

# **Scalable and Computationally Reproducible Approaches to Arctic Research**

Matt Jones, Bryce Mecum, Jeanette Clark, Sam Csik

September 19, 2022

# Table of contents

<b>Preface</b>	<b>6</b>
About . . . . .	6
Schedule . . . . .	6
Code of Conduct . . . . .	6
Setting Up . . . . .	8
Download VS Code and Remote - SSH Extension	8
Log in to the server . . . . .	8
Install extensions on the server . . . . .	9
Test your local setup (Optional) . . . . .	10
About this book . . . . .	12
<b>1 Welcome and Introductions</b>	<b>13</b>
<b>2 Remote Computing</b>	<b>16</b>
2.1 Introduction . . . . .	16
2.2 Servers & Networking . . . . .	16
2.3 IP addressing . . . . .	18
2.4 Bash Shell Programming . . . . .	18
2.5 Connecting to a remote computer via a shell . . . . .	20
2.6 Exercise 1: Connect to a server & practice bash commands . . . . .	20
2.6.1 Launch your Terminal Program . . . . .	20
2.6.2 Connect to a remote server . . . . .	21
2.6.3 Practice . . . . .	21
2.7 Exercise 2: Write a bash shell script . . . . .	23
<b>3 Python Programming on Clusters</b>	<b>26</b>
3.1 Introduction . . . . .	26
3.2 Starting a project . . . . .	26
3.3 Virtual Environments . . . . .	27
3.4 Brief overview of python syntax . . . . .	30
3.5 Jupyter notebooks . . . . .	34
3.5.1 Load libraries . . . . .	34

3.5.2	Read in a csv . . . . .	35
3.6	Functions . . . . .	40
<b>4</b>	<b>Pleasingly Parallel Programming</b>	<b>43</b>
4.1	Introduction . . . . .	43
4.2	Why parallelism? . . . . .	45
4.3	Processors (CPUs) and Cores . . . . .	45
4.4	Modes of parallelization . . . . .	47
4.5	Task parallelism with <code>concurrent.futures</code> . . . . .	48
4.6	Approaches to parallelization . . . . .	49
4.7	<code>concurrent.futures</code> . . . . .	49
4.8	<code>parsl</code> . . . . .	49
4.9	When to parallelize . . . . .	50
<b>5</b>	<b>Parallel Pitfalls and their solutions</b>	<b>52</b>
5.1	Summary . . . . .	52
5.2	Further Reading . . . . .	52
<b>6</b>	<b>Documenting and Publishing Data</b>	<b>53</b>
6.1	Introduction . . . . .	53
<b>7</b>	<b>Group Project: Staging and Preprocessing</b>	<b>54</b>
7.1	Introduction . . . . .	54
7.2	Staging and Tiling . . . . .	56
<b>8</b>	<b>Software Design I</b>	<b>58</b>
<b>9</b>	<b>Data Structures and Formats for Large Data</b>	<b>59</b>
9.1	Learning Objectives . . . . .	59
9.2	Introduction . . . . .	59
9.3	Working with Large Data . . . . .	60
9.4	NetCDF/HDF Overview . . . . .	60
9.5	Introduction to Xarray . . . . .	62
9.6	Exercise . . . . .	62
<b>10</b>	<b>Parallelization with Dask</b>	<b>63</b>
10.1	Learning Objectives . . . . .	63
10.2	Introduction . . . . .	63
10.2.1	Notes . . . . .	63
10.3	Dask Tutorial . . . . .	64
10.4	Conclusion . . . . .	64

<b>11 Spatial and Image Data Using GeoPandas</b>	<b>65</b>
11.1 Introduction . . . . .	65
11.2 Pre-processing raster data . . . . .	66
11.2.1 Check extents . . . . .	70
11.3 Calculate total distance per fishing area . . . . .	72
<b>12 Data Futures: Parquet and Arrow</b>	<b>78</b>
12.1 Introduction . . . . .	78
12.2 Row major vs column major . . . . .	79
12.2.1 Row major versus column major files . . .	79
12.3 Parquet . . . . .	80
12.4 Arrow . . . . .	81
12.5 Example . . . . .	82
<b>13 Software Design II</b>	<b>85</b>
<b>14 Group Project: Data Processing</b>	<b>86</b>
<b>15 Data Ethics</b>	<b>87</b>
<b>16 Google Earth Engine</b>	<b>88</b>
16.1 Introduction (15-20min) . . . . .	88
16.2 Getting started with Google Earth Engine (GEE) on your own machine (40 min) . . . . .	89
16.3 Visualize global precipitation data using Google Earth Engine . . . . .	91
16.4 INGMAR'S DEMONSTRATION HERE?(30-40 min) . . . . .	93
16.5 Conclusion/Summary . . . . .	93
16.6 Other Resources . . . . .	94
<b>17 Group Project: Visualization</b>	<b>95</b>
<b>18 Workflows for data staging and publishing</b>	<b>96</b>
18.1 NSF policy for large datasets . . . . .	96
18.2 Data transfer tools . . . . .	97
18.3 Documenting large datasets . . . . .	99
<b>19 What is Cloud Computing Anyways?</b>	<b>100</b>
<b>20 Reproducibility and Containers</b>	<b>101</b>
20.1 Outline . . . . .	101

20.2 Hands-off Demo . . . . .	102
-------------------------------	-----

# Preface

## About

This 5-day in-person workshop will provide researchers with an introduction to advanced topics in computationally reproducible research in python and R, including software and techniques for working with very large datasets. This includes working in cloud computing environments, docker containers, and parallel processing using tools like parsl and dask. The workshop will also cover concrete methods for documenting and uploading data to the Arctic Data Center, advanced approaches to tracking data provenance, responsible research and data management practices including data sovereignty and the CARE principles, and ethical concerns with data-intensive modeling and analysis.



## Schedule

### Code of Conduct

Please note that by participating in this activity you agree to abide by the [NCEAS Code of Conduct](#).

	<b>Monday</b>	<b>Tuesday</b>	<b>Wednesday</b>	<b>Thursday</b>	<b>Friday</b>
08:00-08:30	Coffee (optional)	Coffee (optional)	Coffee (optional)	Coffee (optional)	Coffee (optional)
08:30-09:00	<b>1. Welcome and Course Overview</b> (Matt)				
09:00-09:30		<b>6. Group project I</b> Data staging and pre-processing (Jeanette)	<b>10. Spatial and Image Data using GeoPandas</b> (Jeanette)		
09:30-10:00	<b>2. Remote computing</b> (Sam)			<b>15. Google Earth Engine</b> (Ingmar, Sam)	<b>19. What is cloud computing anyways?</b> (Matt)
10:00-10:30			<b>11. Data futures: Parquet and Arrow</b> (Jeanette)		
10:30-11:00	BREAK	BREAK	BREAK	BREAK	BREAK
11:00-11:30	<b>3. Python programming on clusters</b> (Jeanette)	<b>7. Software design I</b> (Matt)	<b>12. Software Design II</b> (Carmen)	<b>16. Billions of Ice Wedge Polygons</b> (Chandi)	<b>20. Reproducibility redux via containers</b> (Matt) <b>Survey Feedback Q &amp; A</b>
11:30-12:00					
12:00-12:30	Lunch	Lunch	Lunch	Lunch	
12:30-13:00					Adjourn
13:00-13:30					
13:30-14:00	<b>4. Pleasingly Parallel Programming</b> (Matt)	<b>8. Data structures and formats for large data</b> (Carmen)	<b>13. Group project II Parallel data processing</b> (Jeanette)	<b>17. Group project III Visualizing big geospatial data</b> (Jeanette)	
14:00-14:30					
14:30-15:00					
15:00-15:30	Break	Break	Break	Break	
15:30-16:00	<b>5. Documenting and Publishing Data</b> (Daphne)	<b>9. Parallelization with Dask</b> (Carmen)	<b>14. Data Ethics</b> (Tash)	<b>18. Workflows for data staging and publishing</b> (Jeanette)	
16:00-16:30			Breather Catch-up		
16:30-17:00	Q&A	Q&A	Q&A	Q&A	

# Setting Up

In this course, we will be using Python (> 3.0) as our primary language, and VS Code as our IDE. Below are instructions on how to get VS Code set up to work for the course. If you are already a regular Python user, you may already have another IDE set up. We strongly encourage you to set up VS Code with us, because we will use your local VS Code instance to write and execute code on one of the NCEAS servers.

## Download VS Code and Remote - SSH Extension

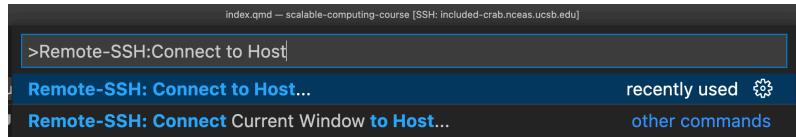
First, [download VS Code](#) if you do not already have it installed.

You'll also need to download the [Remote - SSH extension](#).

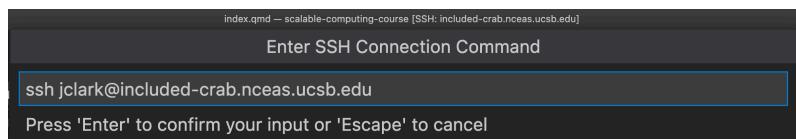
## Log in to the server

To connect to the server using VS Code follow these steps, from the VS Code window:

- open the command palette (Cmd + Shift + P)
- enter “Remote SSH: Connect to Host”

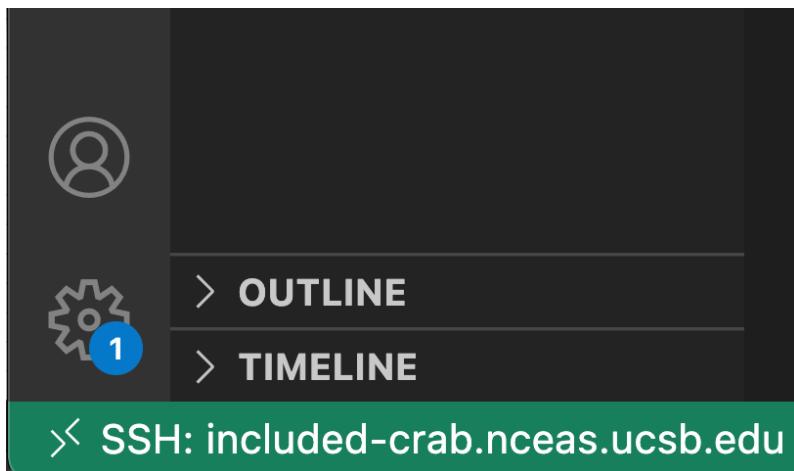


- select “Add New SSH Host”
- enter the ssh command to connect to the host as if in a terminal (`ssh username@included-crab.nceas.ucsb.edu`)
  - Note: you will only need to do this step once



- select the SSH config file to update with the name of the host. You should select the one in your user directory (eg: /Users/jclark/.ssh/config)
- click “Connect” in the popup in the lower right hand corner
  - Note: If the dialog box does not appear, reopen the command palette (Cmd + Shift + P), type in “Remote-SSH: Connect to Host...”, choose included-crab.nceas.ucsb.edu from the options of configured SSH hosts, then enter your password into the dialog box that appears
- enter your password in the dialog box that pops up

When you are connected, you will see in the lower left hand corner of the window a green bar that says “SSH: included-crab.nceas.ucsb.edu.”



## **Install extensions on the server**

After connecting to the server, in the extensions pane (View > Extensions) search for, and install, the following extensions:

- Python
- Jupyter
- Jupyter Keymap
- Pylance

Note that these extensions will be installed on the server, and not locally.

## Test your local setup (Optional)

We are going to be working on the server exclusively, but if you are interested in setting up VS Code to work for you locally with Python, you can follow these instructions. This local setup section summarizes the official VS Code tutorial. For more detailed instructions and screenshots, see the [source material](#). If you already use VS Code for Python you can skip this.

Locally (not connected to the server), check to make sure you have Python installed if you aren't sure you do. File > New Window will open up a new VS Code window locally.

To check your python, from the terminal run:

```
python3 --version
```

If you get an error, it means you need to install Python. Here are instructions for getting installed, depending on your operating system. Note: There are many ways to install and manage your Python installations, and advantages and drawbacks to each. If you are unsure about how to proceed, feel free to reach out to the instructor team for guidance.

- Windows: Download and run an installer from [Python.org](#).
- Mac: Install using [homebrew](#). If you don't have homebrew installed, follow the instructions from their webpage.
  - `brew install python3`

After you run your install, make sure you check that the install is on your system PATH by running `python3 --version` again.

Next, install the [Python extension for VS Code](#).

Open a terminal window in VS Code from the Terminal drop down in the main window. Run the following commands to

initialize a project workspace in a directory called `training`. This example will show you how to do this locally. Later, we will show you how to set it up on the remote server with only one additional step.

```
mkdir training  
cd training  
code .
```

Next, select the Python interpreter for the project. Open the **Command Palette** using Command + Shift + P (Control + Shift + P for windows). The Command Palette is a handy tool in VS Code that allows you to quickly find commands to VS Code, like editor commands, file edit and open commands, settings, etc. In the Command Palette, type “Python: Select Interpreter.” Push return to select the command, and then select the interpreter you want to use (your Python 3.X installation).

To make sure you can write and execute code in your project, [create a Hello World test file](#).

- From the File Explorer toolbar, or using the terminal, create a file called `hello.py`
- Add some test code to the file, and save

```
msg = "Hello World"  
print(msg)
```

- Execute the script using either the Play button in the upper-right hand side of your window, or by running `python3 hello.py` in the terminal.
  - For more ways to run code in VS Code, see the [tutorial](#)

Finally, to test Jupyter, download the [Jupyter extension](#). You’ll also need to install `ipykernel`. From the terminal, run `pip install ipykernel`.

You can create a test Jupyter Notebook document from the command palette by typing “Create: New Jupyter Notebook” and selecting the command. This will open up a code editor pane with a notebook that you can test.

## About this book

These written materials reflect the continuous development of learning materials at the Arctic Data Center and NCEAS to support individuals to understand, adopt, and apply ethical open science practices. In bringing these materials together we recognize that many individuals have contributed to their development. The primary authors are listed alphabetically in the citation below, with additional contributors recognized for their role in developing previous iterations of these or similar materials.

