EE3801
Data Engineering Principles

A/Prof. Bharadwaj Veeravalli & *Dr. Yen Shih-Cheng*

Dept of ECE, NUS *EDIC, CDE, NUS*

elebv@nus.edu.sg *shihcheng@nus.edu.sg*

Lesson Plan – *What you will learn in this module?*

Part 1 - A/Prof Bharadwaj V

- Introduction to Data Engineering (DE) and computing platforms, DE concepts data wrangling and ETL/ELT, Data warehouses, Data pipelining concepts, current day big data architectures, design example of a BDP, introduction to Cluster and Cloud/DC platform architectures, resource sharing in cluster platforms

Part 2 - Dr. Yen Shih-Cheng

- Amazon Web Services Elastic Compute Cloud, Parallel Cluster, Elastic Block System file system, SLURM Workload Manager, data pipeline construction and operation, high-throughput data visualization and inspection, vectorization techniques in data analysis.

(c) Bharadwaj V, Dept of ECE, NUS & SC Yen, EDIC, NUS (2023)

Assessment – Marks Distribution

- Labs 1 to 3: 35%
- Quiz 1: 15%
- Labs 4 to 8: 35%
- Quiz 2: 15%

Part 1: All Lectures, Lab sessions, and Quiz are held in this venue LT3(Mon)/LT7(Thurs).

Part 2: All Lectures and Quiz will be held in LT7 on Thursdays, while Lab sessions will be held in E2-03-08/09 on Mondays.

Part 1 – Contents, Assessments & Marks Distribution

Part 1: 50%

- Lab sessions 35% (3 Lab sessions)

Quiz 1 (Chapters 1, 2, and 4): 15%

Date: Oct 2, 2023

Time: 12pm-1pm (60 mins)

Venue: LT3

Part 1 Lab Schedule

LAB 1
Aug 30 - Lab 1 assignment release
Aug 31 - Lab 1 briefing by GAs (45 mins)
Sept 7 - Deadline Lab 1

LAB 2
Sept 6 - Lab 2 assignment release
Sept 7 -Lab 2 briefing by GAs (45 mins)
Sept 14 - Deadline Lab 2

Work on your individual PCs/Laptops

Lab 3:
Sept 13 Lab 3 assignment release
Sept 14 - Lab 3 briefing (45 mins)
Sept 21 - Deadline Lab 3

(Your first ride on AWS! ☺)

(c) Bharadwaj V, Dept of ECE, NUS & SC Yen, EDIC, NUS (2023)

Part 2 - Assessments & Marks Distribution

Part 2: 50%

- Lab sessions: 35%

  5 Lab sessions all equally weighted

- Quiz 2 (hands on DE): 15%

  Date: Nov 16, 2023 (Thursday)

  Time: 3 pm

  Venue: LT7

(c) Bharadwaj V, Dept of ECE, NUS & SC Yen, EDIC, NUS (2023)

Lesson Plan – Part 2

Week	Monday	Thursday
7	Quiz 1 (Part 1)	Lecture
8	Lab 4	Lecture
9	Lab 5	Lecture
10	Lab 6	Lecture
11	Lab 7	Lecture
12	Lab 8	Lecture
13	Public Holiday	Quiz 2 (Part 2)

