# EEPS 440/640

# Geospatial Data Science (GDS)

# Introduction

- Instructor: Dr. Noemi Vergopolan

# Motivation

There are big environmental challenges that we want to take on…

Here are some examples from Hydrology

# Agriculture

When should I irrigate my crops? And how much?

Where should the irrigation water come from?

What type of irrigation system should I use?

# Floods

Where will it flood during a hurricane?

How can we mitigate the impacts of floods?

What is the risk that my house is flooded?

# Water management

Where should I place a reservoir?

How much water should I release?

How will a changing climate impact water inflows into existing reservoirs?

# Wildfires

Which conditions trigger wildfires?

Are wildfires caused by climate change?

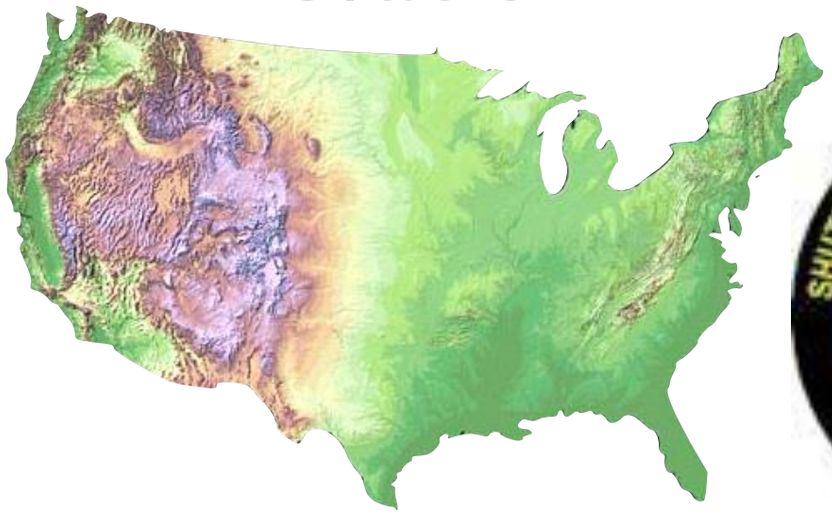
What is the risk that my house is burned?

How do we address those questions?

