DAP03: Big Data Analytics
Summer Semester 2020

Due to the unfortunate impacts of the NCOV-19 virus, we will be teaching the BigData course as an online-only module.

To help prepare for this, we have set up some important teaching resources, and converted the labs to be easier to work on in a remote setting. Logistics can be broken down into two sections:

1. Class Meetings

Class lectures will take place via BigBlueButton (https://bigbluebutton.org/). We will have two sessions per day – one at 9:15, and one at 13:30. These will likely run between 1.5-2 hours each, and will consist of talking through a lab exercise as a group. **Feel free to ask questions during this period! If you have a question, likely several other people do too.**

After each lecture/lab period, we will be available via Email/Chat during the rest of the morning/afternoon to answer more specific questions. If you feel confident in the material we have gone over in a session, feel free to start working on the homework assignments for that lab.

Homework assignments are necessary (and mandatory) for course completion, but no letter grade is given. It is expected that you receive at least 80% of the available points. Only if you successfully pass the homework assignments, you can participate in the final project. These are exercises are meant to help you get a deeper understanding of the material, and to prepare you for the final projects.

The homeworks will be due on May 13th, and there will be a review session on the morning of May 15th where we will go over our solutions to the homework problems, and have time for questions.

2. Lab Exercises

In this class, we will need to use and process large amounts of data. While all of the exercises are doable on a normal laptop – indeed, one of the goals of the course is to learn proper data and memory management strategies – we realize that not everyone has a powerful PC to rely on. To help with this, we will use **Jupyter Notebooks** to run the lab exercises, backed up by the PCPool on Campus Golm. In essence, you will be able to edit your code and see results on your personal laptop, while the data will be stored and processed on one of the Lab PCs. **You will be given a personal login to one of the lab computers – this account is for your use throughout the BigData course**.

If you are more comfortable working on your own laptop, you can also of course run all exercises via Spyder, or via a Jupyter notebook running on your own laptop.

A. Steps for using Jupyter on Your Own Laptop

If you want to use your own laptop, first you will need to install several modules. This can be done via *conda.* On the course Moodle, there is a .yml file with the conda environment used on the lab computers. You can also run these commands:

```
conda create -n BigData python=3.7
conda activate BigData
conda install gdal rasterio matplotlib numpy dask dask-ml xarray pytables pandas geopandas shapely
    requests cartopy scikit-learn pysal h5py statsmodels netcdf4 pyproj spyder scikit-learn jupyter
    descartes
conda install -c conda-forge earthengine-api
conda install -c conda-forge pyhdf
```

Once you have your conda environment set up, simply open your conda prompt and run:

```
conda activate BigData
jupyter notebook --generate-config
jupyter notebook password
```

These steps configure the notebook for us, and let us set a password. We will need this when we start up our Jupyter notebook on our laptops.

```
jupyter notebook --no-browser --port=8888
```

Then, open your web browser and go to:

localhost:8888

You can then download the required data from Moodle, and run the code on your own PC.

Alternatively, you can simply open Spyder and follow along on the Lab PDFs from Moodle. You would still need to have all required modules installed, as listed above.

B. Steps for using Jupyter on the Remote Logins

To use the remote servers for more heavy computation, you need to follow a few extra steps.

On your own PC, you do not need to have any programs installed -- we will use Jupyter over SSH.

Important Note: *If you are on a Linux or Mac PC, you should be able to use ssh without installing any other packages. If you are on a* ***Windows*** *PC, you will need to install some extra software to set this up. We recommend either (1) Git (https://gitforwindows.org/), (2) PuTTY (https://www.putty.org/), or (3) Cygwin (https://www.cygwin.com/). Any of these three will work – all you need to do is be able to run the 'ssh' commands listed below.*

To start a Jupyter notebook on your remote PC, you need to log in via SSH:

```
ssh my_username@my_pc_ip_address
```

Once logged in, you should start a 'screen' session, which keeps your Jupyter notebook alive even if you lose your internet connection.

```
screen
```

Next, to start your Jupyter notebook running, we need to decide on which port to send the data out of. **Important: Use the Four-Digit Port number (e.g., 8888) that was assigned to you along with your PC IP address, username and password.** If you accidentally use the same port as someone else on the same PC, you would be trying to edit each other's labs!

