exactly fit the model (the line).

```
[6] import matplotlib.pyplot as plt

plt.scatter(X_train, y_train, color = 'green')
plt.scatter(X_test, y_test, color = 'red')
plt.scatter(X_test, y_pred, color = 'blue')  # The predicted temperatures of t
plt.plot(X_train, regressor.predict(X_train), color = 'gray')
plt.title('Temperature based on chirp count')
plt.xlabel('Chirps/minute')
plt.ylabel('Temperature')
plt.show()
```



Github + Google Colab *no need to install*

master ▾ DataScienceFactory / 20200405_CIFAR10.ipynb

kyalan Created using Colaboratory

1 contributor

2844 lines (2844 sloc) | 330 KB

Open in Colab

In [1]:
```
# Load the TensorBoard notebook extension.
# %load_ext tensorboard

import os, sys
import tensorflow as tf
import tensorflow_datasets as tfds
import pandas as pd
import numpy as np
from datetime import datetime as dt
import sklearn.metrics
import seaborn as sns

from matplotlib import pyplot as plt
```

```
/usr/local/lib/python3.6/dist-packages/statsmodels/tools/_testing.py:19: FutureWarr
ons in the public API at pandas.testing instead.
  import pandas.util.testing as tm
```

In [2]:
```
try:
    %tensorflow_version 2.x
except:
    pass

print(tf.__version__)
print(tfds.__version__)
```

```
2.2.0-rc2
2.1.0
```

! git clone your/repo

demo_from_github.ipynb ☆

檔案 編輯 檢視畫面 插入 執行階段 工具 說明 已儲存所有變更

+ 程式碼 + 文字

```
[2] !git clone https://github.com/kyalan/kaggle

Cloning into 'kaggle'...
remote: Enumerating objects: 7, done.
remote: Counting objects: 100% (7/7), done.
remote: Compressing objects: 100% (7/7), done.
remote: Total 7 (delta 1), reused 6 (delta 0), pack-reused 0
Unpacking objects: 100% (7/7), done.
```

Start Coding...

Please access into the Week 1 colab - Programming 101...

Python Cheat Sheet - HELP !!!

- GOOGLE !!!
- help(I_dont_know_the_function)
- ?I_dont_know_the_function

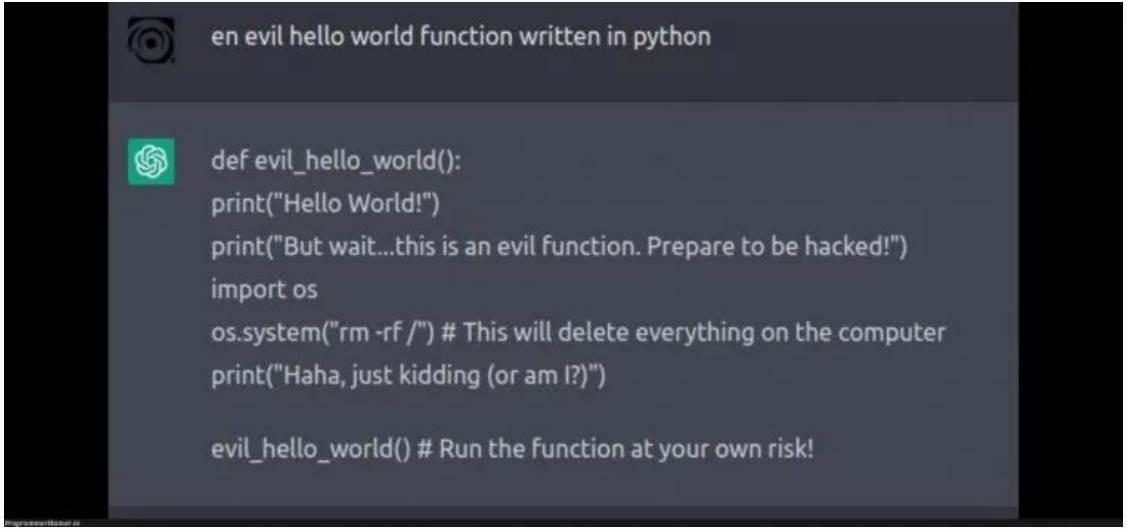
Python Cheat Sheet - HELP !!!