Data, data, and more data

# I. Environmental Data

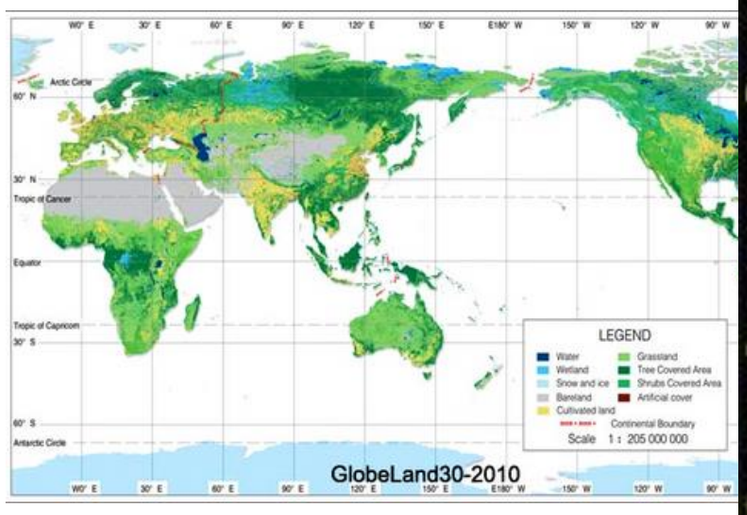
## Elevation

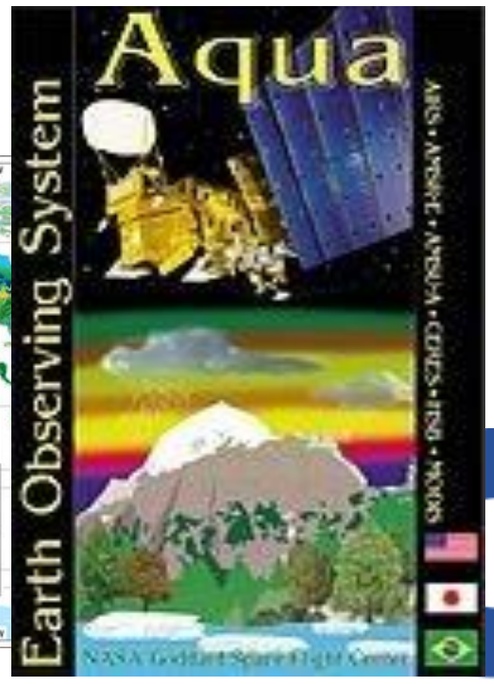


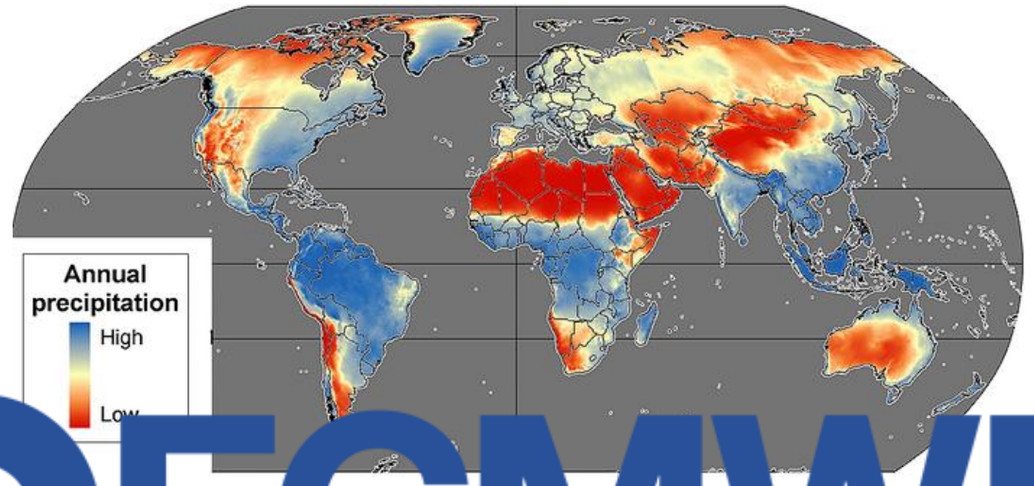

## Soil



## Land Cover




ECMWF
EUROPEAN CENTRE FOR MEDIUM RANGE WEATHER FORECASTS

# I. Environmental Data

Elevation

Soil

Land Cover

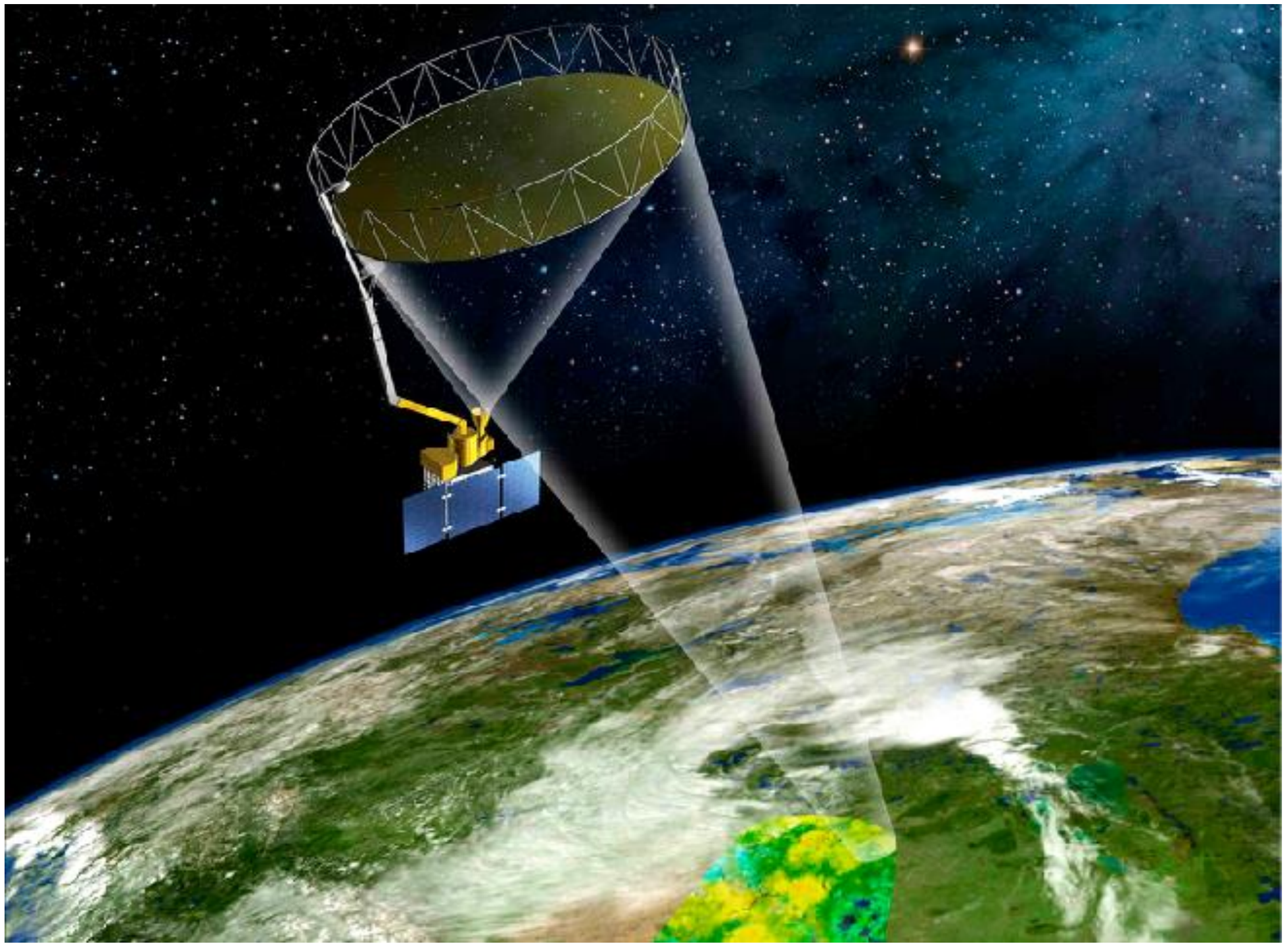
Climate

"Age of Data"