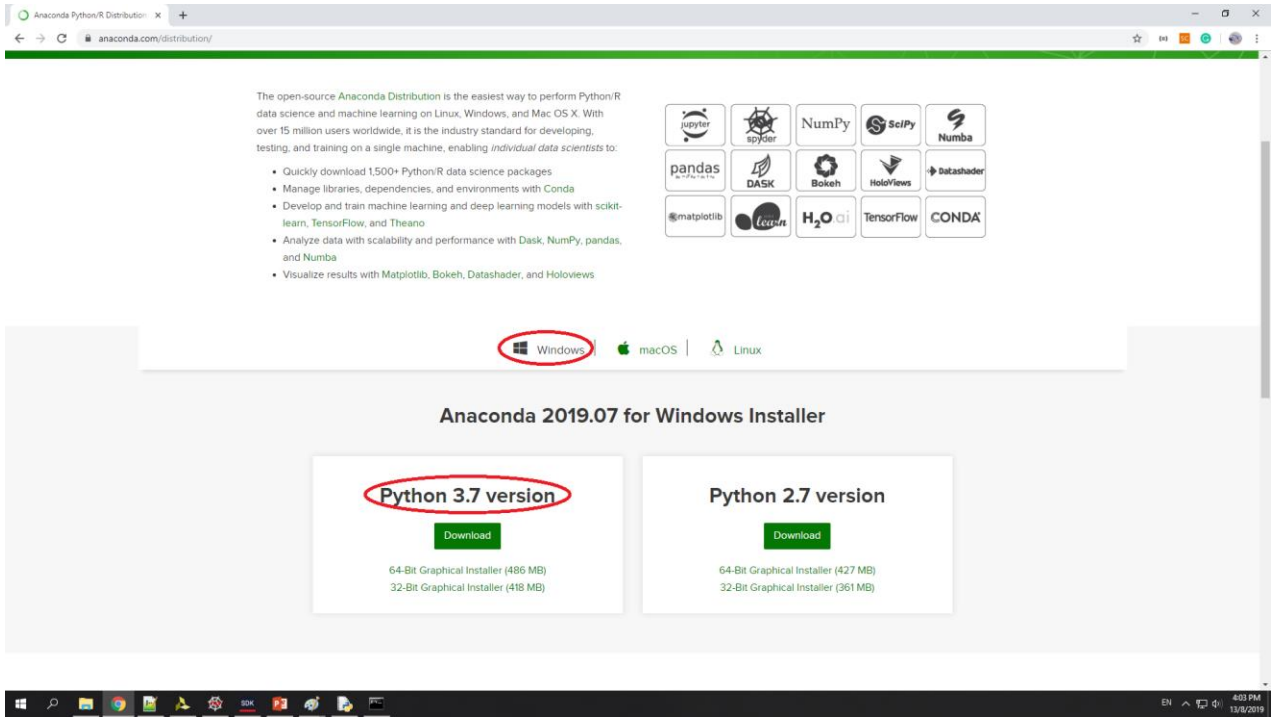
Labs will be held in E2-03-08/09

Lectures and Quiz 2 will be held in LT7

(c) Bharadwaj V, Dept of ECE, NUS & SC Yen, EDIC, NUS (2023)