This work is licensed under a [Creative Commons Attribution 4.0 International License](#).

**Citation:** Matthew B. Jones, Bryce Mecum, S. Jeanette Clark, Samantha Csik. 2022. Scalable and Computationally Reproducible Approaches to Arctic Research.

**Additional contributors:** Amber E. Budden, Natasha Haycock-Chavez, Noor Johnson, Stephanie Hampton, Jim Regetz, Bryce Mecum, Julien Brun, Julie Lowndes, Erin McLean, Andrew Barrett, David LeBauer, Jessica Guo.

This is a Quarto book. To learn more about Quarto books visit <https://quarto.org/docs/books>.

# 1 Welcome and Introductions



This course is one of three that we are currently offering, covering fundamentals of open data sharing, reproducible research, ethical data use and reuse, and scalable computing for reusing large data sets.





# **Reproducible Practices for Arctic Research Using R**

**February 14th - February  
18th, 2022**

*This course will be taught  
virtually*



# **Scalable and Computationally Reproducible Approaches to Arctic Research**

**March 21st - March 25th**  
*Located at the National Center  
for Ecological Analysis and  
Synthesis in Santa Barbara, CA*

# 2 Remote Computing

- Understand the basic architecture of computer networks
- Learn how to connect to a remote computer via a shell
- Become familiarized with Bash Shell programming to navigate your computer's file system, manipulate files and directories, and automate processes

## 2.1 Introduction

Scientific synthesis and our ability to effectively and efficiently work with big data depends on the use of computers & the internet. Working on a personal computer may be sufficient for many tasks, but as data get larger and analyses more computationally intensive, scientists often find themselves needing more computing resources than they have available locally. Remote computing, or the process of connecting to a computer(s) in another location via a network link is becoming more and more common in overcoming big data challenges.

In this lesson, we'll learn about the architecture of computer networks and explore some of the different remote computing configurations that you may encounter, we'll learn how to securely connect to a remote computer via a shell, and we'll become familiarized with using Bash Shell to efficiently manipulate files and directories. We will begin working in the [VS Code](#) IDE (integrated development environment), which is a versatile code editor that supports many different languages.

## 2.2 Servers & Networking

Host computers connect via networking equipment and can send messages to each other over communication protocols (aka

an **Internet Protocol**, or IP). Host computers can take the role of **client** or **server**, though these are not inherent properties of a host (i.e. the same machine can play either role).

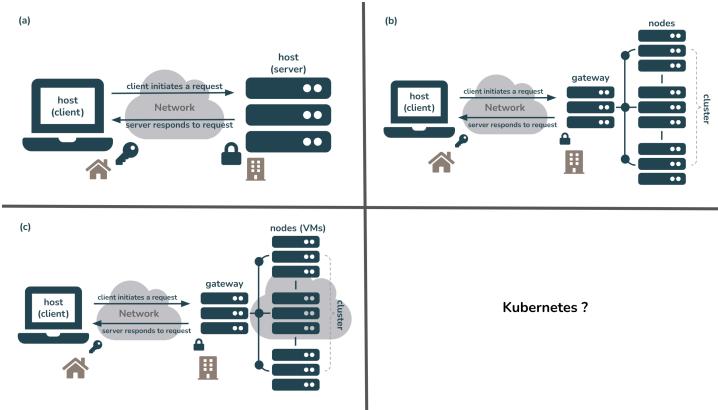
- **Client:** the host *initiating* a request
- **Server:** the host *responding* to a request

**i Note**

Hosts typically have one network address but can have many different ones (for example, adding multiple network cards to a single server increases bandwidth).

**! Important**

GET FEEDBACK ON VISUAL + ADD KUBERNETES CONFIGURATION VISUAL; MAKE LARGER?



**Fig 1.** Different remote computing configurations. (a) A client uses secure shell protocol (SSH) to login/connect to a server over the internet. The client and the server exist in the physical world, but in different locations. (b) A client uses SSH to login/connect to a computing cluster (i.e. a set of computers (nodes) that work together so that they can be viewed as a single system) over the internet. The connection is first made through a gateway node (i.e. a computer that routes traffic from one network to another). The client and the cluster (server) exist in the physical world, but in different locations. (c) A client uses SSH to login/connect to a computing cluster where each node is a virtual machine (VM) hosted by a cloud computing

service (e.g. Amazon Web Services, Google Cloud, Microsoft Azure, etc.). The connection is first made through a gateway node. The client and the gateway are located in the physical world, while the VM nodes are hosted in the cloud.

## 2.3 IP addressing

Hosts are assigned a **unique numerical address** used for all communication and routing called an [Internet Protocol Address \(IP Address\)](#). They look something like this: **128.111.220.7**. Each IP Address can be used to communicate over various “ports”, which allows multiple applications to communicate with a host without mixing up traffic.

Because IP addresses can be difficult to remember, they are also assigned **hostnames**, which are handled through the global [Domain Name System \(DNS\)](#). Clients first look up a hostname in the DNS to find the IP address, then open a connection to the IP address.

**i** Note

The IP address for included-crab.nceas.ucsb.edu is \_\_\_\_\_.

## 2.4 Bash Shell Programming

*What is a shell?* From [Wikipedia](#):

“a computer program which exposes an operating system’s services to a human user or other programs. In general, operating system shells use either a command-line interface (CLI) or graphical user interface (GUI), depending on a computer’s role and particular operation.”

*What is Bash?* Bash, or **Bourne-again Shell**, is a command line tool (language) commonly used to manipulate files and directories. Accessing and using bash is slightly different depending on what type of machine you work on:

- **Mac:** bash via the [Terminal](#), which comes ready-to-use with all Macs
- **Windows:** bash via [Git Bash](#), which needs to be installed

**i** Note

Mac users may have to switch from [Z Shell](#), or zsh, to bash. Use the command `exec bash` to switch your default shell to bash (or `exec zsh` to switch back)

Some commonly used (and very helpful) bash commands:

bash command	what it does
<code>pwd</code>	print your current working directory
<code>cd</code>	change directory
<code>ls</code>	list contents of a directory
<code>tree</code>	display the contents of a directory in the form of a tree structure
<code>mv</code>	move or rename a file
<code>touch</code>	create a new empty file
<code>mkdir</code>	create a new directory
<code>grep</code>	searches a given file(s) for lines containing a match to a given pattern list
<code>awk</code>	
<code>sed</code>	stands for <b>S</b> tream <b>E</b> ditor; a versatile command for editing files
<code>cut</code>	extract a specific portion of text in a file
<code>join</code>	join two files based on a key field present in both
<code>top, htop</code>	view running processes in a Linux system

## 2.5 Connecting to a remote computer via a shell

You can also use a shell to gain accesss to and remotely control other computers (manage/transfer files/etc). To do so, you'll need the following:

- a remote computer (e.g. server) turned on
- the IP address or name of remote computer
- the necessary permissions to access the remote computer

Secure Shell, or SSH, is a network communication protocol that is often used for securely connecting to and running shell commands on a remote host. SSH temendously simplifies remote computing because \_\_\_\_\_, and it is supported out-of-the-box on Linux and Macs. If working on a Windows machine, you'll need \_\_\_\_\_.

## 2.6 Exercise 1: Connect to a server & practice bash commands

Let's connect to a remote computer (a server named **included-crab.nceas.ucsb.edu**) and practice using some of above commands.

### 2.6.1 Launch your Terminal Program

- **MacOS:** navigate to Applications > Utilities and open Terminal
- **Windows:** Navigate to Windows Start > Git and open Git Bash UPDATE: see if this still stands
- **ALTERNATIVELY, from VS Code:** Two options to open a terminal program, if a terminal isn't already an open pane at the bottom of VS Code
  - a) Click on Terminal > New Terminal in top menu bar

- b) Click on the + (dropdown menu) > bash in the bottom right corner

### 2.6.2 Connect to a remote server

You can choose to SSH into the server (included-crab.nceas.ucsb.edu) through the command line by using the `ssh` command, or through VS Code's command palette. If you prefer the latter, please refer back to the [Log in to the server](#) section. Doing so via the command line should look something like this:

```
#| eval: false
samanthacsik:~$ ssh scsik@included-crab.nceas.ucsb.edu
scsik@included-crab.nceas.ucsb.edu's password:
scsik@included-crab:~$
```

**!** Important

You won't see anything appear as you type your password!

**⚠** Warning

#### DO WE NEED THIS SECTION?

3. Change your password

```
#| eval: false
scsik@included-crab:~$ passwd
Changing password for scsik.
(current) UNIX password:
Enter new UNIX password:
Retype new UNIX password:
```

### 2.6.3 Practice

1. Use the `pwd` command to print your current location, or working directory. You'll likely be in your home directory on the server (e.g. `/home/yourusername`).

2. Use the `ls` command to list the contents (any files or subdirectories) of your home directory – you should see a directory named `** ____ **`
3. Use the `cd` command to move the `** ____ **` directory:

```
#| eval: false
# move from /home/yourusername to home/yourusername/_____
cd ___
```

To move *up* a directory level, use two dots, `..`:

```
#| eval: false
# move from /home/yourusername/___ back to /home/yourusername
$ cd ..
```

To quickly navigate back to your home directory from wherever you may be on your computer, use a tilde, `~`:

```
#| eval: false
# e.g. to move from some subdirectory, /home/yourusername/Projects/project1/data, back
$ cd ~

# or use .. to back out three subdirectories
$ cd ../../..
```

4. Create a new directory called, `bash_practice` within your home directory using `mkdir`.

```
#| eval: false
$ mkdir bash_practice
```

Run `ls` to print out the contents of your home directory – you should now see your `bash_practice` directory added to the list.

## 5. TO BE CONTINUED

## 2.7 Exercise 2: Write a bash shell script

As we just demonstrated, we can use bash commands in the terminal to accomplish a variety of tasks like navigating our computer's directories, manipulating/creating/adding files, and much more. However, writing a bash *script* allows us to gather and save our code for automated execution.

Earlier, we created a collection of .txt files and saved them to a new directory called `bash_practice`. Here, we'll write a bash script to iterate over all those files and update \_\_\_\_.

Let's begin by creating a simple bash script that when executed, will print out the message, "Hello, World!" This simple script will help us determine whether or not things are working as expected before writing some more complex (and interesting) code.

1. Open a terminal window and determine where you are by using the `pwd` command. Navigate to where you'd like to save your bash script (your home directory on the server is fine) by using the `cd` command.
2. Next, we'll create a shell script called `hello_world.sh` using the `touch` command:

```
#| eval: false
$ touch hello_world.sh
```

3. There are a number of ways to edit a file or script – here, we'll use [Nano](#), a terminal-based text editor. Open your `hello_world.sh` with nano by running the following in your terminal:

```
#| eval: false
$ nano hello_world.sh
```



You can create and open a file in nano in just one line of code. For example, running `nano hello_world.sh`

is the same as creating the file first using `touch hello_world.sh`, then opening it with nano using `nano hello_world.sh`

4. We can now start to write our script. Some important considerations:

- Anything following a `#` will not be executed as code – these are useful for adding comments to your scripts
- The first line of a Bash script starts with a **shebang**, `#!`, followed by a path to the Bash interpreter – this is used to tell the operating system which interpreter to use to parse the rest of the file. There are two ways to use the shebang to set your interpreter (read up on the pros & cons of both methods on this [Stack Overflow post](#)):

```
#| eval: false

# (option a): use the absolute path to the bash binary
#!/bin/bash

# (option b): use the env utility to search for the bash executable in the user's $PATH environment
#!/usr/bin/env bash
```

5. We'll first specify our bash interpreter using the shebang, which indicates the start of our script. Then, we'll use the `echo` command, which when executed, will print whatever text is passed as an argument. Type the following into your script (which should be opened with nano), then save (Use the keyboard shortcut `control + X` to exit, then type `Y` when it asks if you'd like to save your work. Press `enter/return` to exit nano).

```
#| eval: false
# specify bash as the interpreter
#!/bin/bash

# print "Hello, World!"
$ echo "Hello, World!"
```

6. To execute your script, then run the following in your terminal (be sure that you're in the same working directory as your `hello_world.sh` file or specify the file path to it):

```
#| eval: false  
bash hello_world.sh
```

If successful, “Hello, World!” should be printed in your terminal window.

7.

UPDATE: write a simple shell script that does something – e.g. renaming files with bash loop (e.g. change extension, add date, move them around)

UPDATE: `nohup`, `screen`, `tmux` for starting remote job that you can come back to later; look for tmux lesson in oss training

# 3 Python Programming on Clusters

- Basic Python review
- Using virtual environments
- Writing in Jupyter notebooks
- Writing functions in Python

## 3.1 Introduction

We've chosen to use VS Code in this training, in part, because it has great support for developing on remote machines. Hopefully, your VS Code setup went easily, and you were able to connect to our server `included-crab`. Once connected, the VS Code interface looks just like you were working locally, and connection to the server is seamless.

Other aspects of VS Code that we like: it supports all languages thanks to the extensive free extension library, it has built in version control integration, and it is highly flexible/configurable.

We will also be working quite a bit in Jupyter notebooks in this course. Notebooks are great ways to interleave rich text (mark-down formatted text, equations, images, links) and code in a way that a ‘literate analysis’ is generated. Although Jupyter notebooks are not substitutes for python scripts, they can be great communication tools, and can also be convenient for code development.

## 3.2 Starting a project

To get set up for the course, let's connect to the server again. If you were able to work through the setup for the lesson without

difficulty, follow these steps to connect:

- open the command palette (Cmd + Shift + P)
- enter “Remote SSH: Connect to Host”
- select `included-crab`
- enter your password in the dialog box that pops up

Now we can get set up with a project to work in for the course. Head over to the `scalable-computing-examples` [github repository](#) and fork it to your account.

Back in VS Code, in the terminal clone your fork of the `scalable-computing-examples` repo (`git clone <url-to-forked-repo-here>`) into the top level of your user directory. Run `cd ~/` if you are in some other directory.

To open the project, open the folder into your workspace

- File > Open Folder
- Enter password again if prompted

### 3.3 Virtual Environments

When you install a python library, let's say `pandas`, via `pip`, unless you specify otherwise, `pip` will go out and grab the most recent version of the library, and install it somewhere on your system path (where, exactly, depends highly on how you install python originally, and other factors). This is all great, until you realize that as part of a new project, a new library you are starting to work with requires an older version of `pandas`, what do you do? You need both `pandas` versions for each of your projects. Virtual environments help to solve this issue without making the all too common situation in the comic above even more complicated.

