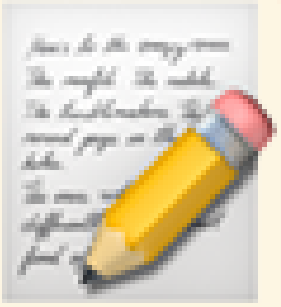
06/26 BRIEFING



DATASET ANALYSIS

Third-Year @ Dept. ATM

Group1 109601003 林群賀

THE PROBLEMS WE HAVE MET

- Cannot download the data
- The process of unzip is too slow
- How to analyze the source dataset

HOW TO DOWNLOAD DATASET?

```
import requests

def download_file(url, save_path):
    response = requests.get(url, stream=True)
    response.raise_for_status()
    with open(save_path, 'wb') as file:
        for chunk in response.iter_content(chunk_size=8192):
            if chunk:
                file.write(chunk)
    print("Download Successfully!!!")

url = "https://ncu365-my.sharepoint.com/personal/ckchang_office
save_path = "path/to/save/location/file.zip"

download_file(url, save_path)
```

HOW TO UNZIP DATASET?

```
import zipfile

def unzip_file(zip_path, extract_path):
    with zipfile.ZipFile(zip_path, 'r') as zip_ref:
        zip_ref.extractall(extract_path)
    print("Unzip Successfully!!!")

zip_path = '/Volumes/HoHo\'s SSD/GS4524/challenge-2018-1.0.0.phy
extract_path = '/Volumes/HoHo\'s SSD/GS4524/'

unzip_file(zip_path, extract_path)
```

THE SOURCE DATASET TOO MESSY.

CLEANUP THE SOURCE CODE

DIVIDE THE DATASET INTO SMALL ONE

286 GB -> 56 GB

THE TOOLS THAT WE CAN TRAIN THE MODEL

- RTX 3080
- Kaggle
  - afordable for **100 GB** dataset
  - need to be **public**

WHAT I HAVE LEARNED YESTERDAY?

- How to use **Web Crawler** wisely
- Analyze the source code
- Divide the huge dataset

WHAT I WANT TO SOLVE TODAY?

- Visualize the source dataset with the materials supplied by Teacher

REFERENCE

- Deep learning for automated sleep staging using instantaneous heart rate
- AUPRC vs. AUC-ROC? [duplicate]
- Index of /physiobank/database/challenge/2018/