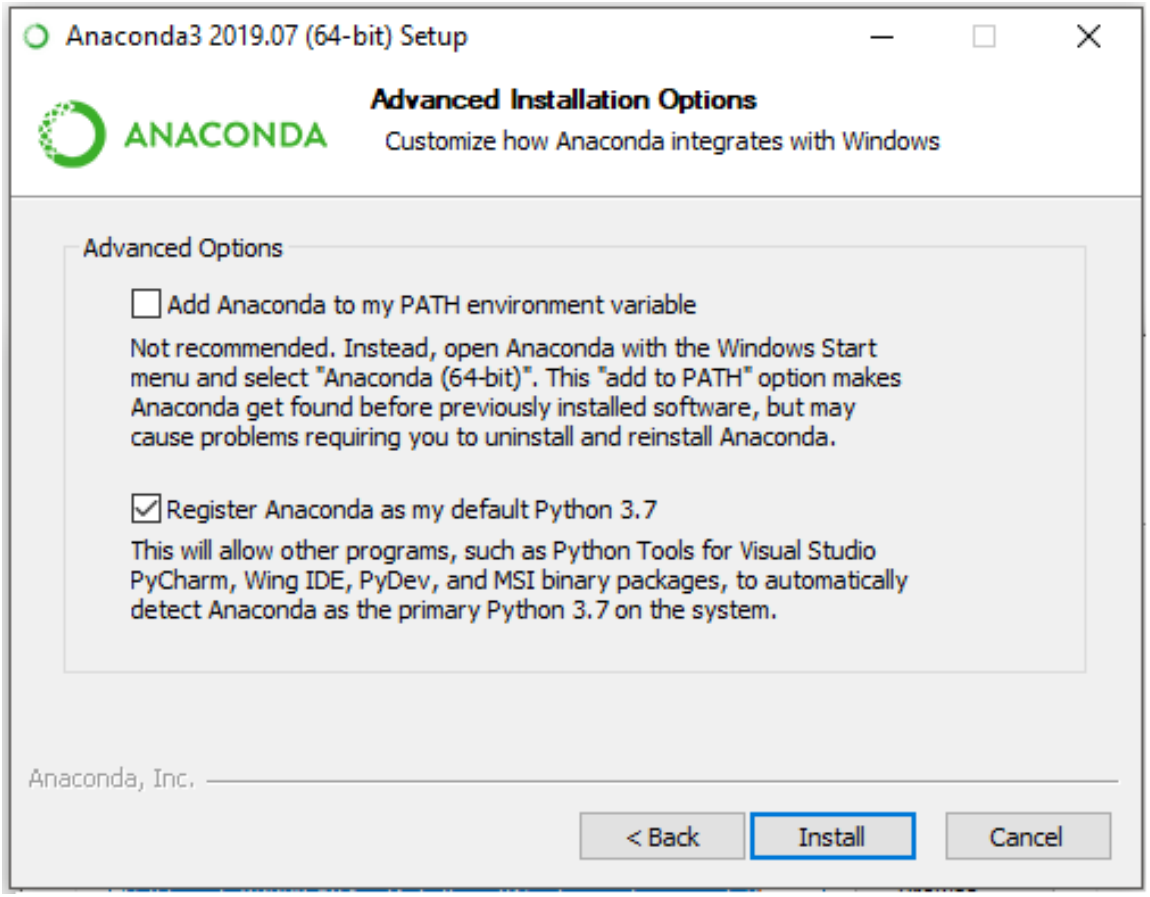
Anaconda Installation

- Download Anaconda for Python 3.x from here: https://www.anaconda.com/distribution/#download-section
- https://docs.conda.io/en/latest/miniconda.html (Miniconda) [~250 MB]
- https://katiekodes.com/setup-python-windows-miniconda/ (Useful)



Anaconda Installation

- Click Next until the screen below. Select the options as shown in the screenshot. Click Install.



Anaconda Installation

- Visit: https://www.datacamp.com/community/tutorials/installing-anaconda-windows

The above site takes you through a step-by-step installation process;