Before we run the Jupyter notebook, we need to do three steps:

```
conda activate
jupyter notebook --generate-config
jupyter notebook password
```

Then, you need to 'cd' into the directory where your data is stored, e.g.,

```
cd /DATA/my_username/BigData_Labs/
```

Make sure that you have your own directory set up on the /DATA drive! Otherwise you will overwrite each other's Jupyter notebooks!

These steps configure the notebook for us, and let us set a password. We will need this when we start up our Jupyter notebook on our laptops.

```
jupyter notebook --no-browser --port=XXXX
```

Replace **XXXX** with your assigned port number. Now you are all set on the server side!

The next step is to tell your computer how to access the Jupyter notebook you started on the remote server. We do this via *port forwarding*.

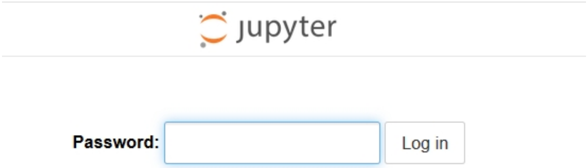
Once again, this is a simple one-line command:

```
ssh -N -f -L localhost:YYYY:localhost:XXXX my_username@my_pc_ip_address
```

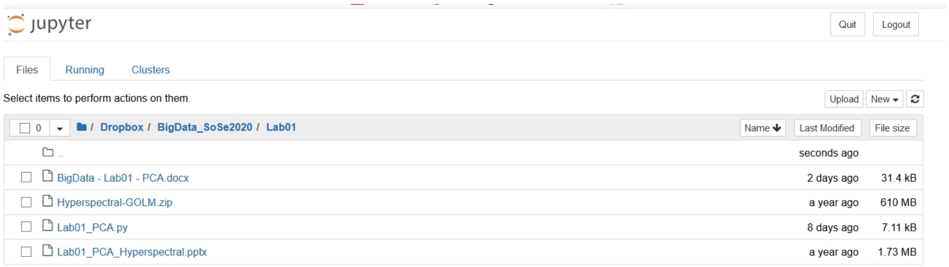
Here, **XXXX** is the same port you used in the command above, and **YYYY** is the port you want your laptop to 'listen' on. This command won't show any output – you can check it is working by opening up your web browser, and navigating to:

localhost:YYYY

You should see something like this:



Once you enter your password, you should see the files on your PC, e.g., something like this:



Select items to perform actions on them.

/ Dropbox / BigData_SoSe2020 / Lab01

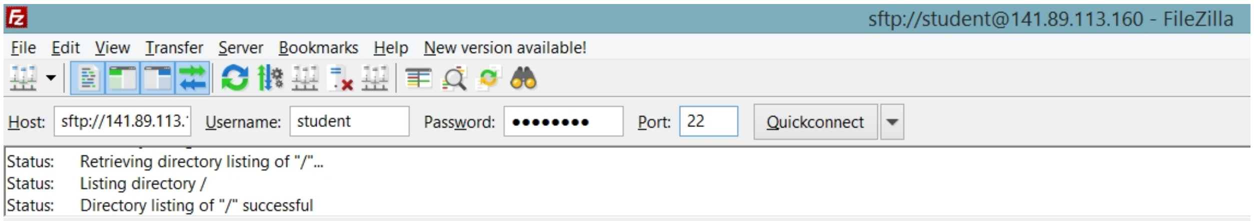
Name	Last Modified	File size
..	seconds ago	
BigData - Lab01 - PCA.docx	2 days ago	31.4 kB
Hyperspectral-GOLM.zip	a year ago	610 MB
Lab01_PCA.py	8 days ago	7.11 kB
Lab01_PCA_Hyperspectral.pptx	a year ago	1.73 MB

You should now be able to access the Lab exercises directly from your browser, while using the computing power of the PC pools.

Data Access (upload/download) for the Lab PCs

At some point you will want to save your own data, and transfer some figures and analysis back to your own laptop, either for some more analysis or to add to your lab assignments. To do this, you will need to use SFTP (https://en.wikipedia.org/wiki/SSH_File_Transfer_Protocol). We recommend using Filezilla (https://filezilla-project.org/), which will work on all computer platforms.

You can use the IP address, username, and password given to you to login to your remote PC.

