

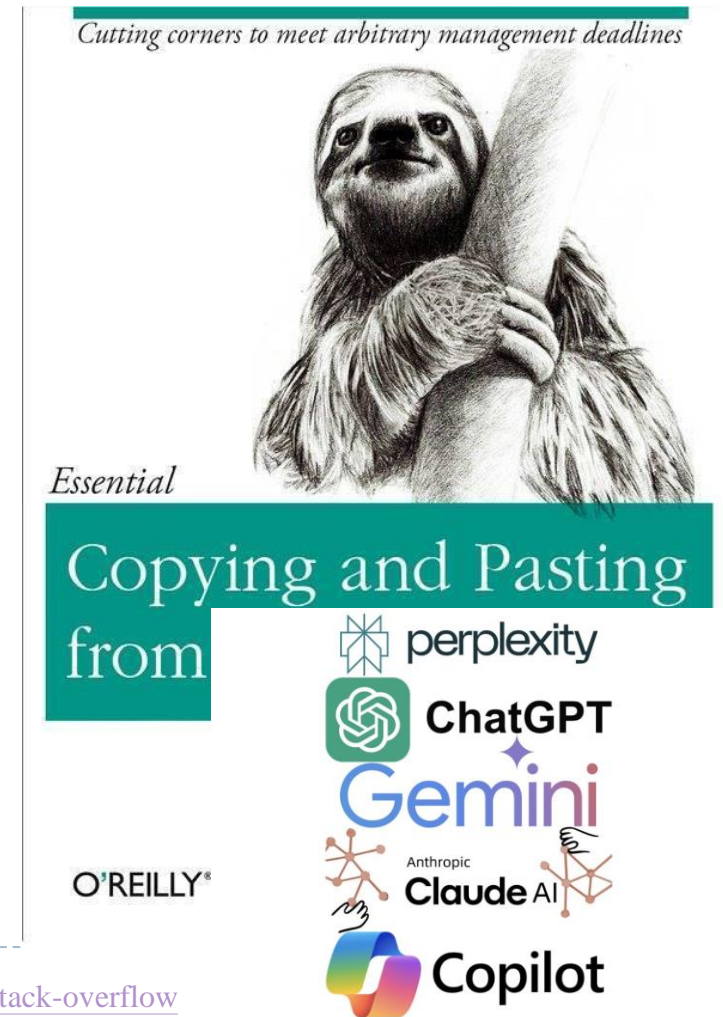
Fundamental tools for data science

Szu-Chi Chung

Department of Applied Mathematics, National Sun Yat-sen University

Motivation

- ▶ As data scientists, we know that computers are great at aiding in repetitive tasks
 - ▶ We have a vast range of tools available at our fingertips that enable us to be more productive and solve more complex problems when working on any computer-related problem
 - ▶ Yet many of us utilize only a tiny fraction of those tools; In this mini-course, I will try my best to help you become familiar with what kind of tools may be useful in your research