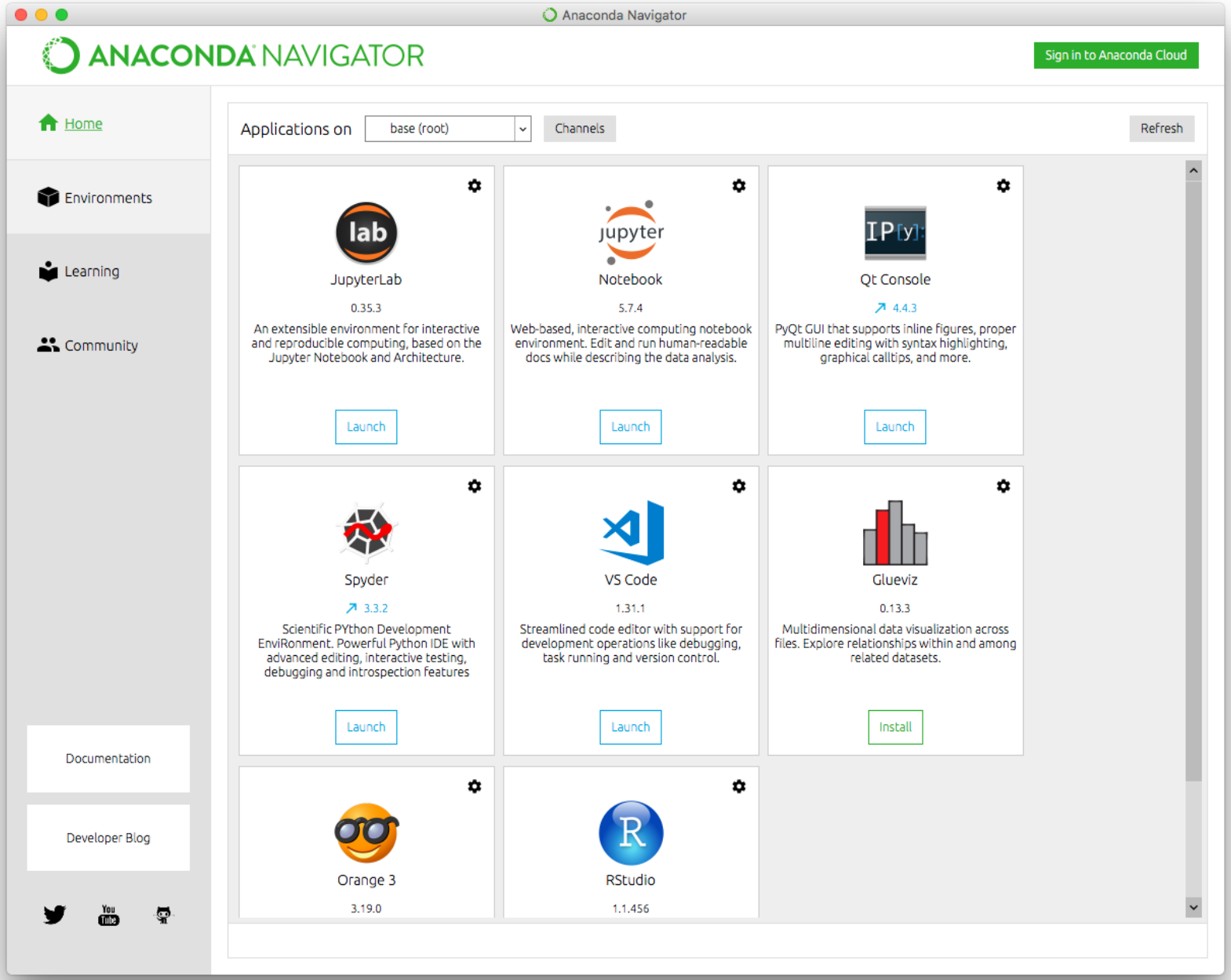
Note that with Anaconda lots of packages (~400+) will be installed; So, if you need other packages, you need to install on your own:

Use:

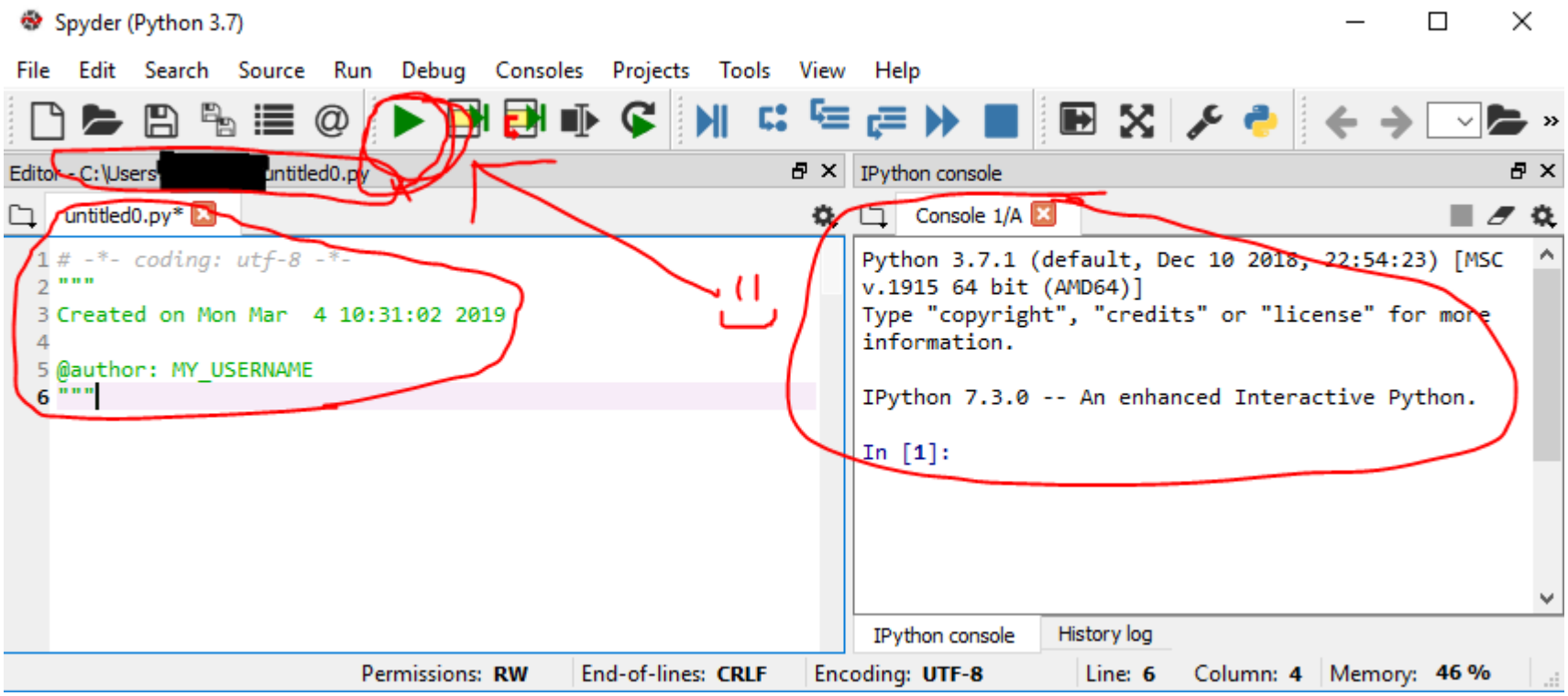
- In Miniconda – IDLE is available; Juypter notebook is NOT available; So, just like other pkgs, you can install using ***conda install jupyter*** should install it;
- With Anaconda *Jupyter, Spyder, and IDLE* are already built-in.
- Second part of this module exclusively uses Spyder editor; For the first part you can use either IDLE or Spyder;

Installing packages required:

- pip install *pkg-name* *(or)* > conda install *pkg-name*


Anaconda Navigator
ANACONDA NAVIGATOR
Sign in to Anaconda Cloud
Home
Environments
Learning
Community
Documentation
Developer Blog
Applications on
base (root)
Channels
Refresh
JupyterLab
0.35.3
An extensible environment for interactive and reproducible computing, based on the Jupyter Notebook and Architecture.
Launch
Notebook
5.7.4
Web-based, interactive computing notebook environment. Edit and run human-readable docs while describing the data analysis.
Launch
Qt Console
4.4.3
PyQt GUI that supports inline figures, proper multiline editing with syntax highlighting, graphical calltips, and more.
Launch
Spyder
3.3.2
Scientific PYthon Development EnviRonment. Powerful Python IDE with advanced editing, interactive testing, debugging and introspection features
Launch
VS Code
1.31.1
Streamlined code editor with support for development operations like debugging, task running and version control.
Launch
Glueviz
0.13.3
Multidimensional data visualization across files. Explore relationships within and among related datasets.
Install
Orange 3
3.19.0
RStudio
1.1.456


spyder 3
Initializing...


Spyder (Python 3.7)
File Edit Search Source Run Debug Consoles Projects Tools View Help
Editor - C:\Users untitled0.py
untitled0.py*
-*- coding: utf-8 -*-
"""
Created on Mon Mar 4 10:31:02 2019
@author: MY_USERNAME
"""
IPython console
Console 1/A
Python 3.7.1 (default, Dec 10 2018, 22:54:23) [MSC v.1915 64 bit (AMD64)]
Type "copyright", "credits" or "license" for more information.
IPython 7.3.0 -- An enhanced Interactive Python.
In [1]:
IPython console
History log
Permissions: RW
End-of-lines: CRLF
Encoding: UTF-8
Line: 6
Column: 4
Memory: 46 %

What next?