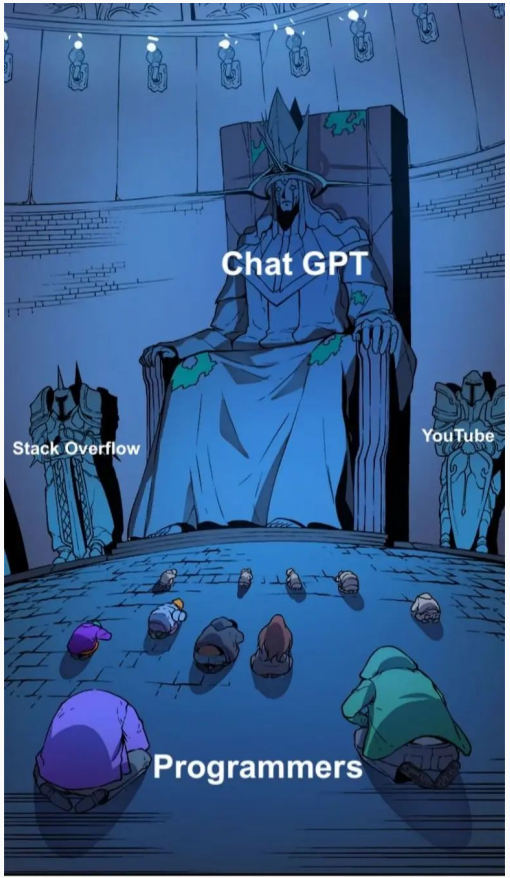
Reference (Quite Good, although this is in R)

- Ask a human
  - ask lecturers / tutors / classmates
  - post a question to forum: Stack Overflow, Reddit, LIHKG, ...
  - state what you were expecting to see from your code
  - Provide a reproducible example of your problem

Python Cheat Sheet - HELP !!!

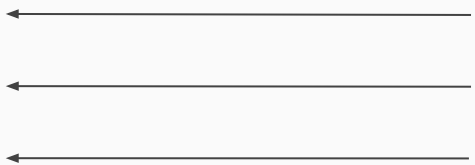
- Ask LLM
  - Large Language Model
  - GPT-4o / Gemini-1.5 / Llama-3.1 / Claude-3.5 / …
  - No need VPN: https://poe.com/