Course Outline

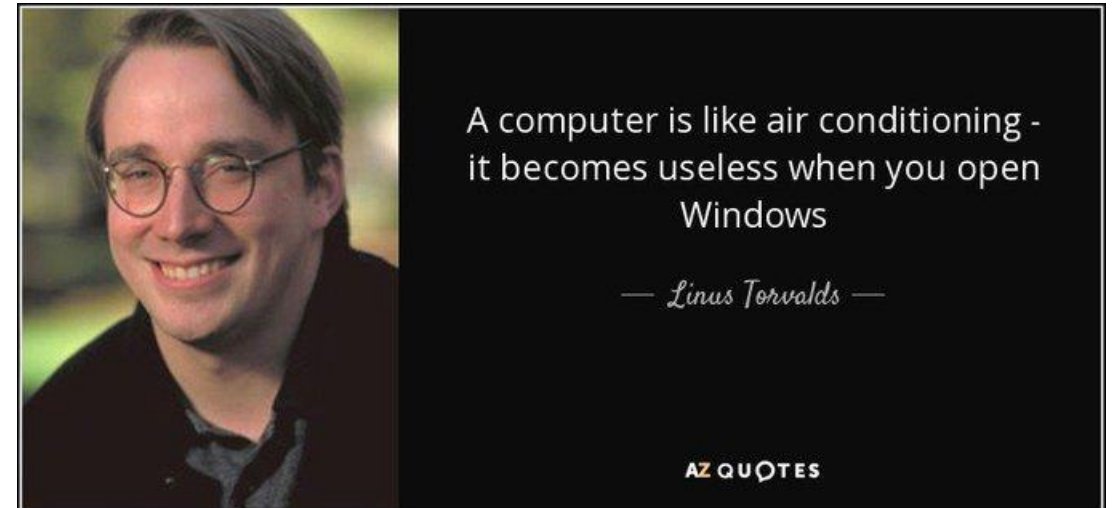
- ▶ Specifically, what we will cover includes the following topics
 - ▶ (Pre) Install Python envs/Windows WSL and Mac/Colab/GitHub/Kaggle
 - 1. Bash commands to help you be comfortable with the command line
 - ▶ <https://github.com/RehanSaeed/Bash-Cheat-Sheet>
 - ▶ https://oit.ua.edu/wp-content/uploads/2020/12/Linux_bash_cheat_sheet-1.pdf
 - 2. Basic Python with ChatGPT/Copilot/Gemini
 - 3. Colab to help you exploit the GPU power and conduct an interactive experiment
 - 4. Kaggle/GitHub search/Paperwithcode to help you explore the datasets and code from data science community



