

Github introduction

1. Create a Github.com account
2. Add your account to [google sheet](#)

UNIX practice

Go into your bigdata folder. If you have not used the cluster before then you will be in the `gen220` project. Or you may have your own lab bigdata folder.

```
cd ~/bigdata # this should work but if it doesn't
cd /bigdata/gen220/$USER # will go into your bigdata folder for the class
# if the above doesn't work you are likely already in a lab group on HPCC
cd /bigdata/$GROUP/$USER # should work since $USER is your login and $GROUP is your primary
```

Now you want to make a folder for your work for this class

```
# you can make a folder for GEN220
mkdir gen220
go into that folder
cd gen220
# now use git to checkout the class data folder
git clone https://github.com/biodataprogram/GEN220_data.git
# now go into this folder
cd GEN220_data
```

Look around in the folder. Go into the `tabular` folder where I've stored some tab or comma delimited data. You will later need to copy a file from this folder into your homework folder.

For your homework:

1. Accept the homework 1 problem - (see link in Canvas).
2. Create a folder to work in on HPCC (or your own computer): `mkdir -p ~/bigdata/gen220/homework` and then `cd ~/bigdata/gen220/homework`
3. Checkout the GEN220 data folder `git clone https://github.com/biodataprogram/GEN220_data.git`
4. Checkout the homework 1 github repository created in step 2. (if you [setup SSH keys](#) in github)

```
git clone git@github.com:biodataprogram/2023_hw1-YOURGITHUBID
```

OR for the https will need to [create a token as your password](#))

```
git clone https://github.com/biodataprogram/2023_hw1-YOURGITHUBID.git
```

6. Go into your folder (`cd 2023_hw1-YOURGITHUBID`).
7. Create a script in there called `filesize.sh`
8. Add some code to this script which achieve the directions at the bottom of this

9. To submit your homework (and you can do this more than once, you don't need to wait till you are sure you are done, this is just saving the file.
10. in the 2023_hw1-YOURGITHUBID folder, say you want to save your results back to github

```
# this step saves a version of the code
git commit -m "This is a homework 1 solution" filesize.sh
# this step will push the data from HPCC or your computer UP to the github site
# this step will request your username (YOURGITHUBID) and your password (that TOKEN I mentioned)
# if you have setup github account with SSH keys then it will ask you for your SSH key password
git push
```

9. you can repeat

Tasks for Homework 1

1. copy the `threatened-species.csv.gz` file
2. Write a script called `filesize.sh`, that script should do the following:
 - print out the size of the `threatened-species.csv.gz` using `du` or `ls -l`
 - Uncompress the file with `gunzip` while leaving the original alone by adding the `'-k'` option (`gunzip -k threatened-species.csv.gz`), add a `-f` option so it will overwrite the new file as well in case you run this more than one time.
 - Print out the size of the new uncompressed file with `du` or `ls -l`
 - print out the number lines in the uncompressed file size
 - BONUS: if you can, print out the compression ratio (compressed filesize / uncompressed size), you can do this if you capture the sizes in two variables and run this - requires using `cut/awk` to capture the file sizes.

```
python -c "print($COMPRESSED / $UNCOMPRESSED)"
```