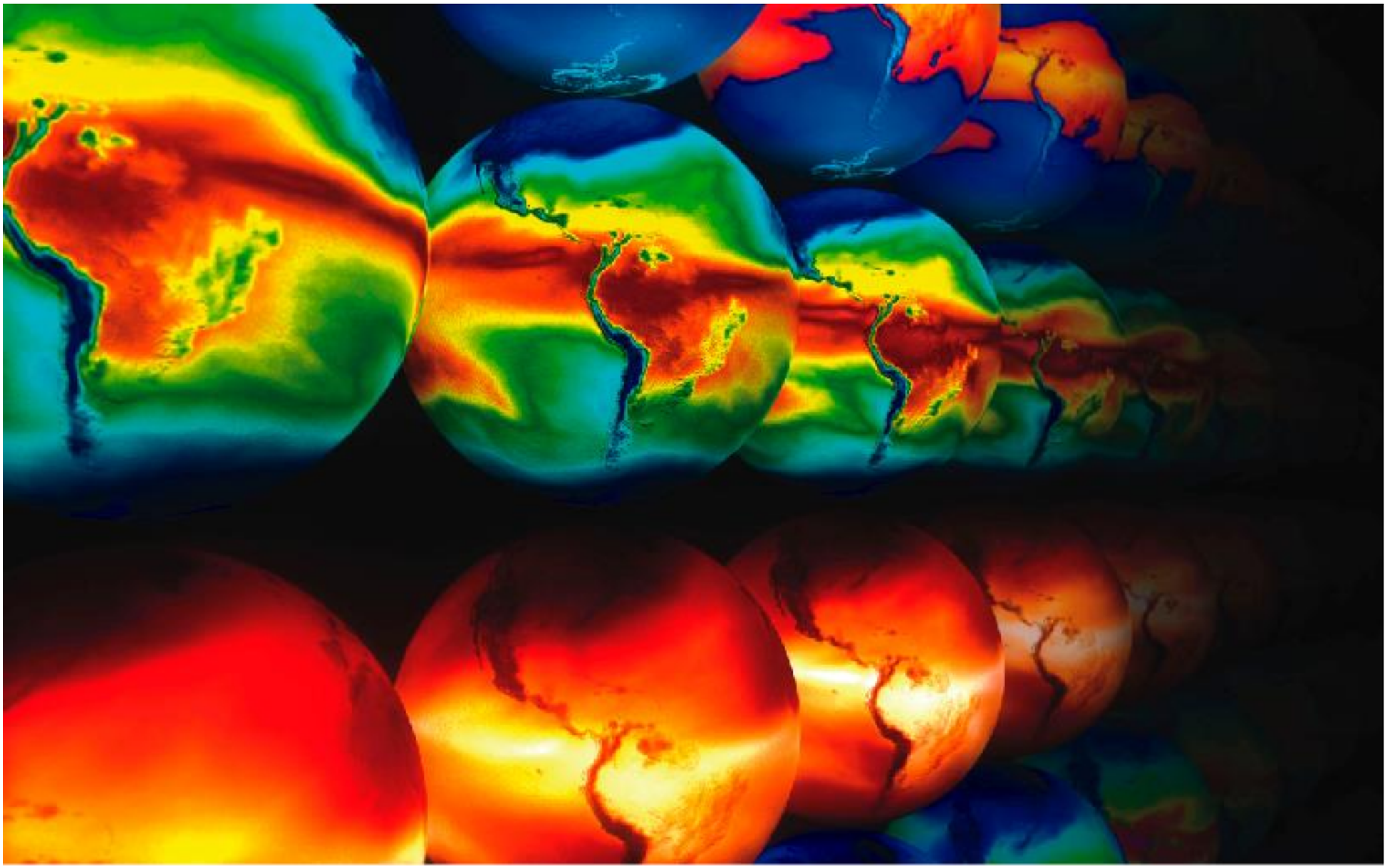
# What Petabytes of data feels like

# And the problem is even much worse with geospatial data

# Example: Satellite remote sensing



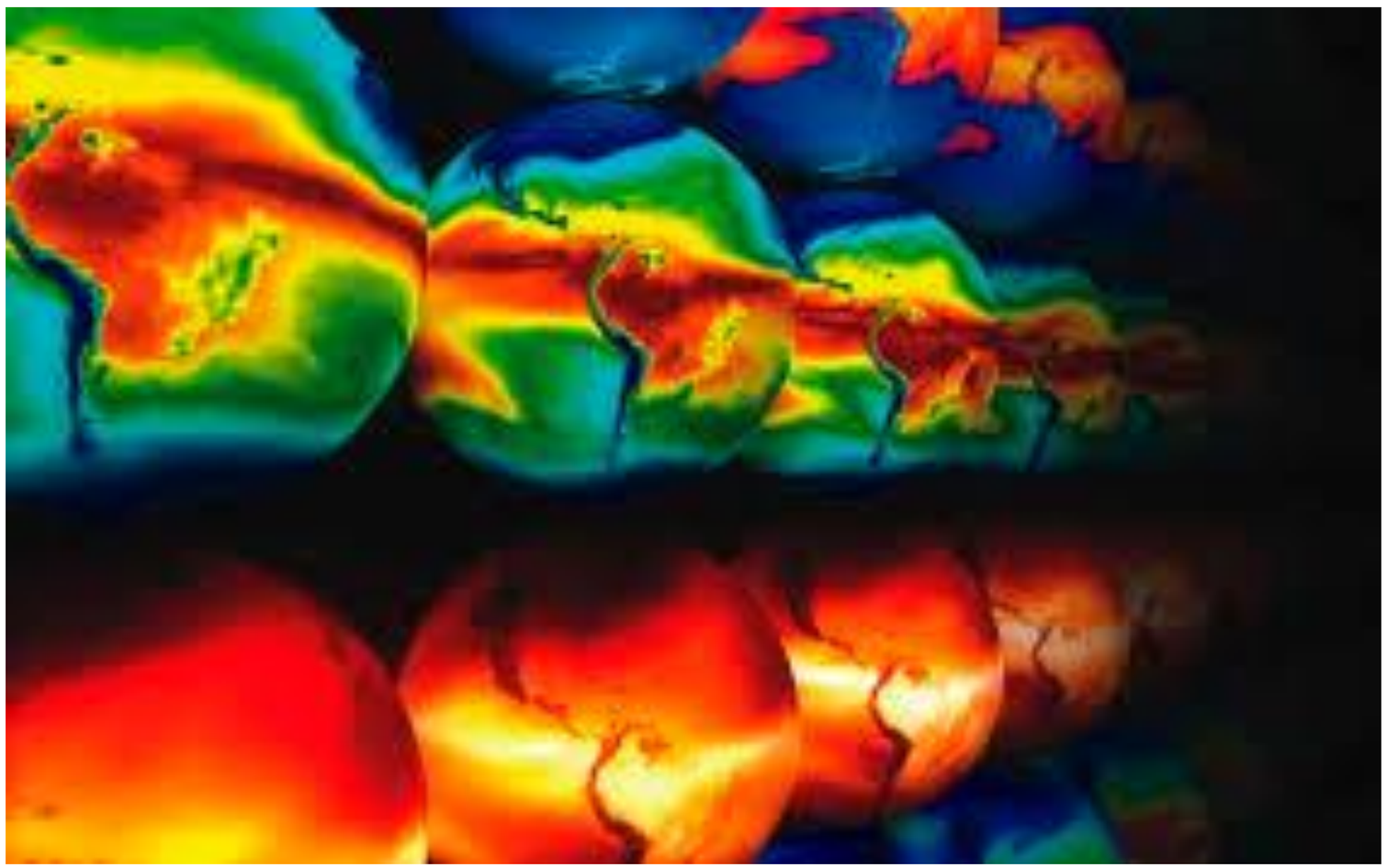
# Example: Climate model output



This course aims to give you the tools and knowledge on how to tackle these data

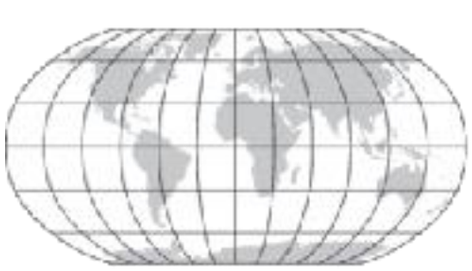
# What topics will we cover?

# Topics: Analyze spatial environmental data



How can we analyze terabytes/petabytes of climate model output and/or satellite data?