Shell

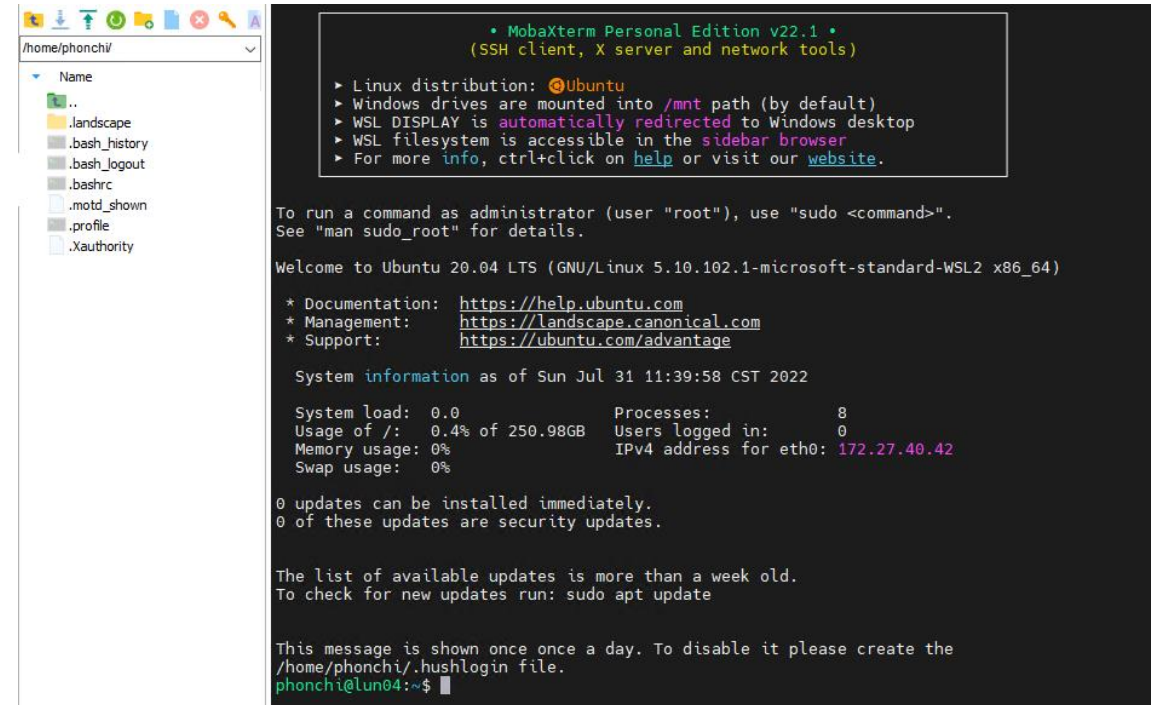
What is the shell

- ▶ Computers these days have a variety of interfaces for giving them commands
 - ▶ Fanciful graphical user interfaces, voice interfaces, and even AR/VR are everywhere
 - ▶ These are great for 80% of use-cases, but they are often fundamentally restricted in what they allow you to do — you cannot press a button that isn't there or give a voice command that hasn't been programmed. To take full advantage of the tools your computer provides, we have to go old-school and drop down to a textual interface: The Shell
- ▶ In this lecture, we will focus on the Bourne Again SHell, or “bash”
 - ▶ This is one of the most widely used shells. To open a shell prompt (where you can type commands), you first need a terminal. Your device probably shipped with one installed, or you can install it



Using the shell

- ▶ You will see a *prompt*.
 - ▶ At this prompt, you can type a command, which will then be interpreted by the shell
 - ▶ Tab can be used for auto-completing
 - ▶ The program will be searched under the *\$PATH* variable
 - ▶ *env* will list all environment variables
 - ▶ *export* can be used to set environment variables
 - ▶ *explorer.exe* . (open . in mac) to open the directory



The screenshot shows a MobaXterm window with a terminal running Ubuntu 20.04 LTS via WSL2. The left sidebar displays the file explorer for the user's home directory, listing files like .landscape, .bash_history, .bash_logout, .bashrc, .motd_shown, .profile, and .Xauthority. The terminal output includes the MobaXterm version (v22.1), system information (Ubuntu 20.04 LTS, GNU/Linux 5.10.102.1-microsoft-standard-WSL2 x86_64), and system status (0 updates available). The prompt is `phonchi@lun04:~$`.

```
• MobaXterm Personal Edition v22.1 •
(SSH client, X server and network tools)

▶ Linux distribution: Ubuntu
▶ Windows drives are mounted into /mnt path (by default)
▶ WSL DISPLAY is automatically redirected to Windows desktop
▶ WSL filesystem is accessible in the sidebar browser
▶ For more info, ctrl+click on help or visit our website.

To run a command as administrator (user "root"), use "sudo <command>".
See "man sudo_root" for details.

Welcome to Ubuntu 20.04 LTS (GNU/Linux 5.10.102.1-microsoft-standard-WSL2 x86_64)

 * Documentation:  https://help.ubuntu.com
 * Management:    https://landscape.canonical.com
 * Support:        https://ubuntu.com/advantage

System information as of Sun Jul 31 11:39:58 CST 2022

System load:  0.0          Processes:      8
Usage of /:   0.4% of 250.98GB Users logged in: 0
Memory usage: 0%          IPv4 address for eth0: 172.27.40.42
Swap usage:   0%

0 updates can be installed immediately.
0 of these updates are security updates.

The list of available updates is more than a week old.
To check for new updates run: sudo apt update

This message is shown once once a day. To disable it please create the
/home/phonchi/.hushlogin file.
phonchi@lun04:~$
```

1. Navigating Directories

- ▶ A path on the shell is a delimited list of directories, separated by / on Linux and macOS and \ on Windows. On Linux and macOS, the path / is the “root” of the file system, under which all directories and files lie, whereas on Windows there is one root for each disk partition (e.g., C:\)
 - ▶ Absolute path starts with /
 - ▶ Relative path starts with .. or .
- ▶ To see what lives in a given directory, we use the *ls* command
- ▶ Usually, running a program with the *--help* flag will print some help text that tells you what flags and options are available
 - ▶ *man* and *tldr* are also useful
 - ▶ If this is the first time you initialize the system try *sudo apt update* and *sudo apt install tldr* (or *tldr --update*)