A virtual environment is a folder structure which creates a symbol link (pointer) to all of the libraries that you need into the folder. The three main components will be: the python distribution itself, its configuration, and a site-packages directory (where your

libraries like `pandas` live). So the folder is a self contained directory of all the version-specific python software you need for your project.

Virtual environments are very helpful to create reproducible workflows, and we'll talk more about this concept of reproducible environments later in the course. Perhaps most importantly though, virtual environments also help you maintain your sanity when python programming. Because they are just folders, you can create and delete new ones at will, without worrying about bungling your underlying python setup.

In this course, we are going to use `virtualenv` as our tool to create and manage virtual environments. Other virtual environment tools used commonly are `conda` and `pipenv`. One reason we like using `virtualenv` is there is an extension to it called `virtualenvwrapper`, which provides easy to remember wrappers around common `virtualenv` operations that make creating, activating, and deactivating a virtual environment very easy.

First we will create a `.bash_profile` file to create variables that point to the install locations of python and `virtualenvwrapper`. `.bash_profile` is just a text file that contains bash commands that are run every time you start up a new terminal. Although setting up this file is not required to use `virtualenvwrapper`, it is convenient because it allows you to set up some reasonable defaults to the commands (meaning less typing, overall), and it makes sure that the package is available every time you start a new terminal.

### 3.3.0.1 Setup

- In VS Code, select ‘File > New Text File’
- Paste this text into the file:

```
export VIRTUALENVWRAPPER_VIRTUAWORKON_HOME=$HOME/.virtualenvs
source /usr/share/virtualenvwrapper/virtualenvwrapper.sh
```

The first line points `virtualenvwrapper` to the directory where your virtual environments will be stored. We point it to a hidden directory (`.virtualenvs`) in your home directory. The last

line sources a bash script that ships with `virtualenvwrapper`, which makes all of `virtualenvwrapper` commands available in your terminal session.

- Save the file in the top of your home directory as `.bash_profile`.
- Restart your terminal (Terminal > New Terminal)
- Check to make sure it was installed and configured correctly by running this in the terminal:

```
mkvirtualenv --version
```

It should return some content that looks like this (with more output, potentially).

```
virtualenv 20.13.0+ds from /usr/lib/python3/dist-packages/virtualenv/__init__.py
```

### 3.3.0.2 Course environment

Now we can create the virtual environment we will use for the course. In the terminal run:

```
mkvirtualenv -p python3.9 scomp
```

Here, we've specified explicitly which python version to use by using the `-p` flag, and the path to the python 3.9 installation on the server. After making a virtual environment, it will automatically be activated. You'll see the name of the env you are working in on the left side of your terminal prompt in parentheses. To deactivate your environment (like if you want to work on a different project), just run `deactivate`. To activate it again, run:

```
workon scomp
```

You can get a list of all available environments by just running:

```
workon
```

Now let's install the dependencies for this course into that environment. To install our libraries we'll use `pip`. As of Python 3.4, `pip` is automatically included with your python installation. `pip` is a package manager for python, and you might have used it already to install common python libraries like `pandas` or `numpy`. `pip` goes out to [PyPI](#), the Python Package Index, to download the code and put it in your `site-packages` directory. Note that on this shared server, your user directory will ahve a `site-packages` directory, in addition to one that our systems administrator manages as the root of the system.

```
pip install -r requirements.txt
```

### 3.3.0.3 Installing locally (optional)

`virtualenvwrapper` was already installed on the server we are working on. To install on your local computer, run:

```
pip3 install virtualenvwrapper
```

And then follow the instructions as described above, making sure that you have the correct paths set when you edit your `.bash_profile`.

## 3.4 Brief overview of python syntax

We'll very briefly go over some basic python syntax and the base variable types. First, open a python script. From the File menu, select New File, type “python”, then save it as ‘python-intro.py’ in the top level of your directory.

In your file, assign a value to a variable using `=` and print the result.

```
x = 4
print(x)
```

To run this code in python we can:

- execute `python python-intro.py` in the terminal
- click the Play button in the upper right hand corner of the file editor
- right click any line and select: “Run to line in interactive terminal”

In that interactive window you can then run python code interactively, which is what we’ll use for the next bit of exploring data types.

There are 5 standard data types in python

- Number (int, long, float, complex)
- String
- List
- Tuple
- Dictionary

We already saw a number type, here is a string:

```
str = 'Hello World!'
print(str)
```

Hello World!

Lists in python are very versatile, and are created using square brackets `[]`. Items in a list can be of different data types.

```
list = [100, 50, -20, 'text']
print(list)
```

[100, 50, -20, 'text']

You can access items in a list by index using the square brackets. Note indexing starts with 0 in python. The slice operator enables you to easily access a portion of the list without needing to specify every index.

```
list[0] # print first element  
list[1:3] # print 2nd until 4th elements  
list[:2] # print first until the 3rd  
list[2:] # print last elements from 3rd
```

100

[50, -20]

[100, 50]

[-20, 'text']

The + and \* operators work on lists by creating a new list using either concatenation (+) or repetition (\*).

```
list2 = ['more', 'things']  
  
list + list2  
list * 3
```

[100, 50, -20, 'text', 'more', 'things']

[100, 50, -20, 'text', 100, 50, -20, 'text', 100, 50, -20, 'text']

Tuples are similar to lists, except the values cannot be changed in place. They are constructed with parentheses.

```
tuple = ('a', 'b', 'c', 'd')  
tuple[0]  
tuple * 3  
tuple + tuple
```

'a'

('a', 'b', 'c', 'd', 'a', 'b', 'c', 'd', 'a', 'b', 'c', 'd')

```
('a', 'b', 'c', 'd', 'a', 'b', 'c', 'd')
```

Observe the difference when we try to change the first value. It works for a list:

```
list[0] = 'new value'  
list
```

```
['new value', 50, -20, 'text']
```

...and errors for a tuple.

```
tuple[0] = 'new value'
```

```
TypeError: 'tuple' object does not support item assignment
```

Dictionaries consist of key-value pairs, and are created using the syntax `{key: value}`. Keys are usually numbers or strings, and values can be any data type.

```
dict = {'name': ['Jeanette', 'Matt'],  
       'location': ['Tucson', 'Juneau']}
```

```
dict['name']  
dict.keys()
```

```
['Jeanette', 'Matt']
```

```
dict.keys(['name', 'location'])
```

To determine the type of an object, you can use the `type()` method.

```
type(list)  
type(tuple)  
type(dict)
```

```
list
```

```
tuple
```

```
dict
```

## 3.5 Jupyter notebooks

To create a new notebook, from the file menu select File > New File > Jupyter Notebook

At the top of your notebook, add a first level header using a single hash. Practice some markdown text by creating:

- a list
- **bold** text
- a link

Use the [Markdown cheat sheet](#) if needed.

You can click the plus button below any chunk to add a chunk of either markdown or python.

### 3.5.1 Load libraries

In your first code chunk, lets load in some modules. We'll use `pandas`, `numpy`, `matplotlib.pyplot`, `requests`, `skimpy`, and `exists` from `os.path`.

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import requests
import skimpy
import os
```

A note on style: There are a few ways to construct import statements. The above code uses three of the most common:

```
import module
import module as m
from module import function
```

The first way of importing will make the module a function comes from more explicitly clear, and is the simplest. However for very long module names, or ones that are used very frequently (like `pandas`, `numpy`, and `matplotlib.plot`), the code in the notebook will be more cluttered with constant calls to longer module names. So `module.function()` instead is written as `m.function()`

The second way of importing a module is a good style to use in cases where modules are used frequently, or have extremely long names. If you import every single module with a short name, however, you might have a hard time remembering which modules are named what, and it might be more confusing for others trying to read your code. Many of the most commonly used libraries for python data science have community-driven styling for how they are abbreviated in import statements, and these community norms are generally best followed.

Finally, the last way to import a single object from a module can be helpful if you only need that one piece from a larger module, but again, like the first case, results in less explicit code and therefore runs the risk of your or someone else misremembering the usage and source.

### 3.5.2 Read in a csv

Create a new code chunk that will download the csv that we are going to use for this tutorial.

- Navigate to Rohi Muthyalu, Åsa Rennermalm, Sasha Leidman, Matthew Cooper, Sarah Cooley, et al. 2022. 62 days of Supraglacial streamflow from June-August, 2016 over southwest Greenland. Arctic Data Center. [doi:10.18739/A2XW47X5F](https://doi.org/10.18739/A2XW47X5F).
- Right click the download button for ‘Discharge\_timeseries.csv’
- Click ‘copy link address’

Create a variable called URL and assign it the link copied to your clipboard. Then use `requests.get` to download the file, and `open` to write it to disk, to a directory called `data/`. We'll write this bundled in an `if` statement so that we only download the file if it doesn't yet exist. First, we create the directory if it doesn't exist:

```
if not os.path.exists('data/'):
    os.mkdir('data/')

if not os.path.exists('data/dischARGE_TIMESERIES.csv'):

    url = 'https://arcticdata.io/metacat/d1/mn/v2/object/urn%3Auuid%3Ae248467d-e1f9-4a32

    data = requests.get(url)
    a = open('data/dischARGE_TIMESERIES.csv', 'wb').write(data.content)
```

Now we can read in the data from the file.

```
df = pd.read_csv('data/dischARGE_TIMESERIES.csv')
df.head()
```

```
/opt/hostedtoolcache/Python/3.9.13/x64/lib/python3.9/site-packages/IPython/core/formatters.py:57
    return method()
```

	Date	Total Pressure [m]	Air Pressure [m]	Stage [m]	Discharge [m <sup>3</sup> /s]	temperature [degree]
0	6/13/2016 0:00	9.816	9.609775	0.206225	0.083531	
1	6/13/2016 0:05	9.810	9.609715	0.200285	0.077785	
2	6/13/2016 0:10	9.804	9.609656	0.194344	0.072278	
3	6/13/2016 0:15	9.800	9.609596	0.190404	0.068756	
4	6/13/2016 0:20	9.793	9.609537	0.183463	0.062804	

The column names are a bit messy so we can use `clean_columns` from `skimpy` to make them cleaner for programming very quickly. We can also use the `skim` function to get a quick summary of the data.

We can see that the `date` column is classed as a string, and not a date, so let's fix that.

```

clean_df = skimpy.clean_columns(df)
skimpy.skim(clean_df)

```

6 column names have been cleaned

skimpy summary									
Data Summary		Data Types							
dataframe	Values	Column Type	Count						
Number of rows	17856	float64	5						
Number of columns	6	string	1						
number									
column_name	NA	NA %	mean	sd	p0	p25	p75	p100	p00
total_pressure_m	0	0	9.9	0.12	9.6	9.8	10	10	9.0
air_pressure_m	0	0	9.6	0.06	9.5	9.6	9.7	9.8	9.0
stage_m	0	0	0.28	0.12	0.00056	0.17	0.37	0.50	0.00
discharge_m_3_s	0	0	0.22	0.19	4.7e-08	0.055	0.35	0.50	0.00
temperature_degrees_	8	0.045	-0.034	0.053	-0.1	-0.1	0	0	0
string									
column_name	NA	NA %	words per row				total words		
date	0	0					2		
End									

```
clean_df['date'] = pd.to_datetime(clean_df['date'])

skimpy.skim(clean_df)
```

# skimpy summary

## Data Summary Data Types

dataframe	Values	Column Type	Count
Number of rows	17856	float64	5
Number of columns	6	datetime64	1

## number

column_name	NA	NA %	mean	sd	p0	p25	p75	p10
total_pressure_m	0	0	9.9	0.12	9.6	9.8	10	9.9
air_pressure_m	0	0	9.6	0.06	9.5	9.6	9.7	9.8
stage_m	0	0	0.28	0.12	0.00056	0.17	0.37	0.45
discharge_m_3_s	0	0	0.22	0.19	4.7e-08	0.055	0.35	0.55
temperature_degrees	8	0.045	-0.034	0.053	-0.1	-0.1	0	0

## datetime

column_name	NA	NA %	first	last	fr
date	0	0	2016-06-13	2016-08-13 23:55:00	5T

End

If we wanted to calculate the daily mean flow (as opposed to the flow every 5 minutes), we need to:

- create a new column with only the date
- group by that variable
- summarize over it by taking the mean

First we should probably rename our existing date/time column to prevent from getting confused.

```
clean_df = clean_df.rename(columns = {'date': 'datetime'})
```

Now create the new date column

```
clean_df['date'] = clean_df['datetime'].dt.date
```

Finally, we use group by to split the data into groups according to the date. We can then apply the `mean` method to calculate the mean value across all of the columns. Note that there are other methods you can use to calculate different statistics across different columns (eg: `clean_df.groupby('date').agg({'discharge_m_3_s': 'max'})`).