# Topics: Map projections



Robinson

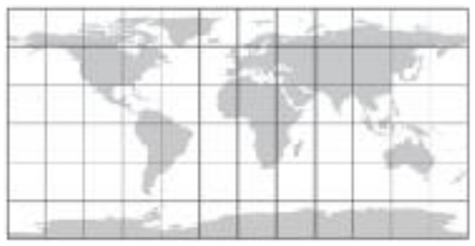
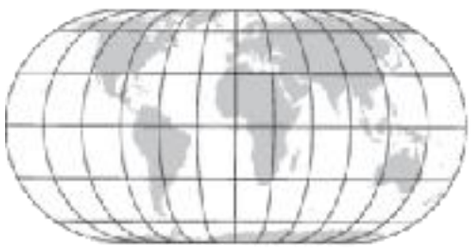


Plate Carrée

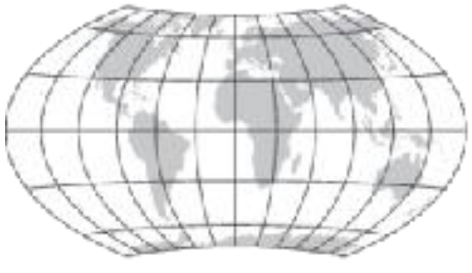
Winkel Tripel



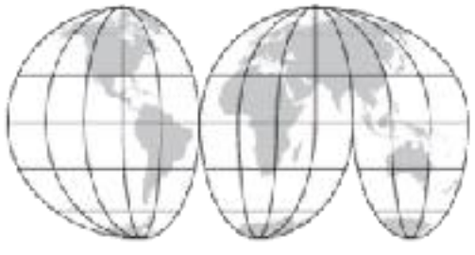
Eckert IV

Mollweide

Mercator



Wagner VII

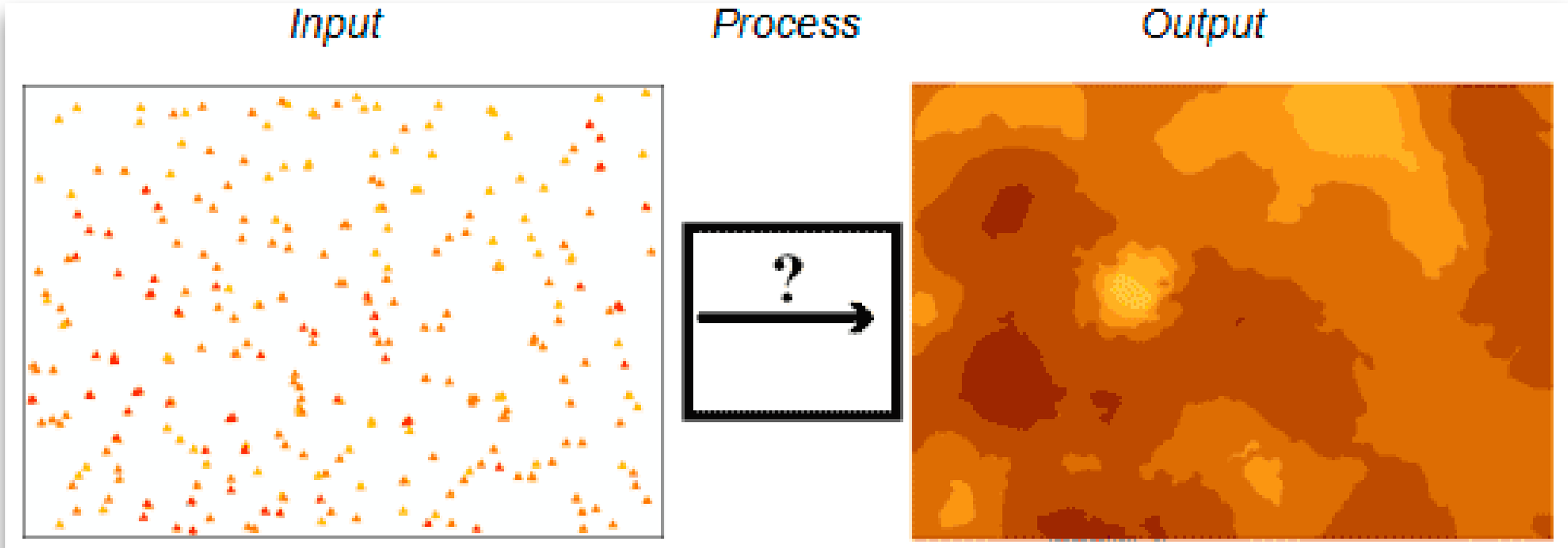


Interrupted Mollweide