Navigating Directories

- ▶ *ls -la* will give you more information about each file or directory present
 - ▶ First, the *d* at the beginning of the line tells us that it is a directory
 - ▶ Then follow three groups of three characters (rwx). These indicate what permissions the owner of the file, the owning group (phonchi), and everyone else respectively has on the relevant item. A - indicates that the given principal does not have the given permission
 - ▶ You can check the groups using *groups*
 - ▶ You can change the permission using *chmod*
- ▶ *cd* can change the directory
 - ▶ *pwd* prints the current directory

```
phonchi@lun04:~$ ls -la
total 32
drwxr-xr-x 3 phonchi phonchi 4096 Jul 31 11:39 .
drwxr-xr-x 3 root    root    4096 Jul 30 17:36 ..
-rw-r--r-- 1 phonchi phonchi  51 Jul 31 11:39 .Xauthority
-rw-r--r-- 1 phonchi phonchi   3 Jul 30 19:13 .bash_history
-rw-r--r-- 1 phonchi phonchi 220 Jul 30 17:36 .bash_logout
-rw-r--r-- 1 phonchi phonchi 3771 Jul 30 17:36 .bashrc
drwxr-xr-x 2 phonchi phonchi 4096 Jul 30 17:37 .landscape
-rw-r--r-- 1 phonchi phonchi   0 Jul 31 11:39 .motd shown
-rw-r--r-- 1 phonchi phonchi 807 Jul 30 17:36 .profile
```


2. File operations

- ▶ *mkdir* can be used to make new directories while *mkdir -p* can be used to create nested directories
- ▶ *touch* can be used to create files, while *cat/head/tail* can be used to print the content of a file
- ▶ *rm* can be used to remove files and *rm -rf* can be used to remove directories recursively
- ▶ *cp* can be used to copy files while *mv* can be used for moving files or renaming files
 - ▶ *cp -r* can be used to copy directories
- ▶ *ln -s* can be used to create a symbolic link

3. Connecting programs

- ▶ In the shell, programs have two primary “streams” associated with them: their input stream and their output stream
 - ▶ Normally, a program’s input and output are both your terminals. That is, your keyboard as input and your screen as output. However, we can also rewire those streams!
 - ▶ The simplest form of redirection is `> file` (Overwrite)
- ▶ You can also use `>>` to append to a file. Where this kind of input/output redirection really shines is in the use of pipes. The `|` operator lets you “chain” programs such that the output of one is the input of another
- ▶ The *xargs* command will execute a command using STDIN as arguments. For example, `ls | xargs rm` will delete the files in the current directory

4. Finding files/code

- ▶ All UNIX-like systems come packaged with *find*, a great shell tool to find files. *find* will recursively search for files matching some criteria
 - ▶ *locate* is another useful tool, but you need to install it with apt on Ubuntu
- ▶ Finding files by name is useful, but quite often, you want to search based on file content
 - ▶ A common scenario is wanting to search for all files that contain some pattern, along with where in those files said pattern occurs. To achieve this, most UNIX-like systems provide *grep*
 - ▶ *grep* "this" file.txt
 - ▶ *grep* "this" . -r

Finding shell commands

- ▶ You may want to find specific commands you typed at some point. Typing the up arrow will give you back your last command, and if you keep pressing it, you will slowly go through your shell history
 - ▶ The `history` command will let you access your shell history
 - ▶ In most shells, you can make use of `Ctrl+R` to perform the backward search
- ▶ `Ctrl+A` can move the cursor to the front while `Ctrl+E` can move the cursor to the end. `Ctrl+U` lets you clear and start from beginning!