Before Aug 30, 2023:

- Install Python (*miniconda / Anaconda*)
- Editors – IDLE / Spyder
- Python packages that will/may be used:
  - – Numpy;
  - -- Scipy;
  - -- Pandas;
  - -- Matplotlib* / Seaborn; (* Recommended)

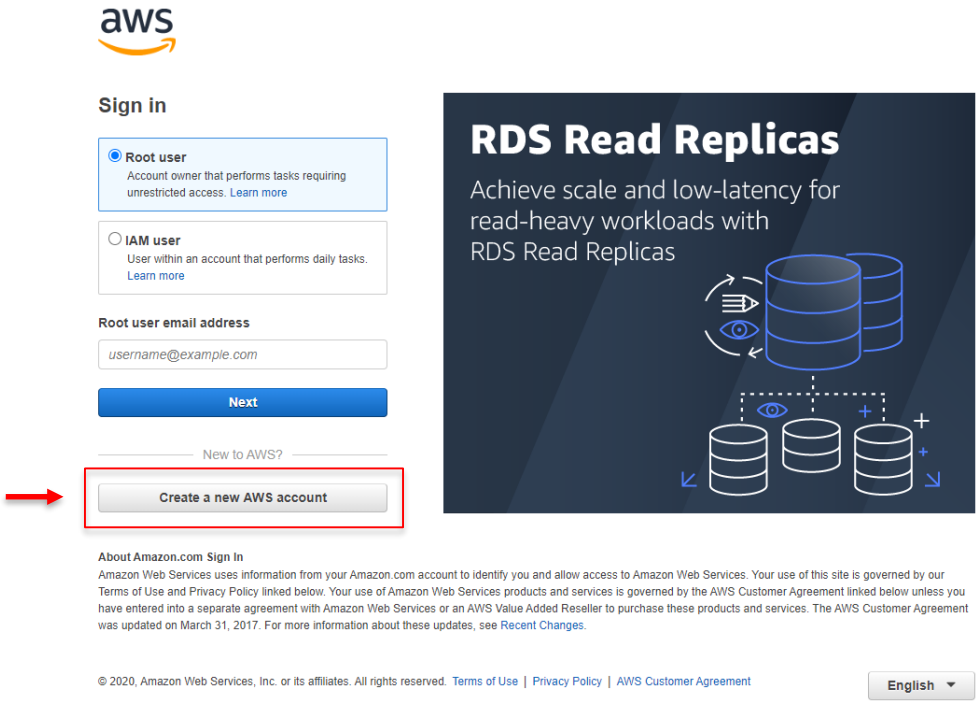
What next?

Before Aug 30, 2023:

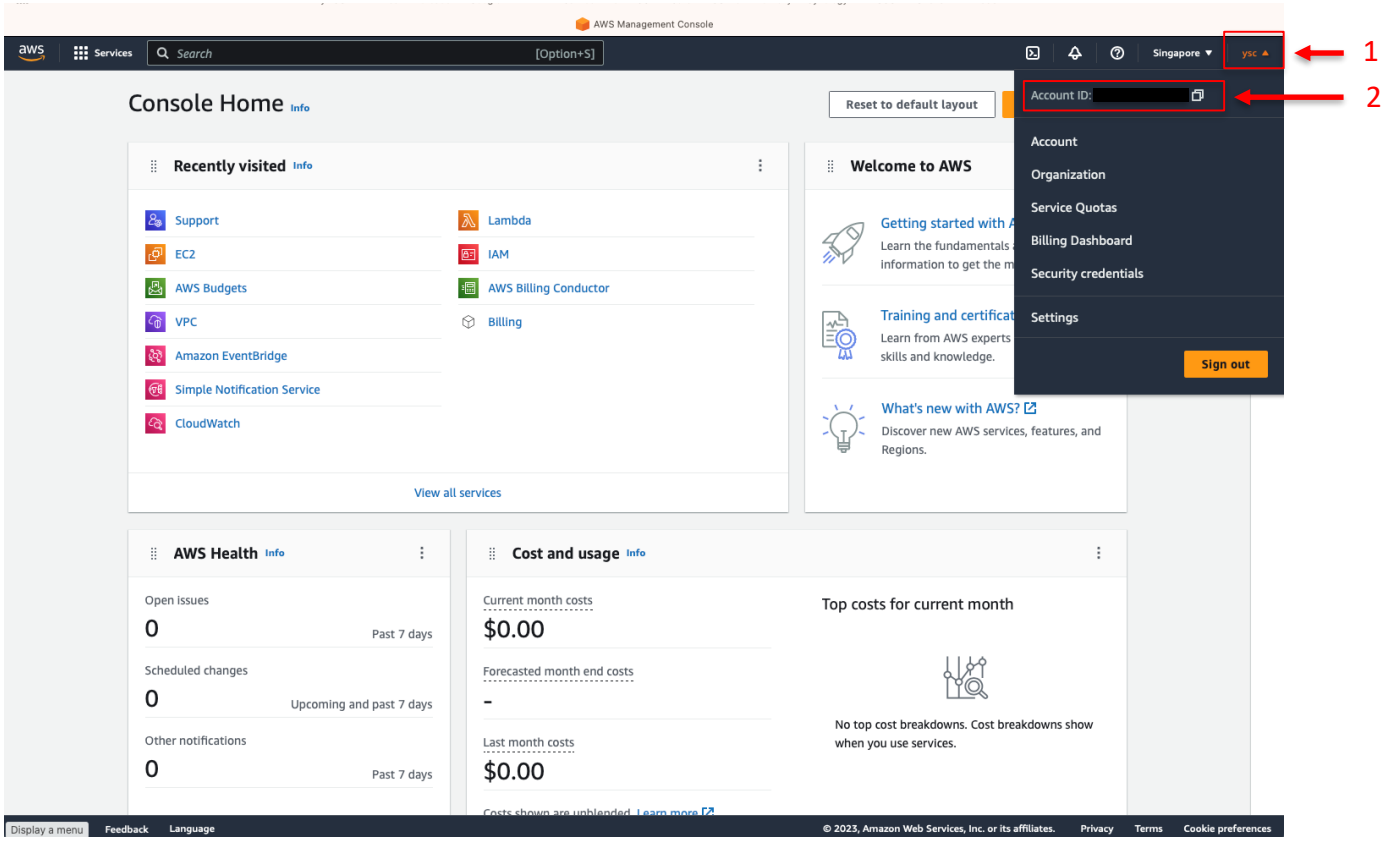
- Get ready by revisiting your *vanilla* Python fundamentals!
- Python
  - All data structures: Lists, tuples, sets, dictionaries;
  - Handling Python arrays – 1D & 2D array;
  - Basic python methods used in the above data structures;
  - Slicing concepts;
  - Converting one data structure to other types
- Go over the definitions of *certain basic statistical quantities* – mean, median, mode, std. deviation, normal distribution, etc.

Sign up for AWS

- https://console.aws.amazon.com/console/home?nc2=h_ct&src=header-signin



Get Account Number from AWS Management Console



Submit AWS and Zoom Information

- Canvas Ungraded Survey (under Quizzes)
  - Information for Labs
    - Submit 12-digit AWS account number
    - Indicate what kind of computer you will be using to complete the labs (e.g. Windows, Mac, or Linux)
  - Submit by Aug 27 (Sun), 11:59 pm

Annex – Quick review of basic Python methods

(c) Bharadwaj V, Dept of ECE, NUS & SC Yen, EDIC, NUS (2023)

Python - List methods

(Note: This is not an exhaustive list)

FUNCTION	DESCRIPTION
append()	Add an element to the end of the list
extend()	Add all elements of a list to the another list
insert()	Insert an item at the defined index
remove()	Removes an item from the list
pop()	Removes and returns an element at the given index
clear()	Removes all items from the list
index()	Returns the index of the first matched item
count()	Returns the count of number of items passed as an argument
sort()	Sort items in a list in ascending order
reverse()	Reverse the order of items in the list
copy()	Returns a copy of the list
len()	Returns the length of the list (# of elements in the list)

DIY!

Create an empty list and demonstrate some of the above methods; Observe how elements are accessed!