Goode Homolosine

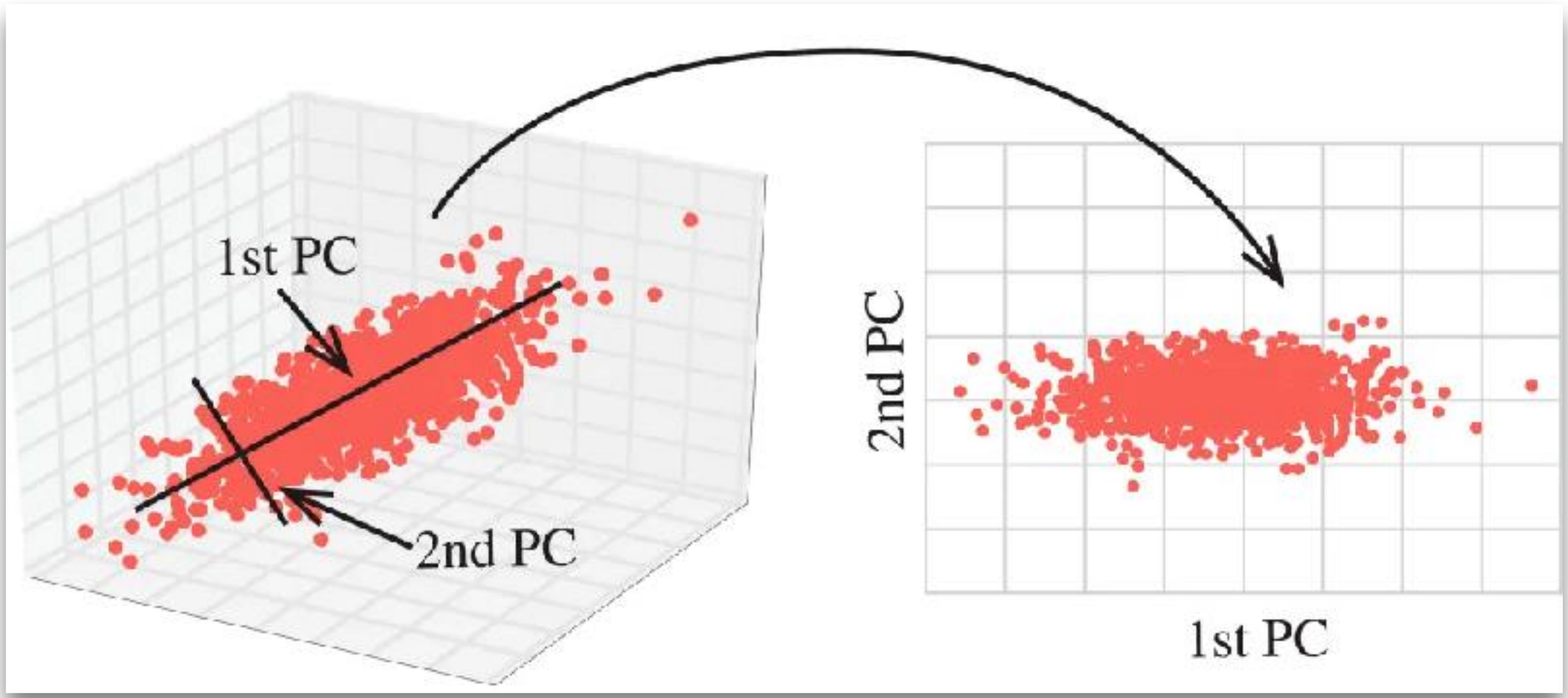
How can we transform a map to a different projection?

# Topics: Spatial interpolation/ prediction




How do you create a spatial map from point data?

# Topics: Reduce dimensions




How do we compress data?

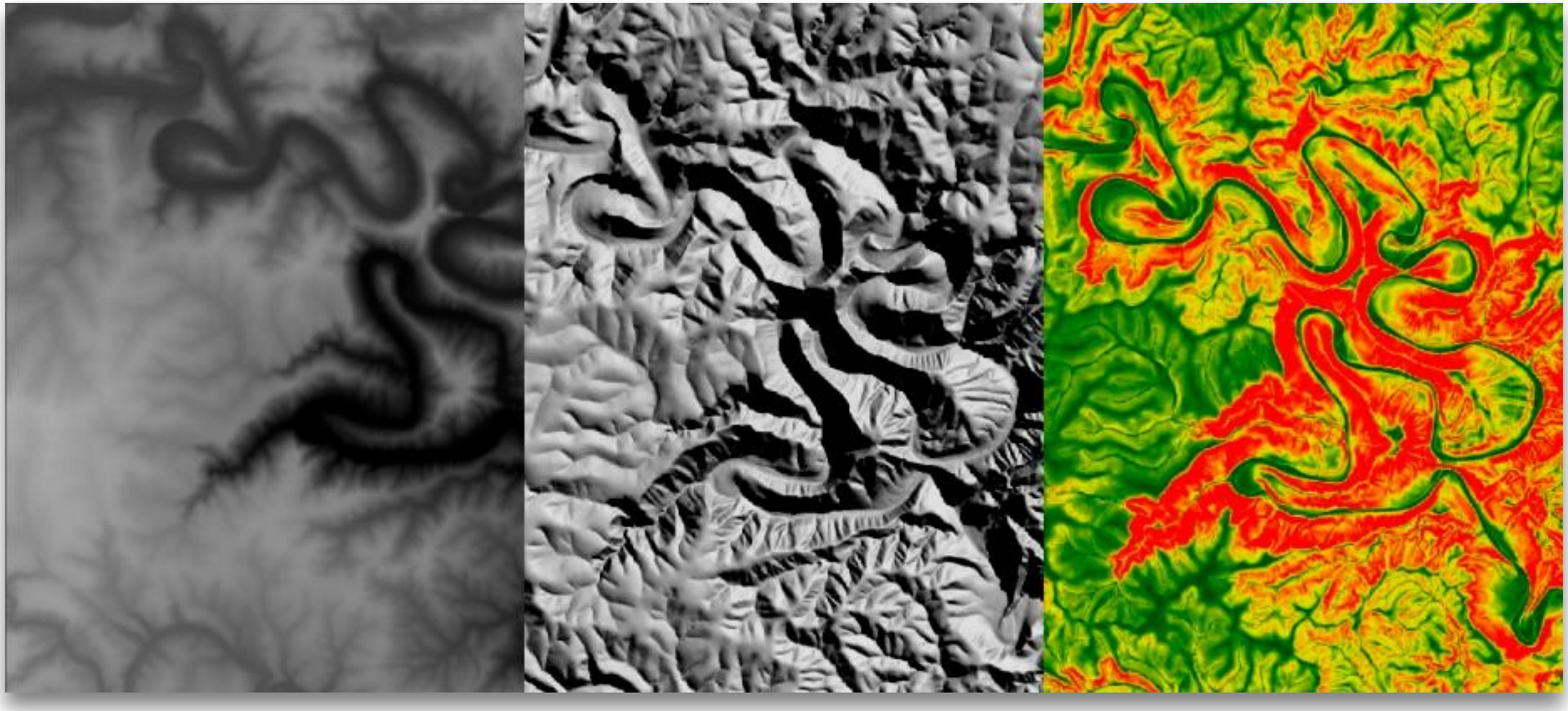
What are the most important features in the data?