5. Job Control

- ▶ *top/htop* will list all process
- ▶ When typing *Ctrl – C* this prompts the shell to deliver a SIGINT signal to the process and many programs can be stopped
- ▶ & suffix in command will run the command in the background
- ▶ *jobs* will list all background jobs
- ▶ *ps – aux* will list all processes
- ▶ A more generic signal for asking a process to exit gracefully is to use the *kill* command, with the syntax *kill – 9 < PID >*
- ▶ *tmux* can be used as multiplexers

6. Dotfiles

- ▶ Many programs are configured using plain-text files known as dotfiles (because the file names begin with a .
 - ▶ Shells are one example of programs configured with such files. On startup, your shell will read many files to load its configuration
 - ▶ For bash, editing your `~/.bashrc`
 - ▶ <https://github.com/Bash-it/bash-it>
- ▶ Use *source* to activate the dotfiles

Try the GUI Programs

- ▶ xclock
- ▶ xeyes
- ▶ xcalc
- ▶ gedit



Colab

Colab

- ▶ Jupyter Notebooks have rapidly grown in popularity among data scientists to become the standard for quick prototyping and exploratory analysis
 - ▶ For example, [Netflix](#) based all of their machine learning workflows on them, effectively building a whole notebook infrastructure to leverage them as a unifying layer for scheduling workflows
- ▶ Remember DO NOT store input data in your drive and load from there. The input/output is very slow (store at /content/ instead). Your output data should be stored in your google drive so that it can be accessed next time.



Kaggle

Kaggle

- ▶ Kaggle is an online community of data scientists and machine learners owned by Google, Inc
 - ▶ Kaggle allows users to find and publish data sets, explore and build models in a web-based data-science environment, work with other data scientists and machine learning engineers, and enter competitions to solve data science challenges
 - ▶ Kaggle got its start by offering machine learning competitions and now also offers a public data platform, a cloud-based workbench for data science
- ▶ You can search the competitions, datasets and notebooks using the site search on the top bar of the website
 - ▶ But you can get fine-grained control using the search in each panel from the sidebar on the left-hand

Kaggle

▶ The flexibility of selecting a script or notebook

▶ Scripts

- ▶ The first type is a script. Scripts are files that execute everything as code sequentially. To start a script, click on “Create Notebook” and select “Script”. This will open the Scripts editing interface
- ▶ From here you may select what type of script you would like to execute. You may write scripts in R or in Python.