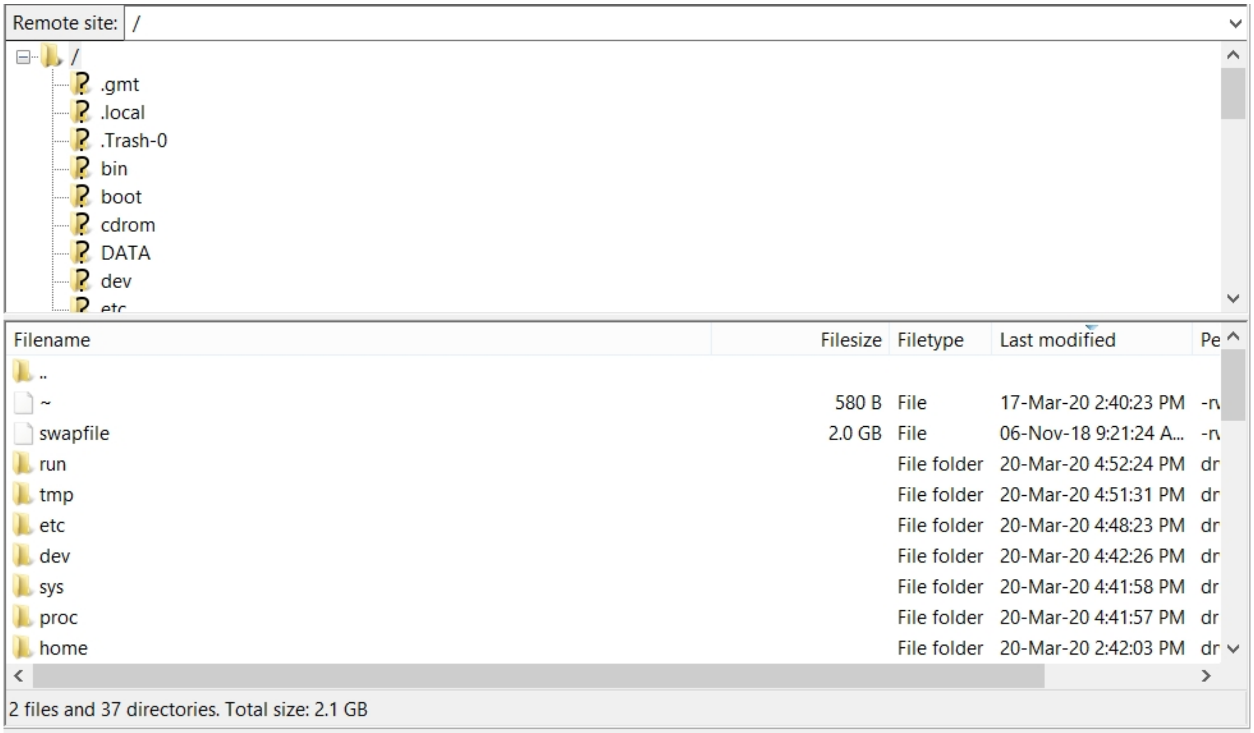


sftp://student@141.89.113.160 - FileZilla

Host: sftp://141.89.113.' Username: student Password: •••••••• Port: 22

```
Status:     Retrieving directory listing of "/"...
Status:     Listing directory /
Status:     Directory listing of "/" successful
```

You can then see all of the files on the remote system, and copy data back and forth:



Remote site: /

```
/
    .gmt
    .local
    .Trash-0
    bin
    boot
    cdrom
    DATA
    dev
    etc
```

Filename	Filesize	Filetype	Last modified	Pe
..				
~	580 B	File	17-Mar-20 2:40:23 PM	-rv
swapfile	2.0 GB	File	06-Nov-18 9:21:24 A...	-rv
run		File folder	20-Mar-20 4:52:24 PM	dr
tmp		File folder	20-Mar-20 4:51:31 PM	dr
etc		File folder	20-Mar-20 4:48:23 PM	dr
dev		File folder	20-Mar-20 4:42:26 PM	dr
sys		File folder	20-Mar-20 4:41:58 PM	dr
proc		File folder	20-Mar-20 4:41:57 PM	dr
home		File folder	20-Mar-20 2:42:03 PM	dr

2 files and 37 directories. Total size: 2.1 GB