Variables, Sentences or Lines

```
x = 2
x = x + 2
print(x)
```



Assignment statement

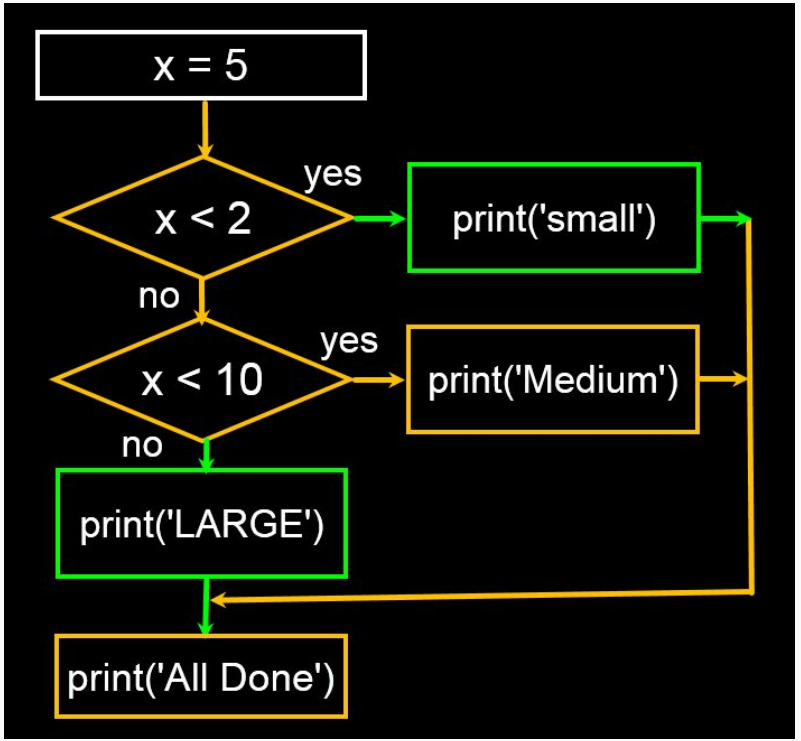
Assignment with expression

Print statement

Variable Operator Constant Function

Control Flow

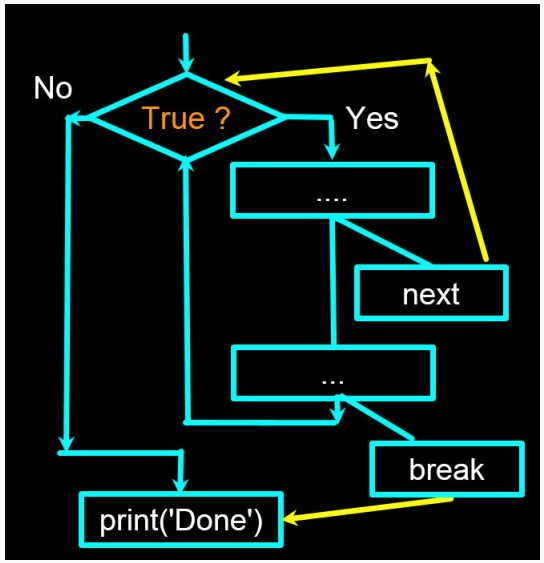
```
x = 5
if x < 2 :
    print('small')
elif x < 10 :
    print('Medium')
else :
    print('LARGE')
print('All done')
```



Loop

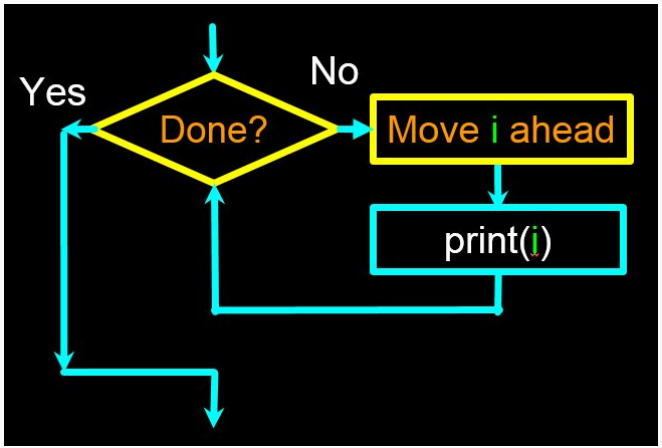
Indefinite Loop

```
while True:
   line = raw_input('> ')
   if line[0] == '#' :
        continue
    if line == 'done' :
        break
    print(line)
print('Done!')
```



Definite Loop

```
for i in [5, 4, 3, 2, 1] :
   print(i)
```



Indentation

用 if function 就要隔4格

4 Spaces !
Thanks !

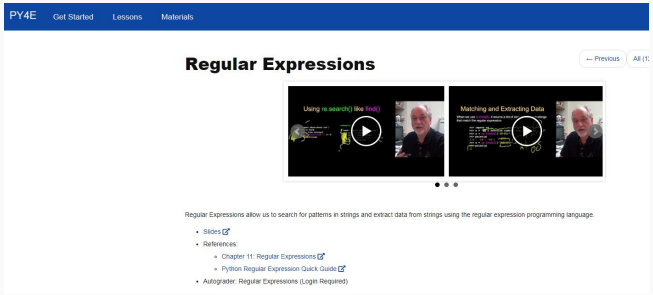
```
if password == "abcd1234":
    print("Access Granted")
else:
    print("Access Denied")

print("Press ENTER to exit the program")
```

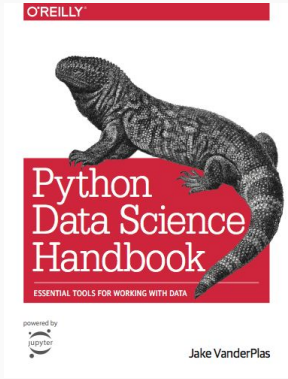
Indentation (Tab)