# Topics: Terrain Analysis

Elevation

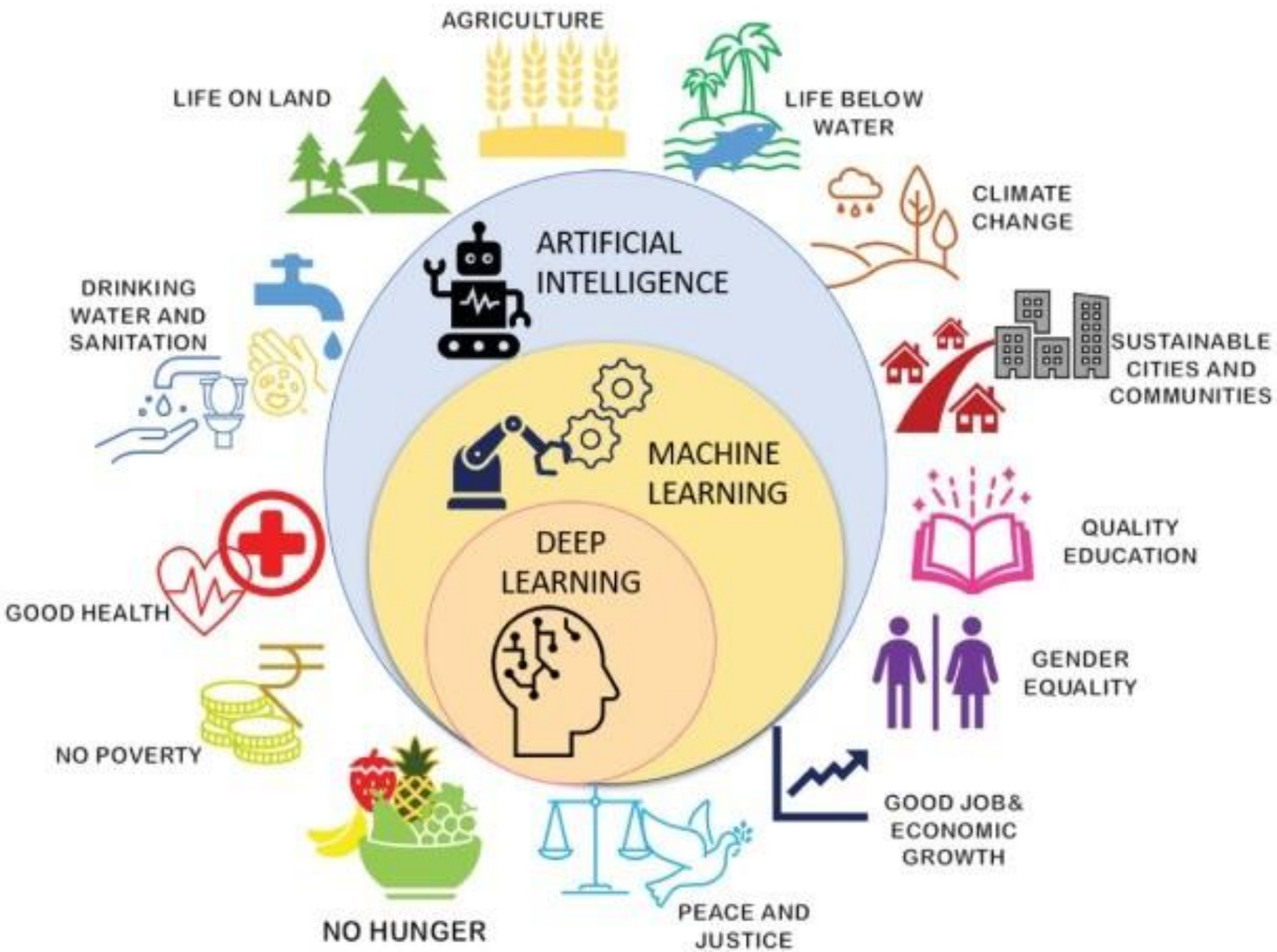
Hillshading

Slope



What can we learn from elevation data?

# Integration of ML methods for geospatial data science



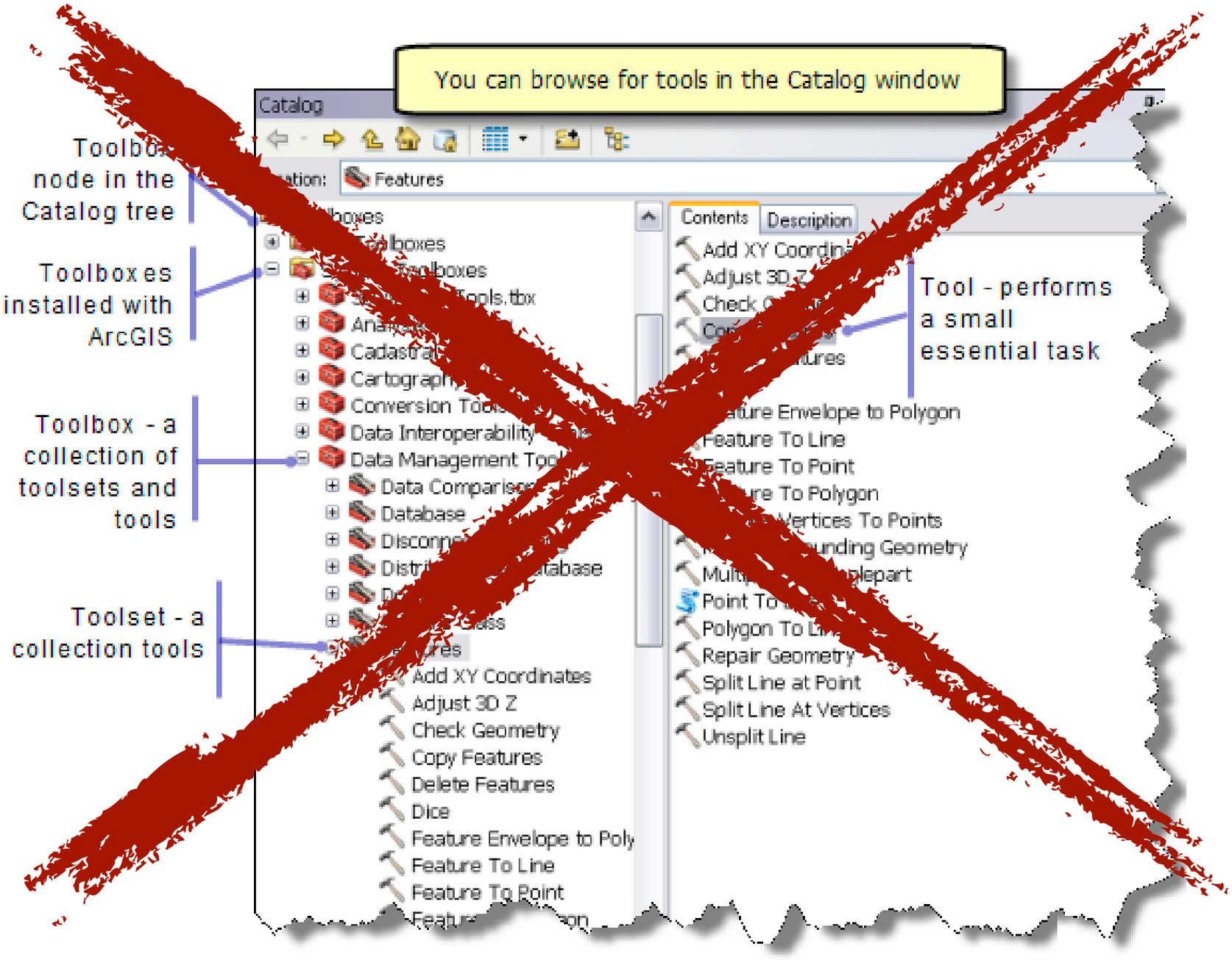
# And more…

But what tools will we use to learn how to create and analyze these large datasets?