```
daily_flow = clean_df.groupby('date', as_index = False).mean()
```

- create a simple plot

```
var = 'discharge_m_3_s'
var_labs = {'discharge_m_3_s': 'Total Discharge'}

fig, ax = plt.subplots(figsize=(7, 7))
plt.plot(daily_flow['date'], daily_flow[var])
plt.xticks(rotation = 45)
ax.set_ylabel(var_labs.get('discharge_m_3_s'))
```

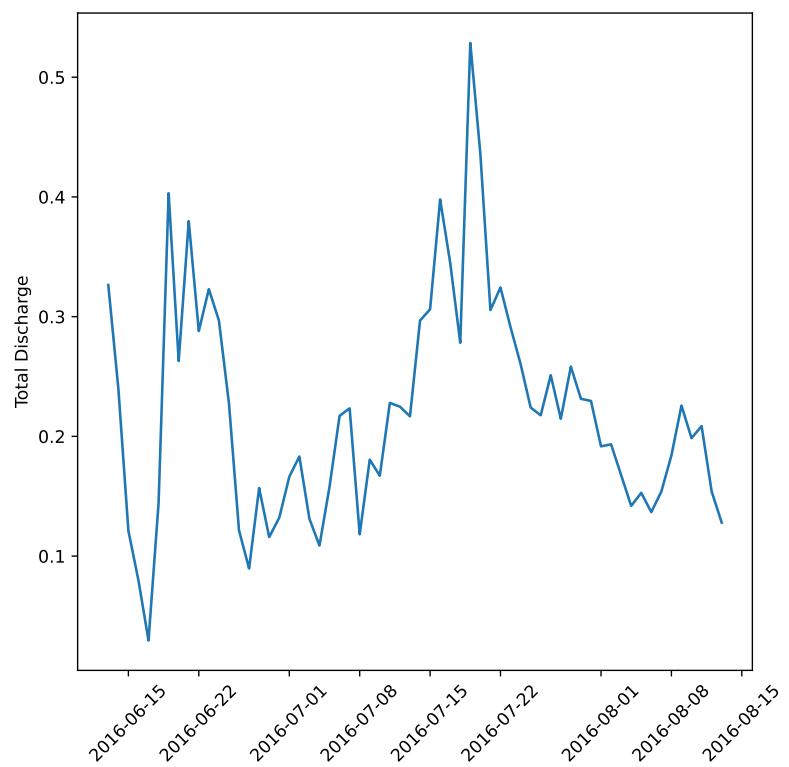
```
(array([16967., 16974., 16983., 16990., 16997., 17004., 17014., 17021.,
       17028.]),
 [Text(0, 0, ''),
  Text(0, 0, '')],
```

```

Text(0, 0, ''),
Text(0, 0, '')
Text(0, 0, '')])

Text(0, 0.5, 'Total Discharge')

```



## 3.6 Functions

The plot we made above is great, but what if we wanted to make it for each variable? We could copy paste it and replace some things, but this violates a core tenet of programming: Don't

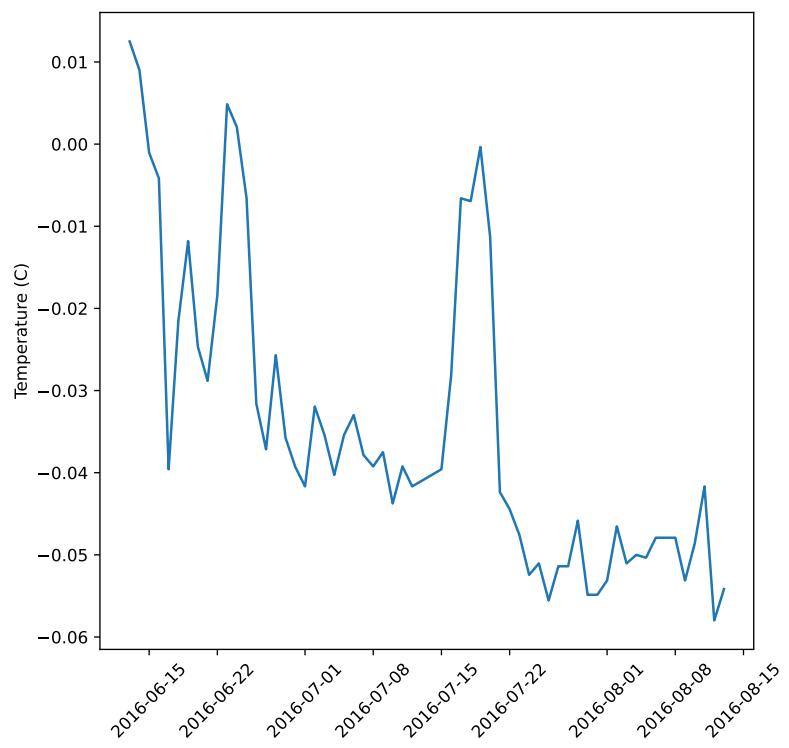
Repeat Yourself! Instead, we'll create a function called `myplot` that accepts the data frame and variable as arguments.

- create `myplot.py`

```
def myplot(df, var):  
  
    var_labs = {'discharge_m_3_s': 'Total Discharge (m^3/s)',  
                'total_pressure_m': 'Total Pressure (m)',  
                'air_pressure_m': 'Air Pressure (m)',  
                'stage_m': 'Stage (m)',  
                'temperature_degrees_c': 'Temperature (C)'}  
  
    fig, ax = plt.subplots(figsize=(7, 7))  
    plt.plot(df['date'], df[var])  
    plt.xticks(rotation = 45)  
    ax.set_ylabel(var_labs.get(var))
```

- load `myplot` into jupyter notebook (`from myplot.py import myplot`)
- replace old plot method with new function

```
myplot(daily_flow, 'temperature_degrees_c')
```



We'll have more on functions in the software design sections.

# 4 Pleasingly Parallel Programming

- Understand what parallel computing is and when it may be useful
- Understand how parallelism can work
- Review sequential loops and map functions
- Build a parallel program using `concurrent.futures`
- Build a parallel program using `parsl`
- Understand Thread Pools and Process pools

## 4.1 Introduction

Processing large amounts of data with complex models can be time consuming. New types of sensing means the scale of data collection today is massive. And modeled outputs can be large as well. For example, here's a 2 TB (that's Terabyte) set of modeled output data from [Ofir Levy et al. 2016](#) that models 15 environmental variables at hourly time scales for hundreds of years across a regular grid spanning a good chunk of North America:

There are over 400,000 individual netCDF files in the [Levy et al. microclimate data set](#). Processing them would benefit massively from parallelization.

Alternatively, think of remote sensing data. Processing airborne hyperspectral data can involve processing each of hundreds of bands of data for each image in a flight path that is repeated many times over months and years.

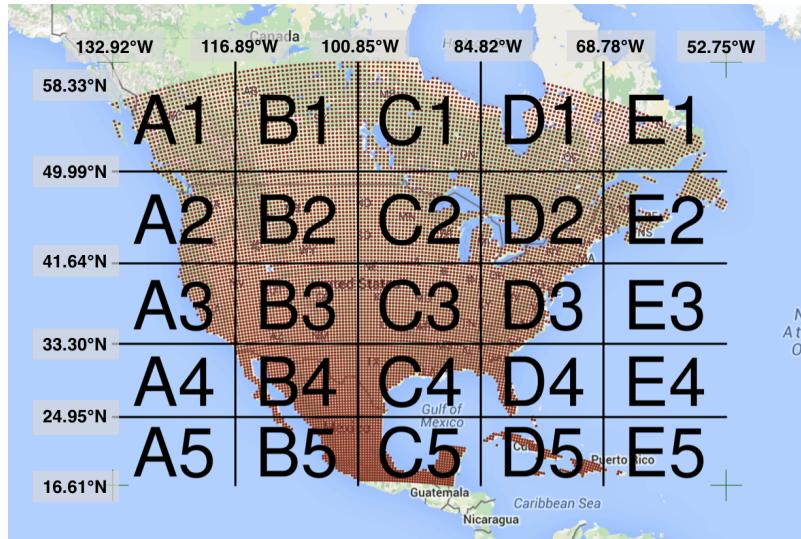


Figure 4.1: Levy et al. 2016. doi:10.5063/F1Z899CZ

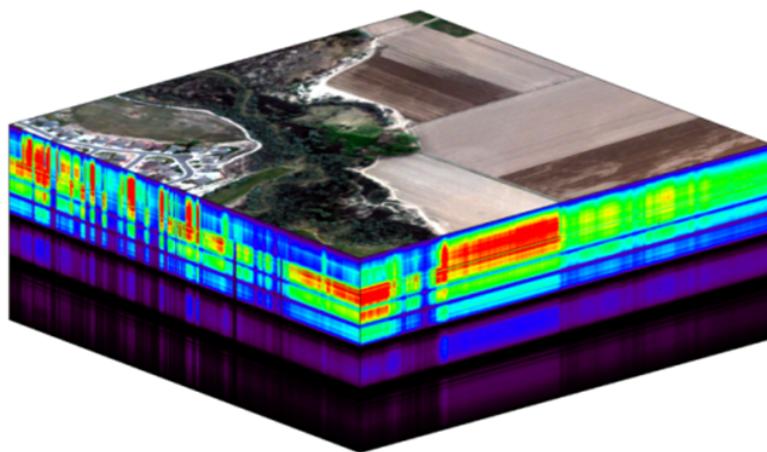


Figure 4.2: NEON Data Cube

## 4.2 Why parallelism?

Much R code runs fast and fine on a single processor. But at times, computations can be:

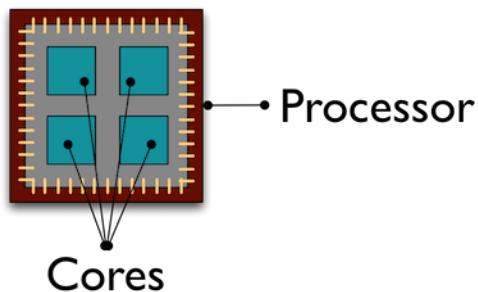
- **cpu-bound**: Take too much cpu time
- **memory-bound**: Take too much memory
- **I/O-bound**: Take too much time to read/write from disk
- **network-bound**: Take too much time to transfer

To help with **cpu-bound** computations, one can take advantage of modern processor architectures that provide multiple cores on a single processor, and thereby enable multiple computations to take place at the same time. In addition, some machines ship with multiple processors, allowing large computations to occur across the entire cluster of those computers. Plus, these machines also have large amounts of memory to avoid **memory-bound** computing jobs.

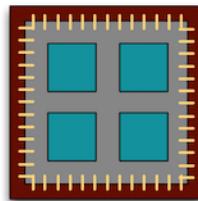
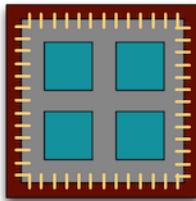
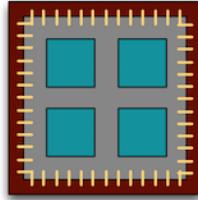
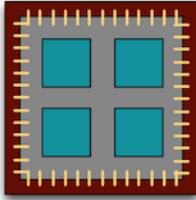
## 4.3 Processors (CPUs) and Cores

A modern CPU (Central Processing Unit) is at the heart of every computer. While traditional computers had a single CPU, modern computers can ship with multiple processors, which in turn can each contain multiple cores. These processors and cores are available to perform computations.

A computer with one processor may still have 4 cores (quad-core), allowing 4 computations to be executed at the same time.



A typical modern computer has multiple cores, ranging from one or two in laptops to thousands in high performance compute clusters. Here we show four quad-core processors for a total of 16 cores in this machine.



You can think of this as allowing 16 computations to happen at the same time. Theoretically, your computation would take 1/16 of the time (but only theoretically, more on that later).

Historically, R has only utilized one processor, which makes it single-threaded. Which is a shame, because the 2017 MacBook Pro that I am writing this on is much more powerful than that:

```
{bash eval=FALSE} jones@powder:~$ sysctl hw.ncpu  
hw.physicalcpu hw.ncpu: 8 hw.physicalcpu: 4
```

To interpret that output, this machine `powder` has 4 physical CPUs, each of which has two processing cores, for a total of 8 cores for computation. I'd sure like my R computations to use all of that processing power. Because its all on one machine, we can easily use *multicore* processing tools to make use of those cores. Now let's look at the computational server `aurora` at NCEAS:

```
{bash eval=FALSE} jones@included-crab:~$ lscpu  
| egrep 'CPU\(\s*\)|per core|per socket' CPU(s):  
88 On-line CPU(s) list: 0-87 Thread(s) per core:  
2 Core(s) per socket: 22 NUMA node0 CPU(s):
```

0,2,4,6,8,10,12,14,16,18,20,22,24,26,28,30,32,34,36,38,40,42,44,46,48,50,52,54,56,58,60,62,64,66  
NUMA node1 CPU(s): 1,3,5,7,9,11,13,15,17,19,21,23,25,27,29,31,33,35,37,39,41,43,45,47,49,51

Now that's more compute power! `included-crab` has 384 GB of RAM, and ample storage. All still under the control of a single operating system.

However, maybe one of these NSF-sponsored high performance computing clusters (HPC) is looking attractive about now:

- JetStream
    - 640 nodes, 15,360 cores, 80TB RAM
  - Stampede2 at TACC is coming online in 2017
    - 4200 nodes, 285,600 cores
  - TODO: update with modern cluster sizes

Note that these clusters have multiple nodes (hosts), and each host has multiple cores. So this is really multiple computers clustered together to act in a coordinated fashion, but each node runs its own copy of the operating system, and is in many ways independent of the other nodes in the cluster. One way to use such a cluster would be to use just one of the nodes, and use a multi-core approach to parallelization to use all of the cores on that single machine. But to truly make use of the whole cluster, one must use parallelization tools that let us spread out our computations across multiple host nodes in the cluster.

## 4.4 Modes of parallelization

- TODO: develop diagram(s) showing
    - Single memory image task parallelization

Serial Launch tasks --> Task 1 --> Task 2 --> Task 3  
--> Task 4 --> Task 5 --> Finish

- Cluster task parallelization

```

Cluster    parallel    Show dispatch to cluster nodes and
reassembly of data    Launch tasks -->                         Marshal
--> Task 1 --> Unmarshal --\                                Marshal
--> Task 2 --> Unmarshal ---\                            Marshal
--> Task 3 --> Unmarshal -----> Finish                  Marshal
--> Task 4 --> Unmarshal ---/                            Marshal
--> Task 5 --> Unmarshal --/

```

- TODO: Should we also include figure with data or functional dependencies?

## 4.5 Task parallelism with concurrent.futures

When you have a list of repetitive tasks, you may be able to speed it up by adding more computing power. If each task is completely independent of the others, then it is a prime candidate for executing those tasks in parallel, each on its own core. For example, let's build a simple loop that downloads the data files that we need for an analysis. First, we start with the serial implementation.