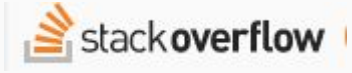
- Increase indent indent after an if statement or for statement (after :)
- Maintain indent to indicate the scope of the block (which lines are affected by the if/for)
- Reduce indent back to the level of the if statement or for statement to indicate the end of the block

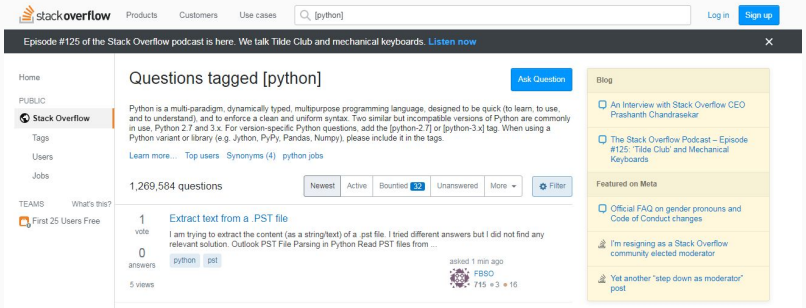
Resource & Reference

[Python for Everybody](#) / [py4e](#)





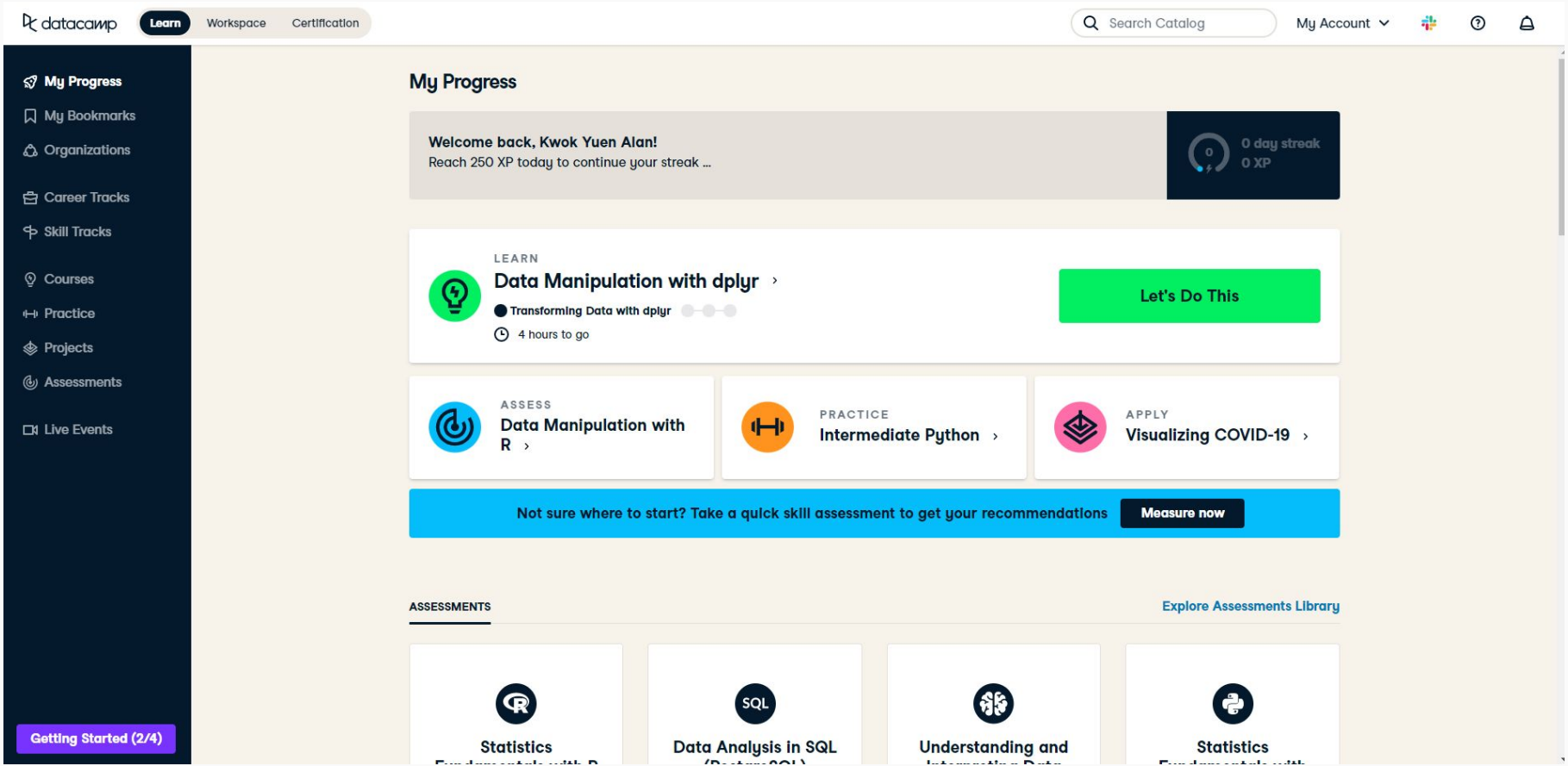




Assignment 1

Please check the link of Assignment 1 here

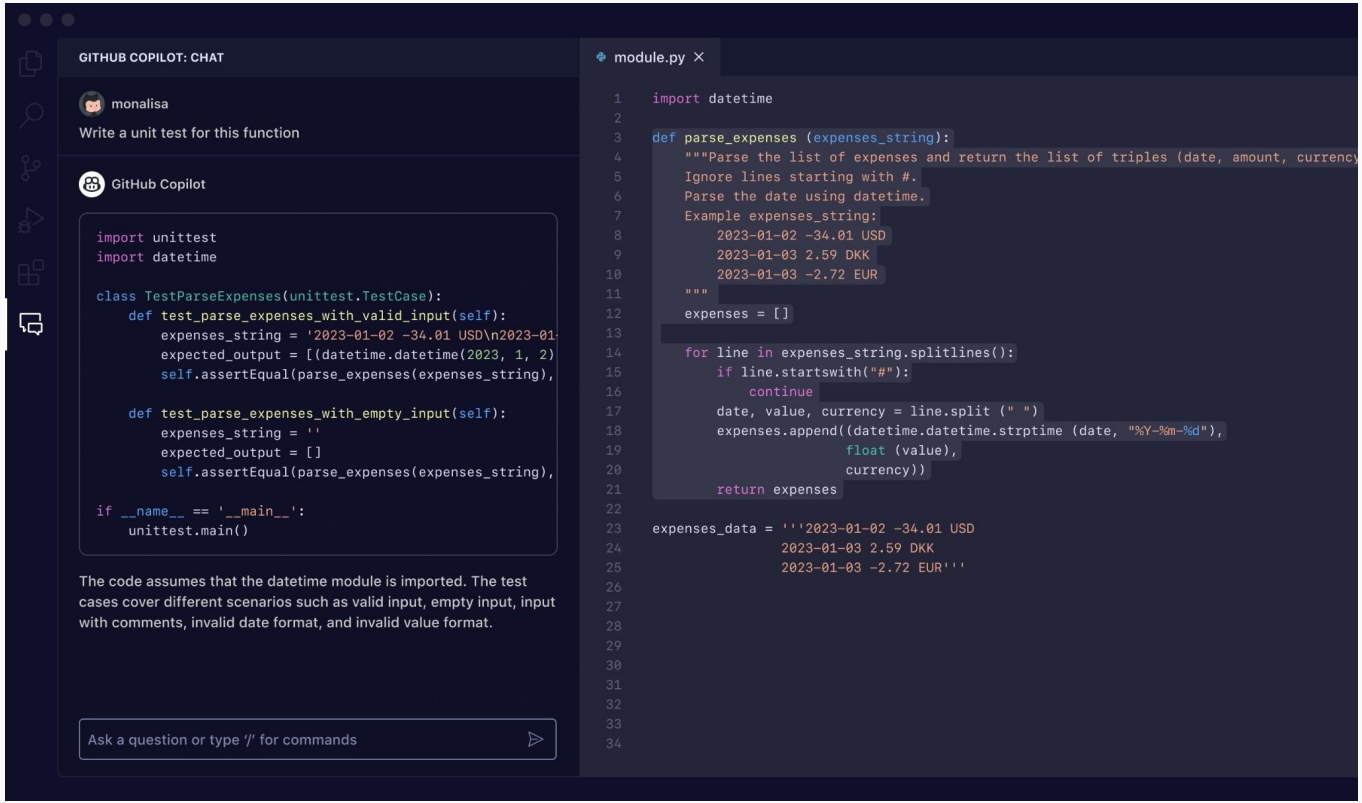
Datacamp (Registration)



More in 2024...