Let’s start off with what we will NOT be using this semester

# This is NOT an ArcGIS course

# This is NOT a "click the button" class




If that was what you were expecting then this may not be the class for you

If you want to learn ArcGIS Rice has resources for you:

GIS/Data Center (GDC)

https://library.rice.edu/places/gisdata-center-gdc

https://wiki.rice.edu/confluence/x/QIAKAg

# Then what will we use?

# Programming, programming, and more programming

If you are terrified of programming…

Don’t be! It is worth the time investment.

However, if you are not willing to dedicate the time and just want to “pass by” then this class is probably not a good fit.

# Which programming languages will we use?

# Primary language: Python


Python level:  Intermediate

# Secondary language: Bash


Bash level: Beginner

# Don’t be afraid…

- We are going to spend the first third of the class learning how to program in Python (and Bash)
- It will get hard, but it will be additive so one can take this class without a background in these programming languages

# Rice Computing Resources

## Python Programming Series

https://library.rice.edu/services/data-workshops

Intro to programming – Jan. 24, 1-3 pm

Loops and conditions – Jan. 31, 1-3 pm

Data analysis and Viz w/ Pandas – Feb. 21, 1-2:30 pm

Text analysis – Apr. 8, 1-2:30 pm

## Bash - The World is at your Command….line:

https://library.rice.edu/courses/world-your-commandline

Jan. 29, 1-4 pm

# Syllabus

# Course Logistics: General Info

## Course Information

Course time: Tu-Th 4–5:15pm
Course classroom: KWG 128

## Contact Information

**Instructor:** Dr. Noemi Vergopolan
**Office:** KWGL 301
**Email:** nv25@rice.edu
**Office hours:** by reservation (calendly.com/noemi-vergopolan) Fridays 10am-12pm

# Course Logistics: General Info

## Grade Policies

The final grade for **undergraduates** will comprise:
10% Class Participation
40% Homeworks
50% Final Project: Proposal (10%), Presentation (10%), and Paper (30%)

The final grade for **graduate** students will comprise:
10% Class Participation
30% Homeworks
60% Final Project: Proposal (10%), Presentation (10%), and Paper (40%)

# Homework

- Completed assignments will be submitted via each student's private GitHub repository
- Each assignment must be submitted before class on the day listed on the schedule

There are no textbooks required

I am are here to help.

That said, Google and ChatGPT will answer almost all questions much faster (Use it!).

# Participation

- Discussion/questions
- Avoid the one-way lecture approach, which relies on your involvement
- Bring your laptop to class to follow along with the interactive lectures

# Final Project

**Components**:

- A 1-page project proposal.
- An ~8-10 page final project paper.
- A final project presentation will be held during the last week of the semester.

**Scope**:

- Students select a topic of personal interest or curiosity. While it may relate to their research, the project must be self-contained and utilize the tools introduced in the class or more advanced ones.

**Expectations**:

- **Undergraduate Students**: Focus on a well-defined, independent project.
- **Graduate Students**: Address strong scientific research questions with rigorous data analysis using advanced methods from the course. The final report should demonstrate novelty and quality comparable to a peer-reviewed scientific article.
- **Evaluation** will consider the project's originality, quality, and alignment with academic standards.

# Tentative schedule

| Timeline | Topic | Software | Assignments |
|---|---|---|---|
| 01/14 | Introduction | Jupyter/GitHub/Bash | |
| 01/16 | Python overview | Python | |
| 01/21 | Multi-dimensional arrays I | NumPy | |
| 01/23 | Probability/Statistics | Scipy | HW 0 |
| 01/28 | Visualizing data | Matplotlib | |
| 01/30 | Data storage | NetCDF/GeoTiff/NetCDF/HDF5/Zarr | |
| 02/04 | Probability/Statistics | | HW 1 |
| 02/06 | Bayesian Statistics | | |

# Tentative schedule

| Timeline | Topic | Software | Assignments |
|---|---|---|---|
| 02/11 | Map Projections I | Cartopy | |
| 02/18 | Map Projections II | GDAL | |
| 02/20 | Multi-dimensional arrays II | CDO/Xarray | |
| 02/25 | Vector Data | OGR/Shapely/GeoPandas | |
| 02/27 | Cluster Analysis I | Scikit-Learn | HW 2 |
| 03/04 | Cluster Analysis II | | |
| 03/06 | Dimensionality Reduction | | |
| 03/11 | Decision Trees | | |
| 03/13 | Random Forests & Boosting | | |
| Spring Break | | | |

# Tentative schedule

| | | | |
|---|---|---|---|
| Spring Break | | | |
| 03/25 | Artificial Neural Networks | | |
| 03/27 | Convolutional Neural Networks | | HW 3 |
| 04/01 | Kriging and Semi-variogram | | Project Proposal |
| 04/03 | Regression Kriging | | |
| 04/08 | Terrain Analysis | | |
| 04/10 | Interactive Visualization | Folium/leaflet | |
| 04/15 | Parallel computing | Numba/Mpi4py/Dask | HW 4 |
| 04/17 | BONUS | | |
| 04/21 | Oral Presentations | | |
| 04/23 | Oral Presentations | | Project Paper |