```

# Use loop for serial execution of tasks

# Tasks are to download data from a dataset

```

The issue with this loop is that we execute each trial sequentially, which means that only one of our 8 processors on this machine are in use. In order to exploit parallelism, we need to be able to dispatch our tasks as functions, with one task going to each processor. To do that, we need to convert our task to a function, and then use the `map()` function to apply that function to all of the members of a set. Here's the same code rewritten to use `map()`, which applies a function to each of the members of a list (in this case the files we want to download):

```
# Use `map` for serial execution of tasks  
  
# Tasks are to download data from a dataset
```

## 4.6 Approaches to parallelization

When parallelizing jobs, one can:

- Use the multiple cores on a local computer through `mclapply`
- Use multiple processors on local (and remote) machines using `makeCluster` and `clusterApply`
  - In this approach, one has to manually copy data and code to each cluster member using `clusterExport`
  - This is extra work, but sometimes gaining access to a large cluster is worth it

## 4.7 concurrent.futures

```
# Loop versus map for parallel execution of tasks  
  
# Using concurrent.futures and ThreadPool  
  
# Tasks are to download data from a dataset
```

## 4.8 parsl

- Overview of parsl and its use of python decorators.

```
# Loop versus map for parallel execution of tasks  
  
# Using parsl decorators and ThreadPool  
  
# Tasks are to download data from a dataset
```

- Configurable Executors in parsl
    - HighThroughputExecutor for cluster jobs
- ```
# Loop versus map for parallel execution of tasks

# Using parsl decorators and ThreadPool

# Tasks are to download data from a dataset
```

## 4.9 When to parallelize

It's not as simple as it may seem. While in theory each added processor would linearly increase the throughput of a computation, there is overhead that reduces that efficiency. For example, the code and, importantly, the data need to be copied to each additional CPU, and this takes time and bandwidth. Plus, new processes and/or threads need to be created by the operating system, which also takes time. This overhead reduces the efficiency enough that realistic performance gains are much less than theoretical, and usually do not scale linearly as a function of processing power. For example, if the time that a computation takes is short, then the overhead of setting up these additional resources may actually overwhelm any advantages of the additional processing power, and the computation could potentially take longer!

In addition, not all of a task can be parallelized. Depending on the proportion, the expected speedup can be significantly reduced. Some propose that this may follow [Amdahl's Law](#), where the speedup of the computation (y-axis) is a function of both the number of cores (x-axis) and the proportion of the computation that can be parallelized (see colored lines):

```
#| eval: false
library(ggplot2)
library(tidyr)
amdahl <- function(p, s) {
  return(1 / ( (1-p) + p/s ))
}
```

```

doubles <- 2^(seq(0,16))
cpu_perf <- cbind(cpus = doubles, p50 = amdahl(.5, doubles))
cpu_perf <- cbind(cpu_perf, p75 = amdahl(.75, doubles))
cpu_perf <- cbind(cpu_perf, p85 = amdahl(.85, doubles))
cpu_perf <- cbind(cpu_perf, p90 = amdahl(.90, doubles))
cpu_perf <- cbind(cpu_perf, p95 = amdahl(.95, doubles))
#cpu_perf <- cbind(cpu_perf, p99 = amdahl(.99, doubles))
cpu_perf <- as.data.frame(cpu_perf)
cpu_perf <- cpu_perf %>% gather(prop, speedup, -cpus)
ggplot(cpu_perf, aes(cpus, speedup, color=prop)) +
  geom_line() +
  scale_x_continuous(trans='log2') +
  theme_bw() +
  labs(title = "Amdahl's Law")

```

So, it's important to evaluate the computational efficiency of requests, and work to ensure that additional compute resources brought to bear will pay off in terms of increased work being done. With that, let's do some parallel computing...

# 5 Parallel Pitfalls and their solutions

- Race conditions
- Deadlocks

## 5.1 Summary

In this lesson, we showed examples of computing tasks that are likely limited by the number of CPU cores that can be applied, and we reviewed the architecture of computers to understand the relationship between CPU processors and cores. Next, we reviewed the way in which traditional `for` loops in R can be rewritten as functions that are applied to a list serially using `lapply`, and then how the `parallel` package `mclapply` function can be substituted in order to utilize multiple cores on the local computer to speed up computations. Finally, we installed and reviewed the use of the `foreach` package with the `%dopar` operator to accomplish a similar parallelization using multiple cores.

## 5.2 Further Reading

Ryan Abernathey & Joe Hamman. 2020. [Closed Platforms vs. Open Architectures for Cloud-Native Earth System Analytics](#). Medium.

# **6 Documenting and Publishing Data**

## **6.1 Introduction**

# 7 Group Project: Staging and Preprocessing

- Get familiarized with the overall group project workflow
- Write a parsl app that will stage and tile the IWP example data in parallel

## 7.1 Introduction

The Permafrost Discovery Gateway is an online platform for archiving, processing, analysis, and visualization of permafrost big imagery products to enable discovery and knowledge-generation. The PDG utilizes and makes available products derived from high resolution satellite imagery from the Polar Geospatial Center, Planet (3m), Sentinel (10 m), Landsat (30 m), and MODIS (250 m). One of these products is a dataset showing Ice Wedge Polygons (IWP) that form in melting permafrost.

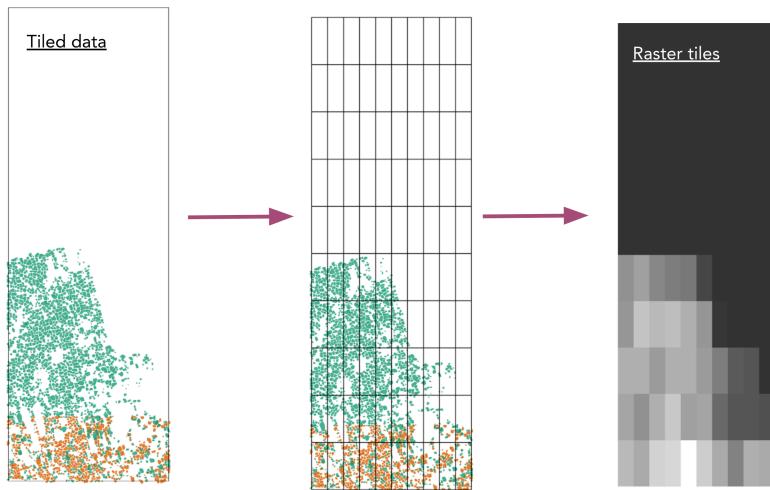
Ice wedges form as a result of thermal contraction during melt/freeze cycles of permafrost. They can form very distinctive geometries clearly visible in satellite images. The PDG is using advanced analysis and computational tools to take high resolution satellite imagery and automatically detect where ice wedge polygons form. Below is an example of a satellite image (left) and the detected ice wedge polygons in geospatial vector format (right) of that same image.



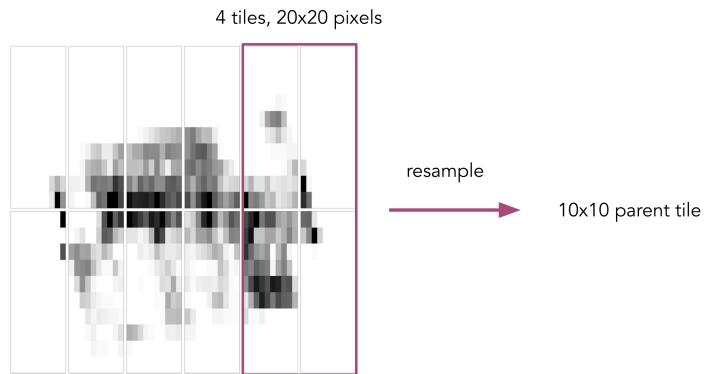
In the group project, we are going to use a subset of the high resolution dataset of these detected ice wedge polygons in order to learn some of the reproducible, scalable techniques that will allow us to process it. Our workflow will start with a set of large geopackage files that contain the detected ice wedge polygons. These files all have irregular extents due to the variation in satellite coverage, clouds, etc. Our first processing step will take these files and “tile” them into smaller files which have regular extents.



In step two of the workflow, we will take those regularly tiled geopackage files and rasterize them. The files will be regularly gridded, and a summary statistic will be calculated for each grid cell (such as the proportion of pixel area covered by polygons).



In the final step of the workflow, we will take the raster files and resample them to create a set of raster tiles at different resolutions. This last step is what will enable us to visualize our raster data dynamically, such that we look at lower resolutions when very zoomed out (and high resolution data would take too long to load), and higher resolution data when zoomed in and the extent is smaller.



## 7.2 Staging and Tiling

Today we will undertake the first step of the workflow, staging and tiling the data.



In your fork of the [scalable-computing-examples](#) repository, open the Jupyter notebook in the `group-project` directory called `session-06.ipynb`. This workbook will serve as a skeleton for you to work in. It will load in all the libraries you need, including a few helper functions we wrote for the course, show an example for how to use the method on one file, and then lays out blocks for you and your group to fill in with code that will run that method in parallel.

In your small groups, work together to write the solution, but everyone should aim to have a working solution on their own fork of the repository. In other words, everyone should type out the solution themselves as part of the group effort. Writing the code out yourself (even if others are contributing to the content) is a great way to get “mileage” as you develop these skills.

## **8 Software Design I**

# 9 Data Structures and Formats for Large Data

- Look into parallel access and NetCDF and whether all wheels can take advantage of the built-in parallel access. Looking at [NetCDF's Parallel I/O page](#) you can see the low-level libraries support it.
- Find dataset(s) to use in the xarray hands on.
  - Jeanette says there on on the Arctic Data Center we store in /var/data. Ask her for more info.
  - Candidate: <https://www.ncei.noaa.gov/products/climate-data-records/snow-cover-extent> (<https://arcticdata.io/catalog/view/doi%3A10.7289%2FV5N014G9>)
  - Candidate: <https://arcticdata.io/catalog/view/doi%3A10.18739%2FA24B2X59C> “Understory micrometeorology across a larch forest density gradient in northeastern Siberia 2014-2020” ~50MB, has a few different axes to filter on
  - Also check out candidates in the Dask lesson TODO section

## 9.1 Learning Objectives

- Learn about the NetCDF data format
- Understand how formats such as NetCDF differ from formats like CSV in terms of random/arbitrary access and how that applies to parallel computing
- Learn how to use the `xarray` package to work with N-Dimensional datasets in NetCDF files

## 9.2 Introduction

TODO

- [Insert amazing image representing “large scale” and “multidimensional”]
- Talk about how the choice of our data formats is at least as important as how we parallelize our code
- Maybe talk about how it’s very common to convert our data from a less efficient format (e.g., CSV) into a more efficient one (NetCDF, Parquet) *before* we parallelize

## 9.3 Working with Large Data

TODO

- Size & Dimension, “N-D”: Talk about common dimensions (ie space-time, where space is xy, or xyz). *This is a great time to ask students about their N-Dimensional experience*
- Data Organization / File Naming?
  - Using hierarchical folder structures
  - Encoding hierarchy into filenames
  - File naming for natural ordering (principle of most sig. first, ie YMD vs DMY)
  - Setting things up for multi-file support in tools like Dask, XArray, Arrow

## 9.4 NetCDF/HDF Overview

TODO

Overview major functionality of NetCDF. Students who aren’t already familiar with NetCDF should come away feeling more able to engage with it and maybe even excited about it.

- NetCDF is for Multidimensional / N-Dimensional data: Storing multidimensional data in common tabular formats such as CSV can get cumbersome, though there are ways to do it by adding columns full of redundant data or splitting data into multiple files. But once you get into 4+ dimensions (e.g., x, y, z, time) formats like NetCDF start looking a lot better.

- NetCDF is “self-describing”, meaning we can encode metadata about every variable such as its units, definitions, and lots of other information
- Cover the [NetCDF data model](#) so students have the lingo
- Random access: While CSVs are ubiquitous and easy to work with, they can get really slow to read into our scripts and programs and can take up a lot of RAM because we have to read and parse the entire CSV into RAM before we can do anything with it. With NetCDF, on the other hand, we don’t need to read the entire file into RAM before we query it. Instead of reading the entire dataset into RAM, we only need to read the first part of a NetCDF to get enough information about the data in the rest of the file in order to query it. We call this random or arbitrary access meaning the file describes itself well enough for us to know where the subset of the file we care about is within the whole file. Random access file formats such as NetCDF are fantastic in scalable computing contexts because our parallel workers don’t need to read our data files completely into RAM before running our queries or other transformation.
- Remote access: See [https://rabernat.github.io/research\\_computing\\_2018/xarray-tips-and-tricks.html](https://rabernat.github.io/research_computing_2018/xarray-tips-and-tricks.html). Because NetCDF supports random access (we don’t have to read the entire file to know where the subset of the data we want is), servers can host more NetCDF files on disk than they could store in RAM, throw something like a THREDDS server in front of them, and we can just query the files without making the server read each file into memory and the server can return just the data we asked for.

Other topics:

- Describe available tooling (python packages, Panoply, others?). This doesn’t necessarily need to be hands on:
  - Command line
  - Python packages: xarray
  - GUI applications: Panoply, others?
  - Some students might have experience with this, what do they use and like/hate?

- ? Talk about CF conventions (climate forecast conventions for metadata). This could be cut for time.
- Talk about NetCDF and its role in data archival
  - Is it a good archival format: Yes! Better than CSV in many (not all) ways. The format is open and well-documented, support for the reading the format is ubiquitous, it's efficient w/ disk space (compared to CSV), it supports remote querying (unlike CSV).

## 9.5 Introduction to Xarray

[Time estimate ~20-30min, link to the dataset and work through dataset with students bit-by-bit]

TODO: Base on <https://docs.xarray.dev/en/stable/getting-started-guide/quick-overview.html> but with a course-appropriate dataset.

Course appropriate datasets:

- Probably a space<->time one, maybe one with x, y, z and time
- Jeanette says there's a nice one One in /var/data on ADC, may have to ask her for more information

The focus here is for students to learn how to open up a NetCDF file, get info on it, read in the data and do some basic map-reduce type operations using xarray. Being able to write out a NetCDF file might be outside the scope here.

## 9.6 Exercise

(30-45min)

TODO: Students work on their own or in pairs to write a script to analyze a NetCDF dataset

# 10 Parallelization with Dask

- Find dataset(s) to use:
  - <https://arcticdata.io/catalog/view/doi%3A10.18739%2FA24B2X59C>  
“Understory micrometeorology across a larch forest density gradient in northeastern Siberia 2014-2020”  
~50MB, has a few different axes to filter on
  - <https://arcticdata.io/catalog/view/doi%3A10.18739%2FA28W38388>  
“River and lake ice phenology data for Alaska and Northwest Canada from 1882 to 2021”
  - Also check out candidates in the xarray lesson TODO section

## 10.1 Learning Objectives

- Learn about the map-reduce
- Learn how to use Dask

## 10.2 Introduction

TODO

### 10.2.1 Notes

- <https://www.dask.org/>
- <https://docs.xarray.dev/en/stable/user-guide/dask.html#dask>
- <https://stephanhoyer.com/2015/06/11/xray-dask-out-of-core-labeled-arrays/>
- <https://examples.dask.org/xarray.html>

- Good example to base exercise on: <https://examples.dask.org/applications/image-processing.html>
- Split-apply-combine
- Dask stuff
  - Lazy eval (compute())
    - \* Grouping compute() calls versus calling compute() multiple times
  - Dask Array
  - Dask DataFrame
  - Skip or just mention Bag, Delayed, Futures? Not sure yet.
  - visualize()
  - Choosing how many chunks to divide work into
  - Task overhead
  - Distributed dask?
    - \* Persist > Dask is convenient on a laptop. It installs trivially with conda or pip and extends the size of convenient datasets from “fits in memory” to “fits on disk”.
- From <https://docs.dask.org/en/stable/>

## 10.3 Dask Tutorial

TODO ## Exercise

TODO (50min)

## 10.4 Conclusion

- Comparison with other libraries Thread pools, process pools, distributed dask clusters

# 11 Spatial and Image Data Using GeoPandas

- Reading raster data with rasterio
- Using geopandas and rasterio to process raster data
- Working with raster and vector data together

## 11.1 Introduction

In this lesson, we'll be working with geospatial raster and vector data to do an analysis on vessel traffic in south central Alaska. If you aren't already familiar, geospatial vector data consists of points, lines, and/or polygons, which represent locations on the Earth. Geospatial vector data can have differing geometries, depending on what it is representing (eg: points for cities, lines for rivers, polygons for states.) Raster data uses a set of regularly gridded cells (or pixels) to represent geographic features.

Both geospatial vector and raster data have a coordinate reference system, which describes how the points in the dataset relate to the 3-dimensional spheroid of Earth. A coordinate reference system contains both a datum and a projection. The datum is how you georeference your points (in 3 dimensions!) onto a spheroid. The projection is how these points are mathematically transformed to represent the georeferenced point on a flat piece of paper. All coordinate reference systems require a datum. However, some coordinate reference systems are “unprojected” (also called geographic coordinate systems). Coordinates in latitude/longitude use a geographic (unprojected) coordinate system. One of the most commonly used geographic coordinate systems is WGS 1984.

Coordinate reference systems are often referenced using a short-hand 4 digit code called an EPSG code. We'll be working with two coordinate referece systems in this lesson with the following codes:

- 3338: Alaska Albers
- 4326: WGS84 (World Geodetic System 1984), used in GPS

In this lesson, we are going to take two datasets:

- [Alaskan commercial salmon fishing statisical areas](#)
- [North Pacific and Arctic Marine Vessel Traffic Dataset](#)

and use them to calculate the total distance travelled by ships within each fishing area.

The high level steps will be

- read in the datasets
- reproject them so they are in the same projection
- extract a subset of the raster and vector data using a bounding box
- turn each polygon in the vector data into a raster mask
- use the masks to calculate the total distance travelled (sum of pixels) for each fishing area

## 11.2 Pre-processing raster data

First we need to load in our libraries. We'll use `geopandas` for vector manipulation, `rasterio` for raster maniupulation, and `shapely` for manipulating geospatial data generally.