(c) Bharadwaj V, Dept of ECE, NUS & SC Yen, EDIC, NUS (2023)

Python – Tuples

(Note: This is not an exhaustive list)

Method	Description
count()	returns occurrences of element in a tuple
index()	returns smallest index of element in tuple
len()	returns Length of an Object
max()	returns largest element
min()	returns smallest element
reversed()	returns reversed iterator of a sequence
slice()	creates a slice object specified by range()
sum()	Add items of an iterable
tuple()	Creates a Tuple

DIY!

Create an empty tuple and demonstrate some of the above methods

(c) Bharadwaj V, Dept of ECE, NUS & SC Yen, EDIC, NUS (2023)

Python – Sets

(Note: This is not an exhaustive list)

DIY!

Method	***Description***
add()	Adds an element to the set
clear()	Removes all the elements from the set
copy()	Returns a copy of the set
difference()	Returns a set containing the difference between two or more sets
discard()	Remove the specified item
intersection()	Returns a set, that is the intersection of two other sets
isdisjoint()	Returns whether two sets have a intersection or not
issubset()	Returns whether another set contains this set or not
issuperset()	Returns whether this set contains another set or not
pop()	Removes an element from the set
remove()	Removes the specified element
union()	Return a set containing the union of sets
update()	Update the set with the union of this set and others

(c) Bharadwaj V, Dept of ECE, NUS & SC Yen, EDIC, NUS (2023)

Python – Dictionary

(Note: This is not an exhaustive list)

Method	Description
clear()	Removes all the elements from the dictionary
copy()	Returns a copy of the dictionary
fromkeys()	Returns a dictionary with the specified keys and values
get()	Returns the value of the specified key
items()	Returns a list containing a tuple for each key value pair
keys()	Returns a list containing the dictionary's keys
pop()	Removes the element with the specified key
popitem()	Removes the last inserted key-value pair
update()	Updates the dictionary with the specified key-value pairs
values()	Returns a list of all the values in the dictionary

DIY!

Python – Arrays

- Some useful array methods include:

append() insert() pop() reverse()
remove() index() count() extend() ... *and more*

- Numpy – Matrix operations

DIY!

Create an array of integers and demonstrate some of the above methods; Observe how elements are accessed!

Ethics in Computer Programming

Following Code of Ethics in Computer Programming is based on defunct International Programmer's Guild.

A programmer must...
...never create or distribute malware.
...never write code that is obfuscated or intentionally difficult to follow.
...never write documentation that is intentionally confusing or inaccurate.
...never reuse copyrighted code unless the proper license is purchased or permission is obtained.
...acknowledge (verbally and in source code comments) the work of other programmers on which the code is based, even if substantial changes are made.
...never write code that is deliberately inefficient with the intent of later claiming credit for making efficiency improvements.
...never intentionally introduce bugs with the intent of later claiming credit for fixing the bugs, or to stimulate the uptake of later versions.
...never write code that intentionally breaks another programmer's code for the purpose of elevating one's status.
...never hide known obstacles to a project's completion during any phase of development, especially the design phase.
...never dishonestly downplay the difficulty of completing a project.

(c) Bharadwaj V, Dept of ECE, NUS & SC Yen, EDIC, NUS (2023)

Ethics in Computer Programming (Cont'd)...

...report any illegal activities of the employer.
...never defame the profession.
...never falsely deny the presence of bugs.
...never reveal the secret corporate knowledge of an employer.
...never accept compensation from multiple parties for the same work unless permission is given.
...never perform competitive work without the employer's knowledge.
...never conceal pertinent information from other members of the development team.
...never conceal from the employer their financial interest in development resources.
...never conceal any conflict of interest that may affect the project.
...never seek external profit from a project that was funded by a second party without permission. If permission is given to resell a product, the work should be discounted.
...never maliciously injure the reputation of an employer or members of the development team.
...never misrepresent their knowledge, experience, or abilities.
...never take credit for another's work.
...never steal software, especially development tools.
...never conceal the deficiencies of other programmers by writing code for them and allowing them to pass it off as their own work.
...never install third-party applications without the user's permission. Preferably not at all.
...stay current on the advancement of the field of Computer Science.
...never force updates on a user without their knowledge and approval.