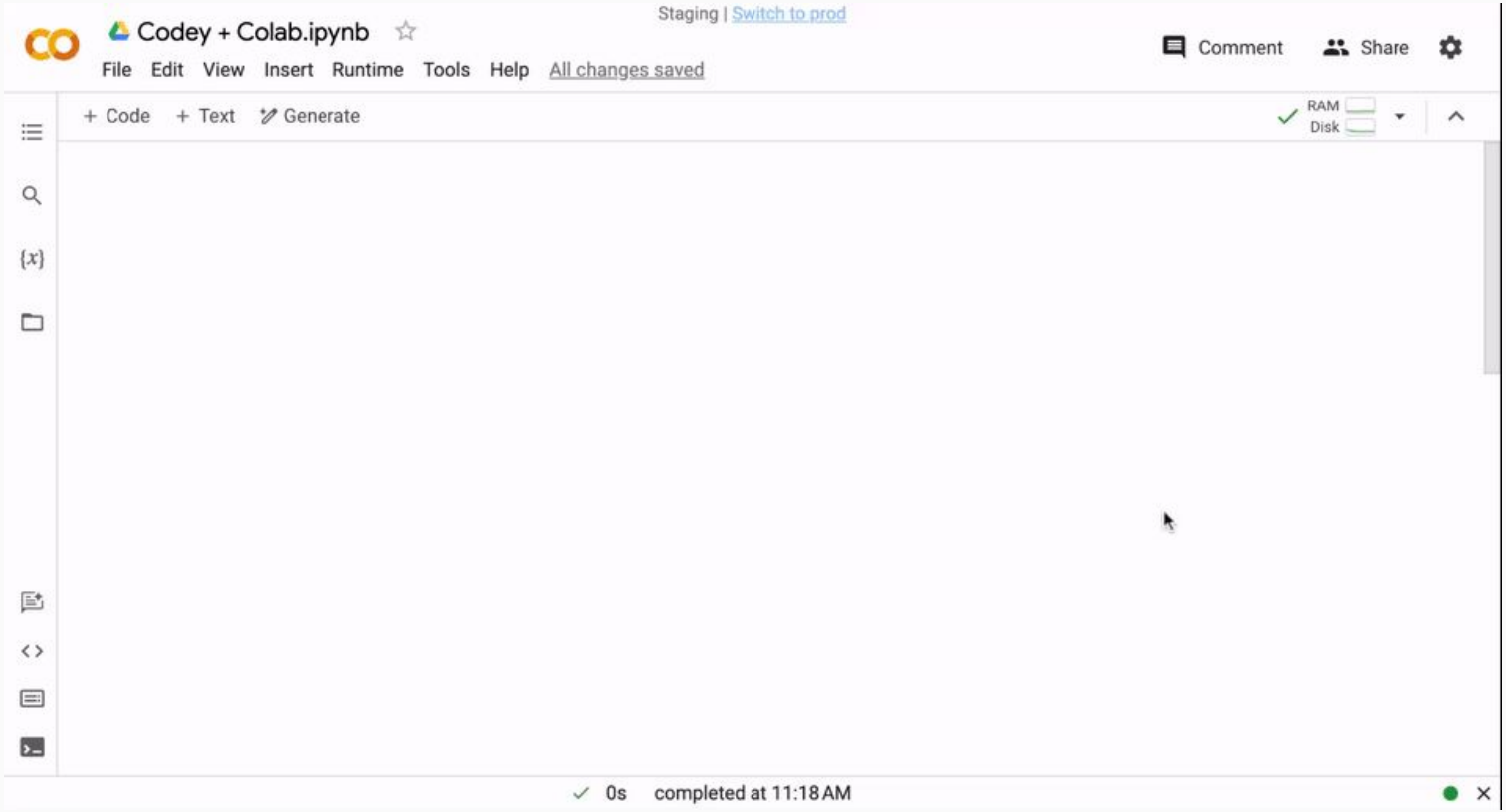
Visual Studio Code + Github Copilot

Github Copilot (get this free)