```
import geopandas as gpd

import rasterio
import rasterio.mask
import rasterio.warp
import rasterio.plot
from rasterio import features
```

```
from shapely.geometry import box
from shapely.geometry import Polygon

import requests

import matplotlib.pyplot as plt
from matplotlib import style

import pandas as pd
import numpy as np
```

First, we'll use `requests` to download the ship traffic raster from [Kapsar et al.](#). We grab a one month slice from August, 2020 of a coastal subset of data with 1km resolution. To get the URL in the code chunk below, you can right click the download button for the file of interest and select "copy link address."

```
url_sf = 'https://arcticdata.io/metacat/d1/mn/v2/object/urn%3Auuid%3A2cc77a90-080b-4696-b5bc...'

response_sf = requests.get(url_sf)
open("Coastal_2020_08.tif", "wb").write(response_sf.content)
```

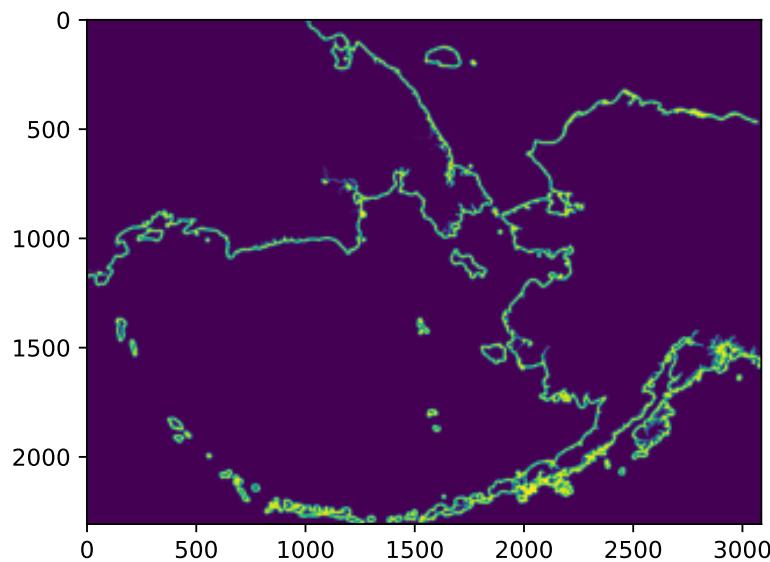
1422727

Open the raster file, plot it, and look at the metadata.

```
with rasterio.open("Coastal_2020_08.tif") as dem_src:
    ships = dem_src.read(1)
    ships_meta = dem_src.profile

    plt.imshow(ships)
    print(ships_meta)

{'driver': 'GTiff', 'dtype': 'float32', 'nodata': -3.3999999521443642e+38, 'width': 3087, 'height': 2057, 'affine': Affine(0.0, -999.9687691991521, 2711703.104608573), 'tiled': False, 'compress': 'lzw', 'interleave': 'band'}
```



Now download a vector shapefile of commercial fishing districts in Alaska.

```
url = 'https://knb.ecoinformatics.org/knb/d1/mn/v2/object/urn%3Auuid%3A7c942c45-1539-4d47-b4'

response = requests.get(url)
open("Alaska_Commercial_Salmon_Boundaries.gpkg", "wb").write(response.content)
```

36544512

Read in the data

```
comm = gpd.read_file("Alaska_Commercial_Salmon_Boundaries.gpkg")

comm.crs
```

```
<Geographic 2D CRS: EPSG:4326>
Name: WGS 84
Axis Info [ellipsoidal]:
- Lat[north]: Geodetic latitude (degree)
- Lon[east]: Geodetic longitude (degree)
```

Area of Use:

- name: World.
- bounds: (-180.0, -90.0, 180.0, 90.0)

Datum: World Geodetic System 1984 ensemble

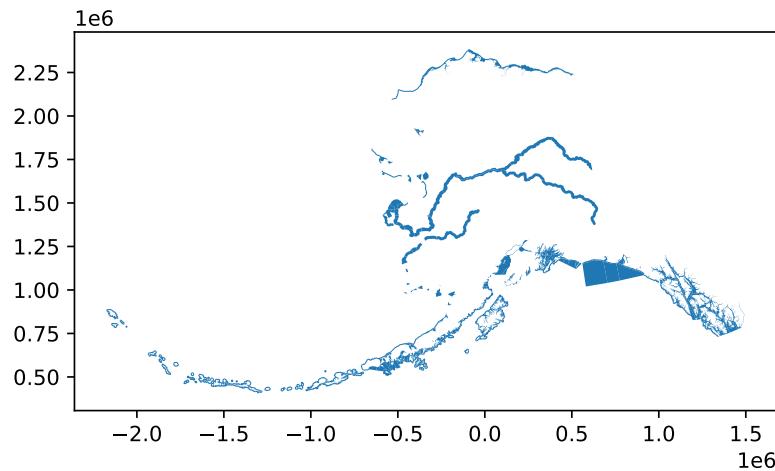
- Ellipsoid: WGS 84
- Prime Meridian: Greenwich

The raster data is in 3338, so we need to reproject this. We use the `to_crs` method on the `comm` object to transform it from 4326 (WGS 84) to 3338 (Alaska Albers).

```
comm.crs
comm_3338 = comm.to_crs("EPSG:3338")

comm_3338.plot()
```

<AxesSubplot:>



We can create a bounding box for the area of interest, and use that to clip the original raster data to just the extent we need. We use the `box` function from `shapely` to create the bounding box, then create a `geoDataFrame` from the points, and convert the WGS 84 coordinates to the Alaska Albers projection.

```

coord_box = box(-159.5, 55, -144.5, 62)

bbox_crop = gpd.GeoDataFrame(
    crs = 'EPSG:4326',
    geometry = [coord_box]).to_crs("EPSG:3338")

```

Now, we can read in raster again cropped to bounding box. We use the `mask` function from `rasterio.mask`. Note that we apply this to the connection to the raster file (with `rasterio.open(...)`), update the metadata associated with the raster, and then write it back out again.

```

with rasterio.open("Coastal_2020_08.tif") as src:
    out_image, out_transform = rasterio.mask.mask(src, bbox_crop["geometry"], crop=True)
    out_meta = src.meta

    out_meta.update({"driver": "GTiff",
                     "height": out_image.shape[1],
                     "width": out_image.shape[2],
                     "transform": out_transform,
                     "compress": "lzw"})

with rasterio.open("Coastal_2020_08_masked.tif", "w", **out_meta) as dest:
    dest.write(out_image)

```

We can also clip the shapefile data to the same bounding box using the `clip` method, passing it as an argument the `geometry` column of our `bbox_crop` geo data frame.

```
comm_clip = comm_3338.clip(bbox_crop['geometry'])
```

### 11.2.1 Check extents

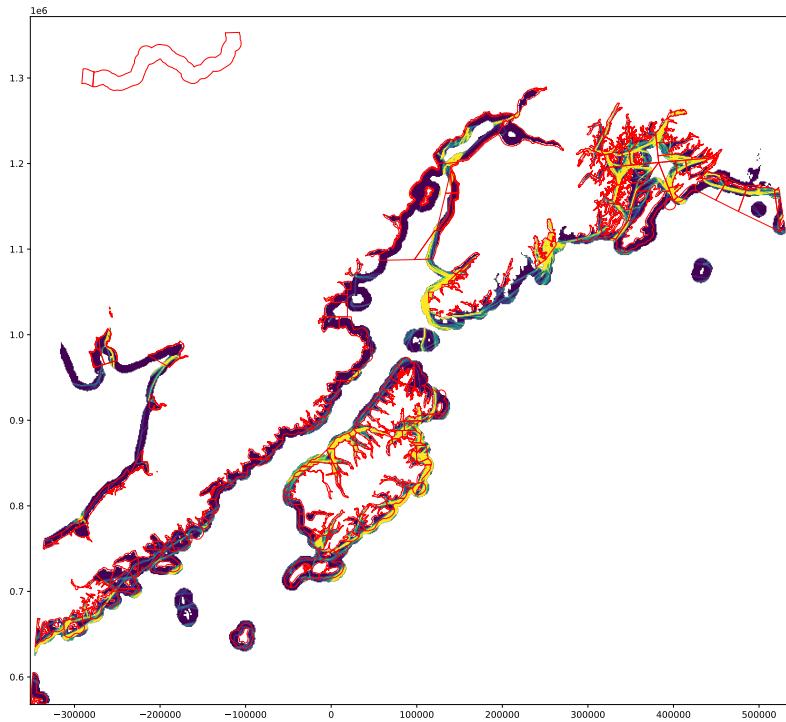
Let's read in the clipped raster data, and make a quick plot to ensure they are in the same extent, and look as expected.

First we read in the cropped data again, since we'll need it later. We also save the `shape` and `transform` attributes of the raster into their own attributes.

```
with rasterio.open('Coastal_2020_08_masked.tif') as src:
    shape = src.shape
    transform = src.transform
    # read in the cropped raster
    ship_arr = src.read(1)
    # turn no data values into actual NaNs
    ship_arr[ship_arr == src.nodata] = np.nan

# set up plot
fig, ax = plt.subplots(figsize=(15, 15))
# plot the raster
rasterio.plot.show(ship_arr,
                    ax=ax,
                    vmin = 0,
                    vmax = 6000,
                    transform = transform)
# plot the vector
comm_clip.plot(ax=ax, facecolor='none', edgecolor='red')
```

<AxesSubplot:>



### 11.3 Calculate total distance per fishing area

In this step, we rasterize each polygon in the shapefile that falls within the bounds of the raster data we are calculating statistics for.

This process returns a dictionary of indexed arrays, where each item corresponds to one polygon (fishing area). The array contains the indices of the original raster that fall within that fishing area. In other words, we are creating a raster mask for each fishing district.

First we'll create an `id` column in the vector data frame. This will help us track unique fishing districts later.

```
comm_3338['id'] = range(0,len(comm_3338))
```

### 11.3.0.1 Zonal statistics over one polygon

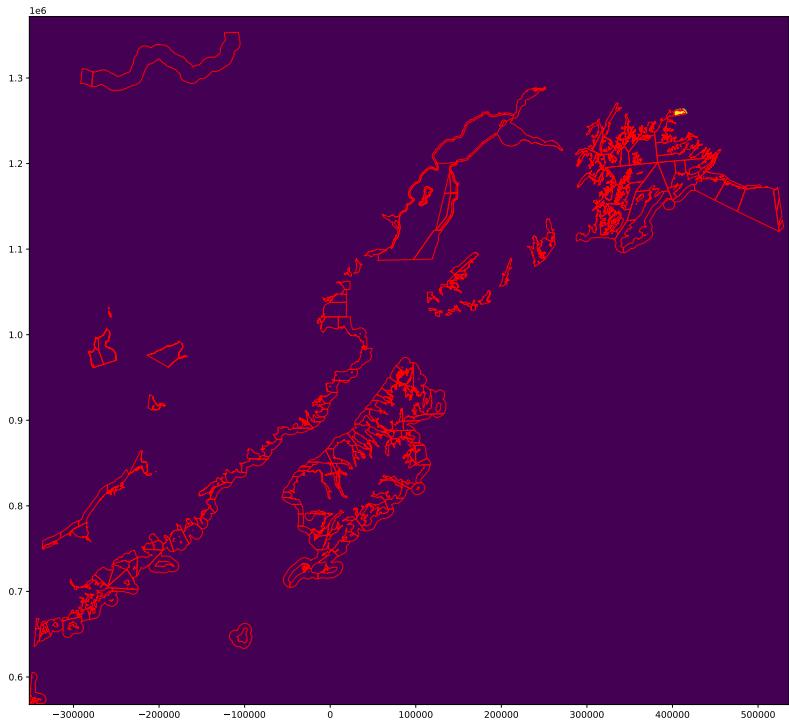
Let's look at how this works over just one fishing area first. We use the `rasterize` method from the `features` module in `rasterio`. This takes as arguments that data to rasterize (in this case the 40th row of our dataset), the shape and transformation the output raster will take (these were extracted from our raster data when we read it in). We also set the `all_touched` argument to true, which means any pixel that touches a boundary of our vector will be burned into the mask.