▶ RMarkdown Scripts

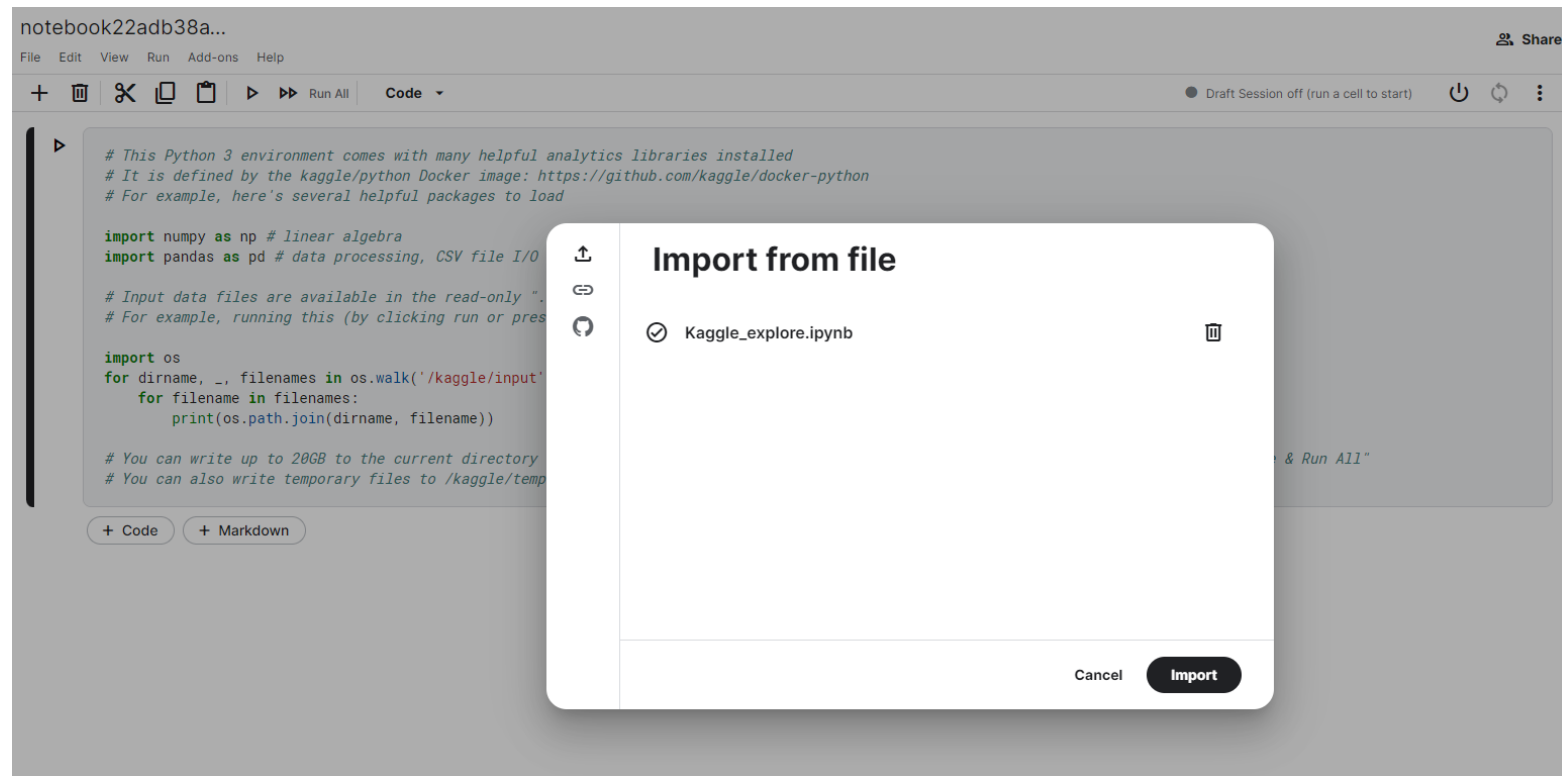
- ▶ RMarkdown scripts are a special type of script that executes not just R code, but RMarkdown code. To start editing an RMarkdown script, click on “Create Notebook”, navigate to the “Scripts” pane, and click on that. Then, in the language dropdown, click on “RMarkdown”.

▶ Notebooks

- ▶ The last type is Jupyter notebooks (usually just “notebooks”). To start a notebook, click on “Create Notebook”, and select “Notebook”. This will open the Notebooks editing interface. Notebooks may be written in either R or Python.

Kaggle kernel (notebook)

► File-> Import Notebook



The GPU accelerator

- ▶ The Resources are listed below
 - ▶ Kaggle GPU: 16G NVIDIA TESLA P100
 - ▶ <https://www.kaggle.com/docs/efficient-gpu-usage>
 - ▶ Limited to 30+ hrs/week depending on usage.
 - ▶ Limited to 12hrs/run

The screenshot shows a Kaggle Notebook interface. The top bar includes the notebook ID 'notebooke587b6bbad', a 'Draft saved' status, and buttons for 'Share', 'Save Version', and '0'. Below this is a menu bar with 'File', 'Edit', 'View', 'Run', 'Add-ons', and 'Help'. A toolbar contains icons for adding, deleting, and running code cells, along with a 'Run All' button. The main code area contains the following Python code:

```
# This Python 3 environment comes with many helpful analytics libraries installed
# It is defined by the kaggle/python Docker image: https://github.com/kaggle/docker-python
# For example, here's several helpful packages to load

import numpy as np # linear algebra
import pandas as pd # data processing, CSV file I/O (e.g. pd.read_csv)

# Input data files are available in the read-only "../input/" directory
# For example, running this (by clicking run or pressing Shift+Enter) will list all files under the input directory

import os
for dirname, _, filenames in os.walk('/kaggle/input'):
    for filename in filenames:
        print(os.path.join(dirname, filename))

# You can write up to 20GB to the current directory (/kaggle/working/) that gets preserved as output when you create a version using "Save & Run All"
# You can also write temporary files to /kaggle/temp/, but they won't be saved outside of the current session
```

At the bottom of the code area, two file paths are listed: `/kaggle/input/data-science-capstone-project-spring-2022/sample_submission.csv` and `/kaggle/input/data-science-capstone-project-spring-2022/mappings.csv`.