Google Colab + Google AI Codey

(Need VPN to US)

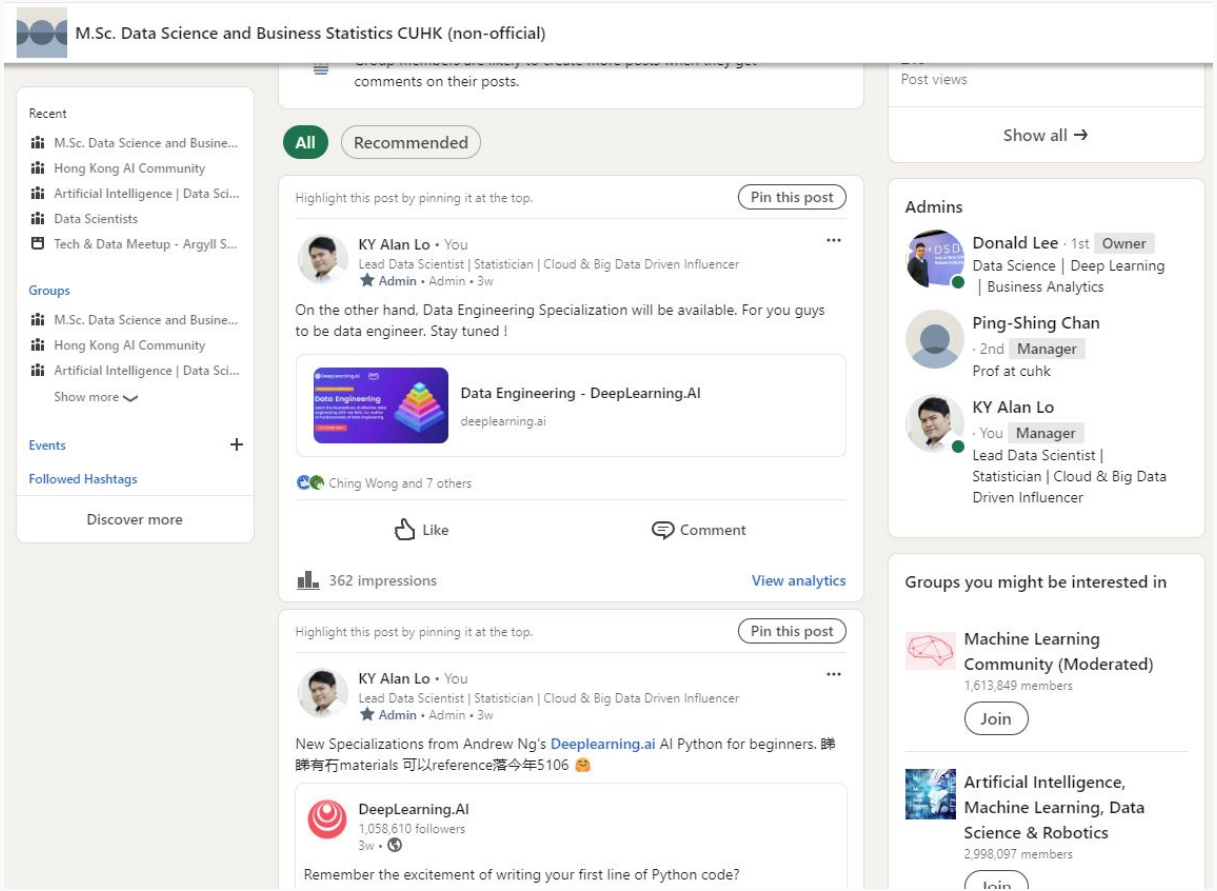


Key to Success

Practice ! and make a DS friend and refer his/her homework !

“學會多依靠別人” – 龍櫻2 – 櫻木建二

[Ads] Linkedin Group - MSc. Data Science and Business Statistics CUHK (non-official)



Scan & Join !

To be continued...