```
r40 = features.rasterize(comm_3338['geometry'][40].geoms,
                         out_shape=shape,
                         transform=transform,
                         all_touched=True)
```

If we have a look at a plot of our rasterized version of the single fishing district, we can see that instead of a vector, we now have a raster with the shape of the district.

```
# set up plot
fig, ax = plt.subplots(figsize=(15, 15))
# plot the raster
rasterio.plot.show(r40,
                    ax=ax,
                    vmin = 0,
                    vmax = 1,
                    transform = transform)
# plot the vector
comm_clip.plot(ax=ax, facecolor='none', edgecolor='red')
```

<AxesSubplot:>



A quick call to `np.unique` shows our unique values are 0 or 1, which is what we expect.

```
np.unique(r40)
```

```
array([0, 1], dtype=uint8)
```

Finally, we need to know is the indices of the original raster where the fishing district is. We can use `np.where` to extract this information

```
r40_index = np.where(r40 == 1)
print(r40_index)
```

```
(array([108, 108, 108, 108, 108, 109, 109, 109, 109, 109, 109,
       109, 109, 110, 110, 110, 110, 110, 110, 110, 110, 110, 110,
       110, 110, 110, 111, 111, 111, 111, 111, 111, 111, 111,
```

```
111, 111, 111, 111, 112, 112, 112, 112, 112, 112, 112, 112, 112,
112, 112, 112, 112, 112, 113, 113, 113, 113, 113, 113, 113,
113, 113, 113, 113, 113, 113, 113, 114, 114, 114, 114, 114,
114, 114, 114, 114, 114, 115, 115, 115, 115, 115, 115, 115, 116]), array([759, 760, 762,
764, 765, 753, 754, 755, 756, 757, 758, 759, 760, 761, 762, 763,
764, 765, 766, 753, 754, 755, 756, 757, 758, 759, 760, 761,
762, 763, 764, 765, 766, 767, 753, 754, 755, 756, 757, 758, 759,
760, 761, 762, 763, 764, 765, 766, 767, 753, 754, 755, 756, 757,
758, 759, 760, 761, 762, 763, 753, 754, 755, 756, 757, 758, 754]))
```

In the last step, we'll use these indices to extract the values of the data from the fishing raster, and sum them to get a total distance travelled.

```
np.nansum(ship_arr[r40_index])
```

```
8387952.0
```

Now that we know the individual steps, let's run this over all of the districts. For each district (with `geometry` and `id`), we run the `features.rasterize` function. If any values equal 1 (some of the districts are outside the bounds of the raster), we calculate the sum of the values of the shipping raster `r_array` based on the indices in the raster where the district is located.

```
distance_dict = {}
for geom, idx in zip(comm_3338['geometry'], comm_3338['id']):
    rasterized = features.rasterize(geom.geoms,
                                    out_shape=shape,
                                    transform=transform,
                                    all_touched=True,
                                    fill=0)
    # only save polygons that have a non-zero value
    if any(np.unique(rasterized)) == 1:
        r_index = np.where(rasterized == 1)
        distance_dict[idx] = np.nansum(ship_arr[r_index])
```

Now we just create a data frame from that dictionary, and join it to the vector data using `pandas` operations.

```
# create a data frame from the result
df = pd.DataFrame.from_dict(distance_dict,
    orient='index',
    columns=['distance'])

# extract the index of the data frame as a column to use in a join
df['id'] = df.index
```

Now we join the result to the original geodataframe.

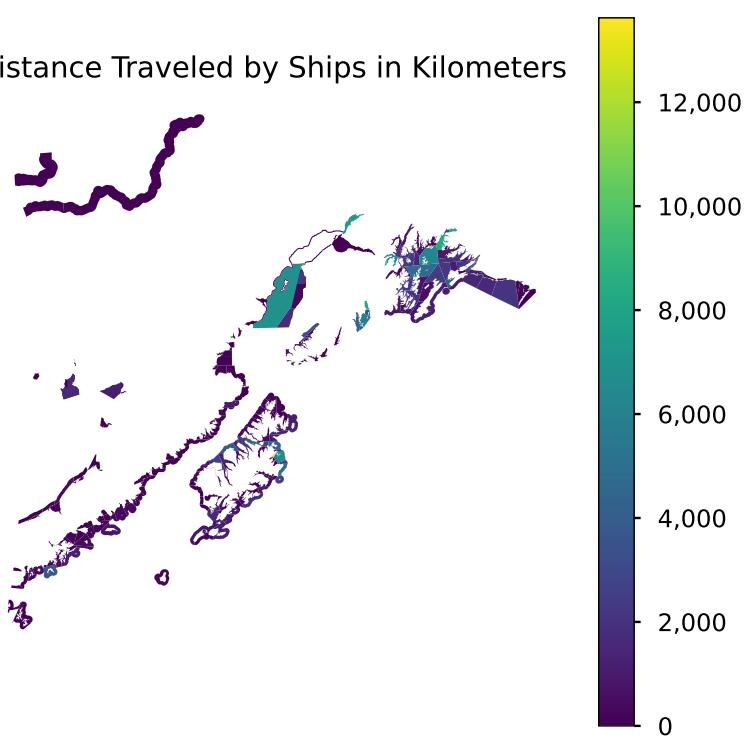
```
# join the sums to the original data frame
res_full = comm_3338.merge(df, on = "id", how = 'inner')
```

Finally, we can plot our result!

```
fig, ax = plt.subplots(figsize=(7, 7))
plt.style.use("seaborn-talk")
ax = res_full.plot(column = "distance", legend = True, ax = ax)
fig = ax.figure
cb_ax = fig.axes[1]
cb_ax.set_yticklabels(["0", "2,000", "4,000", "6,000", "8,000", "10,000", "12,000", "14,000"])
ax.set_axis_off()
ax.set_title("Distance Traveled by Ships in Kilometers")
plt.show()
```

```
/tmp/ipykernel_52923/3303410610.py:6: UserWarning: FixedFormatter should only be used together
  cb_ax.set_yticklabels(["0", "2,000", "4,000", "6,000", "8,000", "10,000", "12,000", "14,000"])
```

Distance Traveled by Ships in Kilometers



# 12 Data Futures: Parquet and Arrow

- The difference between column major and row major data
- Speed advantages to columnar data storage
- How `arrow` enables faster processing

## 12.1 Introduction

System calls are calls that are run by the operating system within their own process. There are several that are relevant to reading and writing data: open, read, write, seek, and close. Open establishes a connection with a file for reading, writing, or both. On open, a file offset points to the beginning of the file. After reading or writing `n` bytes, the offset will move `n` bytes forward to prepare for the next operation. Close closes the connection to the file. Read will read data from the file into a memory buffer, and write will write data from a memory buffer to a file. Seek is used to change the location of the offset pointer, for either reading or writing purposes.

If you've worked with even moderately sized datasets, you may have encountered an "out of memory" error. Memory is where a computer stores the information needed immediately for processes. This is in contrast to storage, which is typically slower to access than memory, but has a much larger capacity. When you `open` a file, you are establishing a connection between your processor and the information in storage. On `read`, the data is read into memory that is then available to your python process, for example.

So what happens if the data you need to read in are larger than your memory? 32GB is a common memory size, but this would be considered a modestly sized dataset by this course's

standards. There are a number of solutions to this problem, which don't involve just buying a computer with more memory. In this lesson we'll discuss the difference between row major and column major file formats, and how leveraging column major formats can increase memory efficiency. We'll also learn about another python library called `pyarrow`, which has a memory format that allows for "zero copy" read.

## 12.2 Row major vs column major

The difference between row major and column major is in the ordering of items in the array.

Take the array:

```
a11 a12 a13  
a21 a22 a23
```

This array in a row-major order would be read in as:

```
a11, a12, a13, a21, a22, a23
```

You could also read it in column-major order as:

```
a11, a21, a12, a22, a13, a33
```

By default, C and SAS use row major order for arrays, and column major is used by Fortran, MATLAB, R, and Julia.

Python uses neither, instead representing arrays as lists of lists, though `numpy` uses row-major order.

### 12.2.1 Row major versus column major files

The same concept can be applied to file formats as the example with arrays above. In row-major file formats, the values (bytes) of each record are read sequentially.

| Name    | Location   | Age |
|---------|------------|-----|
| John    | Washington | 40  |
| Mariah  | Texas      | 21  |
| Allison | Oregon     | 57  |

In the above row major example, data are read in the order: John, Washington, 40, [new line], Mariah, Texas, 21.

This means that getting a subset of all columns would be easy; you can specify to read in only the first X rows. However, if we are only interested in `Name` and `Location`, we would still have to read in all of the rows before discarding the `Age` column.

If these data were organized in a column major format, they might look like this:

```
Name: John, Mariah, Allison
Location: Washington, Texas, Oregon
Age: 40, 21, 57
```

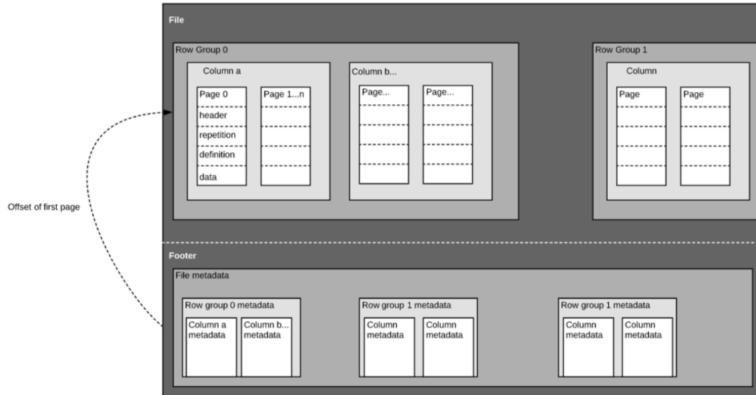
And the read order would first be the names, then the locations, then the age. This means that selecting all values from a set of columns is quite easy (all of the Names and Ages, or all Names and Locations), but reading in only the first few records from each column would require reading in the entire dataset. Another advantage to column major formats is that compression is more efficient since compression can be done across each column, where the data type is uniform, as opposed to across rows with many data types.

## 12.3 Parquet

Parquet is an open-source file format that stores data in a column-major format. The format contains several key components:

- row group
- column

- page
- footer



Row groups are blocks of data over a set number of rows that contain data from the same columns. Within each row group, data are organized in column-major format, and within each column are pages that are typically 1MB. The footer of the file contains metaata like the schema, encodings, unique values in each column, etc.

The parquet format has many tricks to increase storage efficiency, and is increasingly being used to handle large datasets.

## 12.4 Arrow

So far, we have discussed the difference between organizing information in row-major and column-major format, how that applies to arrays, and how it applies to data storage on disk using Parquet.

Arrow is a language-agnostic specification that enables representation of column-major information in memory without having to serialize data from disk. The Arrow project provides implementation of this specification in a number of languages, including Python.

Let's say that you have utilized the Parquet data format for more efficient storage of your data on disk. At some point, you'll need to read that data into memory in order to do analysis on it. Arrow enables data transfer between the on disk Parquet files and in-memory Python computations, via the `PyArrow` library.

`PyArrow` is great, but relatively low level. It supports basic group by and aggregate functions, as well as table and dataset joins, but it does not support the full operations that `pandas` does.

## 12.5 Example

In this example, we'll read in a dataset of fish abundance in the San Francisco Estuary, which is published in csv format on the [Environmental Data Initiative](#). This dataset isn't huge, but it is big enough (3 GB) that working with it locally can be fairly taxing on memory. Motivated by user difficulties in actually working with the data, the [deltafish R](#) package was written using the R implementation of `arrow`. It works by downloading the EDI repository data, writing it to a local cache in parquet format, and using `arrow` to query it. In this example, I've put the Parquet files in a sharable location so we can explore it using `pyarrow`.

First, we'll load the modules we need.

```
import pyarrow.dataset as ds
import numpy as np
import pandas as pd
```

Next we can read in the data using `ds.dataset()`, passing it the path to the parquet directory and how the data are partitioned.

```
deltafish = ds.dataset("/home/jclark/deltafish/fish", format="parquet", partitioning=[ "Speci
```

You can check out a file listing using the `files` method. Another great feature of parquet files is that they allow you to

partition the data across variables of the dataset. These partitions mean that, in this case, data from each species of fish is written to its own file. This allows for even faster operations down the road, since we know that users will commonly need to filter on the species variable. Even though the data are partitioned into different files, `pyarrow` knows that this is a single dataset, and you still work with it by referencing just the directory in which all of the partitioned files live.

```
deltafish.files
```

```
['/home/jclark/deltafish/fish/Taxa=Acanthogobius flavimanus/part-0.parquet',
 '/home/jclark/deltafish/fish/Taxa=Acipenser medirostris/part-0.parquet',
 '/home/jclark/deltafish/fish/Taxa=Acipenser transmontanus/part-0.parquet',
 '/home/jclark/deltafish/fish/Taxa=Acipenser/part-0.parquet'...]
```

You can view the columns of a dataset using `schema.to_string()`

```
print(deltafish.schema.to_string())
```

```
SampleID: string
Length: double
Count: double
Notes_catch: string
Species: string
```

If we are only interested in a few species, we can do a filter:

```
expr = ((ds.field("Species")=="Dorosoma petenense") |
         (ds.field("Species")=="Morone saxatilis") |
         (ds.field("Species")=="Spirinchus thaleichthys"))

fishf = deltafish.to_table(filter = expr)
```

There is another dataset included, the survey information. To do a join, we can just use the `join` method on the `arrow` dataset.

First read in the survey dataset.

```
survey = ds.dataset("/home/jclark/deltafish/survey", format="parquet", partitioning=["SourceID"])

print(survey.schema.to_string())
```

Then do the join, and convert to a pandas `data.frame`.