The right sidebar contains a 'Data' section with 'Add data' and 'data-science-capstone-project-...' under 'Input'. Below this is an 'Output' section with '/kaggle/working' and a refresh icon. Further down are 'Competitions' and 'Settings' sections. The 'Settings' section shows 'Language' set to 'Python', 'Environment' as 'Loading environment...', and 'Accelerator' set to 'GPU' (highlighted with a red box). Below 'Accelerator', 'GPU Quota' is shown as '09:20 / 43 hrs' and 'Internet' is checked.

Save the notebook

► Quick Save

- Skips the top-to-bottom notebook execution and just takes a snapshot of your notebook exactly as it's displayed in the editor. This is a great option for taking a bunch of versions while you're still actively experimenting. You can choose to reserve the output

Save Version

Version Name (optional):

Version 1

9 / 50

✓ Quick Save
Save a version of your notebook the way it currently looks

Advanced Settings Cancel Save

← Version Settings

Save outputs when creating a Quick Save

☐ Never save output

☒ Always save output

☐ Save output for this version

Save & Run All with an accelerator

ACCELERATOR

Run with GPU for all sessions

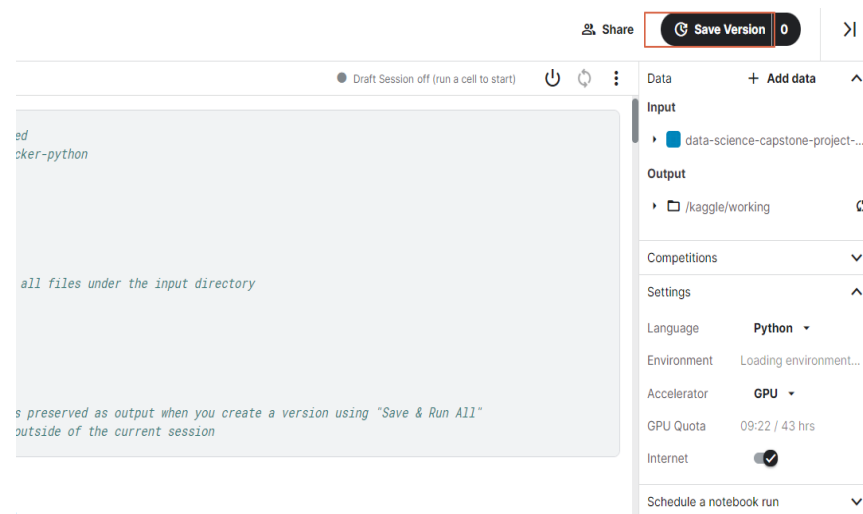
Cancel Save

► Save & Run All

- Creates a new session with a completely clean state and runs your notebook from top to bottom. In order to save successfully, the entire notebook must execute within 12 hours (9 hours for TPU notebooks). Save & Run All is identical to the “Commit” behavior.

Running the code in the background

- ▶ You can also run the code in the background with Kaggle. Firstly, make sure your code is bug-free, as any error in any code block would result in early stopping. Click the “Save Version” button as follows:
 - ▶ *The concept of “Versions” is a collection consisting of a Notebook version, the output it generates, and the associated metadata about the environment.*