```
fish_j = fishf.join(survey, "SampleID").to_pandas()
```

Note that when we did our first manipulation of this dataset, we went from working with a `FileSystemDataset`, which is a representation of a dataset on disk without reading it into memory, to a `Table`, which is read into memory. `pyarrow` has a [number of functions](#) that do computations on datasets without reading them into memory. However these are evaluated “eagerly,” as opposed to “lazily.” These are useful in some cases, like above, where we want to take a larger than memory dataset and generate a smaller dataset (via filter, or group by/summarize), but are not as useful if we need to do a join before our summarization/filter.

More functionality for lazy evaluation is on the horizon for `pyarrow` though, by leveraging [Ibis](#).

## **13 Software Design II**

## **14 Group Project: Data Processing**

Groups will start with the tiled output that they made from last session. The main objective in this lesson is to have them rasterize the tiled shapfiles, which we need to do for visualization (and would also be a useful thing to know how to do in an analysis workflow as well).

The groups will need to:

- run the rasterizer method on one file
- write a parsl app that will run it on all of the files

## **15 Data Ethics**

# 16 Google Earth Engine

SAM NOTES, DELETE LATER - need to make sure that .ipynb that student's will work out of are running in same virtual enviroment as everything else - embed .ipynb into quarto notebook (get book to build from those examples) - use Ryan Abernathey's [post](#) to help frame introduction

- Understand what Google Earth Engine provides and its applications
- Learn about some real-world applications of Google Earth Engine
- Learn how to get started using Google Earth Engine on your own computer
- Learn how to find and access Google Earth Engine Data

## 16.1 Introduction (15-20min)

SAM NOTES, DELETE LATER - have Ingmar help frame utility of GEE in intro

[Google Earth Engine](#) is a geospatial processing platform powered by Google Cloud Platform. It contains over 30 years of satellite imagery and geospatial datasets that are continually updated and available instantly. The Earth Engine API is available in Python (and JavaScript) for anyone with an account to access and analyze data.

ADD SOME COOL IMAGERY HERE

Explore the public [Earth Engine Data Catalog](#) which includes a variety of standard Earth science raster datasets. Browse by [dataset tags](#) or by satellites ([Landsat](#), [MODIS](#), [Sentinel](#)).

SAM NOTES, DELETE LATER - typical: download data and work locally - new way: "Moving compute to data" – GEE is a great example of this

## 16.2 Getting started with Google Earth Engine (GEE) on your own machine (40 min)

SAM NOTES, DELETE LATER

- remove `#| eval: false` once code actually runs to embed outputs
- don't really have time to actually do this in class? include instructions for those who want to try it out on their own later?
- Content borrowed from Dr. Samantha Stevenson's [guide to installing Jupyter/Google Earth Engine on your personal laptop](#).
- PREREQUISITE: need conda (see Earth Engine API installation instructions [here](#)) – haven't added these instructions below yet
- also, not working on my machine at the moment. issue with library installs? starts with `earthengine authenticate`

### 1. Install the Google Earth Engine API

SAM NOTES, DELETE LATER - MOVE MOST OF THIS TO COURSE SETUP (not using conda) BUT KEEP GEE ACTIVATION STEP HERE

- Create an environment where the Google Earth Engine API will live. This ensures that it and its dependent packages will not cause versioning issues with your base environment (or other environments). We'll call our environment `gee_env`.

```
#| eval: false
conda create --name gee_env
```

- Activate your environment so your machine knows where to store subsequent installs.

```
#| eval: false
conda activate gee_env
```

You'll know your environment is activated successfully when `(gee_env)` appears before the prompt in your terminal window (as opposed to `(base)`, for example).

- Install the Google Earth Engine API in your `gee_env`

```
#| eval: false
conda install -c conda-forge earthengine-api
```

## 2. Sign up for a GEE Account

GEE is currently free for educational use. Sign up for an account at <https://signup.earthengine.google.com> (you'll need this to authenticate in the next step).

## 3. Set up GEE Authentication

In order to begin using GEE, you'll need to connect your GEE environment (`gee_env`) to the authentication credentials associated with your Google account. This will need to be done each time you connect to GEE, but should only be done once per session.

- On the command line, type:

```
#| eval: false
earthengine authenticate
```

This should launch a browser window where you can login with your Google account to the Google Earth Engine Authenticator. Following the prompts will generate a code, which you'll then need to copy and paste back onto the command line. This will be saved as an authentication token so you won't need to go through this process again until the next time you start a new session.

4. Install necessary packages (if you don't already have them)

```
#| eval: false
pip install ee # Earth Engine API package
pip install geemap # package for interactive mapping with GEE
pip install pandas # contains useful tools for data manipulation (may not need this)
```

## 16.3 Visualize global precipitation data using Google Earth Engine

*Content for this section was adapted from Dr. Sam Stevenson's [Visualizing global precipitation using Google Earth Engine](#) lesson, given in her [EDS 220 course](#) in Fall 2021.*

1. Import necessary packages

```
import ee # MODULENOTFOUNDERROR
import geemap
import pandas as pd
```

2. Create an interactive basemap

The default basemap is (you guessed it) Google Maps. The following code displays an empty Google Map that you can manipulate just like you would in the typical Google Maps interface. Do this using the `Map` method from the `geemap` library. We'll also center the map at a specified latitude and longitude (here, 40N, 100E), set a zoom level, and save our map as an object called `myMap`.

```
myMap = geemap.Map(center = [40, -100], zoom = 2)
myMap
```

3. Load ERA5 Image Collections from GEE

- NOTE: ADC has worked with these data – took 3 weeks to download

- EE collection is all you need to load and analyze image collection
- precursor to Ingmar's stuff

We'll be using the ERA5 daily aggregates reanalysis dataset, produced by the European Centre for Medium-Range Weather Forecasts (ECMWF), found [here](#), which models atmospheric weather observations. We'll load the `total_precipitation` field (check out the dataset metadata on [here](#)).

The `ImageCollection` method extracts a set of individual images that satisfies some criterion that you pass to GEE through the `ee` package. This is stored as an `ImageCollection` object which can be filtered and processed in various ways. We can pass the `ImageCollection` method arguments to tell GEE which data we want to retrieve. Below, we retrieve all daily ERA5 data (so we can see individual rain events).

```
weatherData = ee.ImageCollection('ECMWF/ERA5/DAILY')
```

#### 4. Select an image to plot

To plot a map over our Google Maps basemap, we need an “Image” rather than an “ImageCollection.” ERA5 contains many different climate variables, so we need to pick what we'd like to plot. We'll use the `.select` method to choose the parameter(s) we're interested in from our `weatherData` object.

```
precip = weatherData.select("total_precipitation")
```

We can look at our `precip` object using the `print` method to see that it's still an “ImageCollection” which contains daily information from 1979 to 2020.

```
print(precip)
```

We want to filter it down to a single field for a time of interest – let's say December 1-2, 2019. We apply the `.filter` method to our `precip` object and apply the `ee.Filter.date` method (from the `ee` package) to filter for data from our chosen date range. We also apply the `.mean` method, which takes whatever precedes it and calculates the average.

```
precip_filtered = precip.filter(ee.Filter.date('2019-12-01', '2019-12-02')).mean()
```

### 5. Add data to map

We can first use the `setCenter` method to tell the map where to center itself. It takes the longitude and latitude as the first two coordinates, followed by the zoom level.

```
Map.setCenter(-152.505706, 59.432367, 2) # Cook Inlet, Alaska (WE CAN CHANGE THIS LOCATION)
```

Next, set a color palette to use when plotting the data layer. The following is a palette specified for precipitation in the GEE description page for ERA5. GEE has lots of color tables like this that you can look up.

```
precip_palette = {  
    'min':0,  
    'max':0.1,  
    'palette': ['#FFFFFF', '#00FFFF', '#0080FF', '#DA00FF', '#FFA400', '#FF0000']  
}
```

Finally, plot our filtered data, `precip_filtered` on top of our basemap using the `.addLayer` method. We'll also pass it our visualization parameters (colors and ranges stored in `precip_palette`, the name of the data field `total precipitation`, and opacity so that we can see the basemap underneath)

```
Map.addLayer(precip_filtered, precip_palette, 'total precipitation', opacity = 0.3)
```

## 16.4 INGMAR'S DEMONSTRATION HERE?(30-40 min)

## 16.5 Conclusion/Summary

- lessons learned
- utilities
- etc.

## 16.6 Other Resources

- [GEE Code Editor](#) is a web-based IDE for using GEE (JavaScript)

## **17 Group Project: Visualization**

Groups will start with razertized tiles from the second group project lesson.

- again, implement the PDG method
- display using either leaflet (folium) or cesium?
- not sure where cesium will be deployed

# **18 Workflows for data staging and publishing**

- NSF archival policies for large datasets
- Data transfer tools
- How to manage co-locating data and code
  - eg: where model runs only has 1 TB of storage but model outputs 10 TB of data
  - workflow tools (pegasus, condor, slurm, snakemake)
- Uploading large datasets to the Arctic Data Center

## **18.1 NSF policy for large datasets**

- there are many different research methods that can generate large volumes of data. Numerical modeling (such as climate or ocean models) and anything generating high resolution imagery are two examples we see very commonly.

The Office of Polar Programs policy requires that metadata files, full data sets, and derived data products, must be deposited in a long-lived and publicly accessible archive.

Metadata for all Arctic supported data sets must be submitted to the NSF Arctic Data Center (<https://arcticdata.io>).

Exceptions to the above data reporting requirements may be granted for social science and indigenous knowledge data, where privacy or intellectual property rights might take precedence.

Such requested exceptions must be documented in the Data Management Plan.

- datasets that are already published on a long lived archive do not need to be replicated to the Arctic Data Center
  - example: a research project accesses many terabytes of VIIRS satellite data. The original satellite data does not need to be published on the Arctic Data Center, but the code that accessed it, and derived products, can be published
- for some numerical models, if the model results can be faithfully reproduced from code, the code that generates the models can be a sufficient archival product, as opposed to the code and the model output
  - if the model is difficult to set up, or takes a very long time to run, we would probably recommend publishing the output as well as code
- the Arctic Data Center is committed to archiving data of any volume

## 18.2 Data transfer tools

- scenario: you need to send a bunch of data to the Arctic Data Center. after getting the credentials, you use `scp` to start the transfer. You know this typically takes around 12 hours so you start it at 5pm right when you leave the office expecting it to be done when you get back. When you arrive, you see there was a short network outage in the middle of the night. The whole job failed so you have to start it again...

There is a better way!

Three key elements to data transfer

- `endpoints`
- `network`
- `transfer tool`

### **18.2.0.1 Endpoints**

The from and to locations of the transfer, an endpoint is a remote computing device that can communicate back and forth with the network to which it is connected. The speed with which an endpoint can communicate with the network varies depending on how it is configured. Performance depends on the CPU, RAM, OS, and disk configuration. Examples:

- NCEAS `datateam` server:
- Standard laptop

### **18.2.0.2 Network speed**

Determines how quickly information can be sent between endpoints, largely dependent on what you pay for. Wired networks get significantly more speed than wireless.

- not all networks are created equal
- server to server (north hall to san diego) versus server to your house

### **18.2.0.3 Transfer tools**

- `scp`
  - uses `ssh` for authentication and transfer
  - if you can `ssh` to a server, you can probably use `scp` to move files without any other setup
  - copies all files linearly and simply. if a transfer fails in the middle, difficult to know exactly what files didn't make it, so you have to start the whole thing over and re-transfer all the files
- `rsync`
  - similar to `scp` but syncs files/directories as opposed to copying
  - if the file already exists on the other side, it is skipped
- `globus`

- parallelizes transfers by utilizing multiple network sockets simultaneously
- is able to fail and restart itself efficiently
- requires more setup, endpoints need to be configured as globus nodes

#### **18.2.0.4 Globus**

- easy to use, as long as your data are accessible via an endpoint configured as a Globus node
- leverage your institutions computing resources! they may be able to help get you access to a data transfer node already configured correctly
- there are paid options to set up a node from your own workstation (Globus Connect Personal - check the naming here, and feature list)
  - remember the other factors though! Globus won't help you overcome a 1 Gb/s laptop connection speed, or a 50 Mb/s network speed

### **18.3 Documenting large datasets**

- the Arctic Data Center is working to support large datasets, but we have performance considerations as well
- self documenting file formats are preferred, to prevent us from needing to document thousands-millions of files in a single metadata document
  - netcdf
  - geotiff, geopackage
- regular, parseable filenames and consistent file formatting is key
- communicate early and often with the Arctic Data Center staff

## **19 What is Cloud Computing Anyways?**

# 20 Reproducibility and Containers

- TODO: Decide about if/how to talk about WholeTale
- TODO: This lesson should be have a wow-factor and emphasize why we're focusing all of this
- TODO: This lesson should be more about wrapping up and tying everything together than showing off new tech
- ~~Learn about software versioning~~
- Become familiar with Docker as a tool to improve computational reproducibility

## 20.1 Outline

- Introduce software reproducibility
  - Motivate the idea with examples and data
  - Talk about software collapse
    - \* <http://blog.khinsen.net/posts/2017/01/13/sustainable-software-and-reproducible-research-dealing-with-software-collapse/>
    - \* <https://xkcd.com/2347/>
- Semantic versioning and the reality of it e.g.,  
<https://pandas.pydata.org/docs/development/policies.html#version-policy>
- MyBinder
- WholeTale?

Examples to look at including:

- <https://numpy.org/neps/nep-0023-backwards-compatibility.html#example-cases>
- <https://github.com/scipy/scipy/issues/16418> > <https://pandas.pydata.org/docs/whatsnew/v1.4.0.html#deprecations>  
DataFrame.append() and Series.append() have been deprecated and will be removed in a future version. Use pandas.concat() instead (GH35407).

Principles to get across:

1. You probably should be thinking about software versioning
  - Know which version of Python your code was written/tested under and keep track of that in a machine-readable way
  - Know the specific versions, of at least the specific MAJOR.MINOR of the packages your code was written+tested under and keep track of them in a machine-readable way (ie requirements.txt)

## 20.2 Hands-off Demo

Show students an example of containerizing a workflow so it runs using a past version of Python and pinned versions of packages. Ideally find an example where behavior changes based on the Python or one or more package versions.