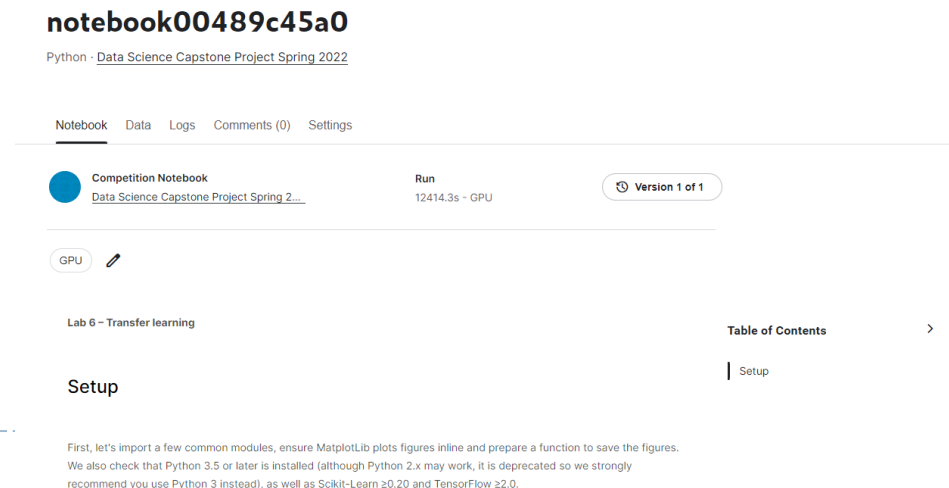
- ▶ Notice that the output is limited to 20G and the max run time is limited to 12hrs

Running the code in the background

- ▶ You can then see the results by clicking the version number:



- ▶ The log will be shown in the Notebook panel



Running the code in the background

- ▶ Your output can be accessed through the “Data” panel, where you can download your data. You can also save your model as a new dataset and import your new dataset into the notebook via “Add data” so that you can modify your code to load your checkpoint:

The screenshot displays the 'Data' panel of a notebook interface. At the top, the notebook is titled 'notebook00489c45a0' with a subtitle 'Python · Data Science Capstone Project Spring 2022'. Below this is a navigation bar with tabs for 'Notebook', 'Data', 'Logs', 'Comments (0)', and 'Settings'. The 'Data' tab is active, showing a list of datasets. The first dataset is 'ev2_fine_tuning_0319.keras' (56.21 MB), which has a 'Submit' button and a download icon. Below the dataset name, there is a message 'Unable to show preview' with a laptop icon and the text 'Previews for binary data are not supported'. To the right of the dataset list, there is a section for 'Input (12.87 MB)' showing 'Data Sources' with a link to 'Data Science Capstone ...'. Below this is an 'Output' section listing 'ev2_fine_tuning_0319.keras' and 'my_keras_model_0319.keras'. At the bottom right, there are buttons for 'Download All', '+ New Dataset', and '+ New Notebook'.



GitHub/Paperwithcode/

Search tips

- ▶ https://www.google.com/advanced_search
 - ▶ <https://www.google.com/search?q=resume+site:cs.cmu.edu+filetype:pdf>
- ▶ <https://github.com/search/>
 - ▶ <https://github.com/search?o=desc&q=dotfiles&s=stars&type=Repositories>
 - ▶ <https://github.com/search?q=GAN>
 - ▶ <https://github.com/search?q=statement+of+purpose>
- ▶ <https://paperswithcode.com/>



Appendix

How to find an interesting fork/Deleted repository?

- ▶ <https://stackoverflow.com/questions/54868988/how-to-determine-which-forks-on-github-are-ahead>
- ▶ <https://softwareheritage.org/>
- ▶ <https://gist.github.com/rjeczalik/81ff08b59d7841970fca82ca39f40a10>

Reference

- ▶ <https://missing.csail.mit.edu/>
- ▶ GUI
 - ▶ <https://docs.microsoft.com/en-us/windows/wsl/setup/environment>
 - ▶ <https://www.thewindowsclub.com/error-0x80370102-the-virtual-machine-could-not-be-started>
 - ▶ HyperV
- ▶ GPU
 - ▶ <https://docs.nvidia.com/cuda/wsl-user-guide/index.html#getting-started-with-cuda-on-wsl>
 - ▶ 需安裝21H2
 - ▶ <https://stackoverflow.com/questions/70011494/why-does-nvidia-smi-return-gpu-access-blocked-by-the-operating-system-in-wsl2>