# Leverage Rice University High-Performance Computing (HPC) Infrastructure

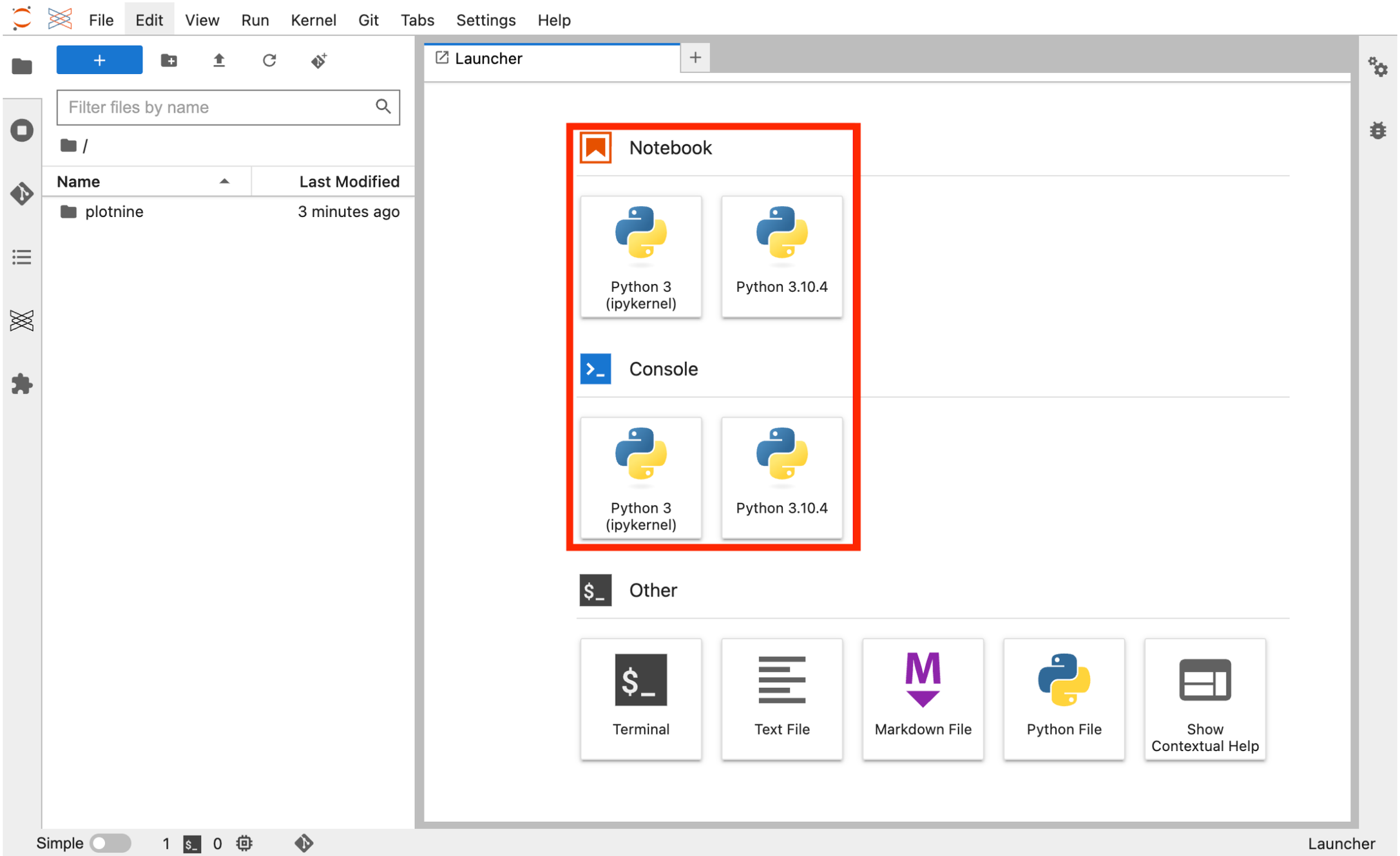
- HPC is a group of powerful computers that solve large problems
- A Docker container in the HPC will host our class (virtual machine)
- We will use **Jupyter Lab**, **Notebooks** and the **Terminal** on the Docker

# Leverage Rice University High-Performance Computing (HPC) Infrastructure

Next week we will cover:

- How to access the docker and run the notebooks
- How to access and submit assignments in the docker

# Jupyter Lab




https://jupyterlab.readthedocs.io/en/stable/

# Jupyter Notebook

jupyter jupyter-interactive-visualization Last Checkpoint: 3 minutes ago (autosaved) Quit

File Edit View Insert Cell Kernel Widgets Help Trusted Python 3 (ipykernel)

Run Markdown

## Interactive Python Visualization Libraries

In [ ]:
```
import numpy as np
import plotly.offline as py
import plotly.figure_factory as ff
from bokeh.models import HoverTool, WheelZoomTool
from bokeh.plotting import figure
from bokeh.io import show, output_notebook
output_notebook()
```

### Plotly

In [ ]:
```
py.init_notebook_mode()
t = np.linspace(-1, 1.2, 2000)
x = (t**3) + (0.3 * np.random.randn(2000))
y = (t**6) + (0.3 * np.random.randn(2000))

colorscale = ['#7A4579', '#D56073', 'rgb(236,158,105)', (1, 1, 0.2), (0.98,0.98,0.98)]

fig = ff.create_2d_density(
    x, y, colorscale=colorscale,
    hist_color='rgb(255, 237, 222)', point_size=3
)

py.iplot(fig, filename='histogram_subplots')
```

https://jupyter-notebook-beginner-guide.readthedocs.io/en/latest/what_is_jupyter.html

# Terminal


Remains one of the best ways to interact more directly with the computer (flexibility!)

# Interacting with the terminal: Bash (Bourne Again SHell)

# Tutorials to learn Bash

https://linuxconfig.org/bash-scripting-tutorial-for-beginners

https://www.codecademy.com/learn/learn-the-command-line

https://devhints.io/bash

https://hackernoon.com/top-10-bash-file-system-commands-you-cant-live-without-4cd937bd7df1

# Top bash commands

1) **ls -** Lists the folder and file names in the current working directory.
2) **cd -** Change Directory to the path specified, for example cd Lectures
3) **mkdir -** Make directories with this command mkdir my_folder
4) **mv -** Moves files and folders. The first argument is the file you want to move, and the second is the location to move it to.
5) **cp -** Copies files and folders cp my_file ./projects . The flag -r recursively copies subfolders and files.
6) **rm -** Removes files and folders rm my_folder . Using -r will again recursively delete subfolders, -f force deletes, and -rf for a recursive force delete.

# Let’s talk about version control…

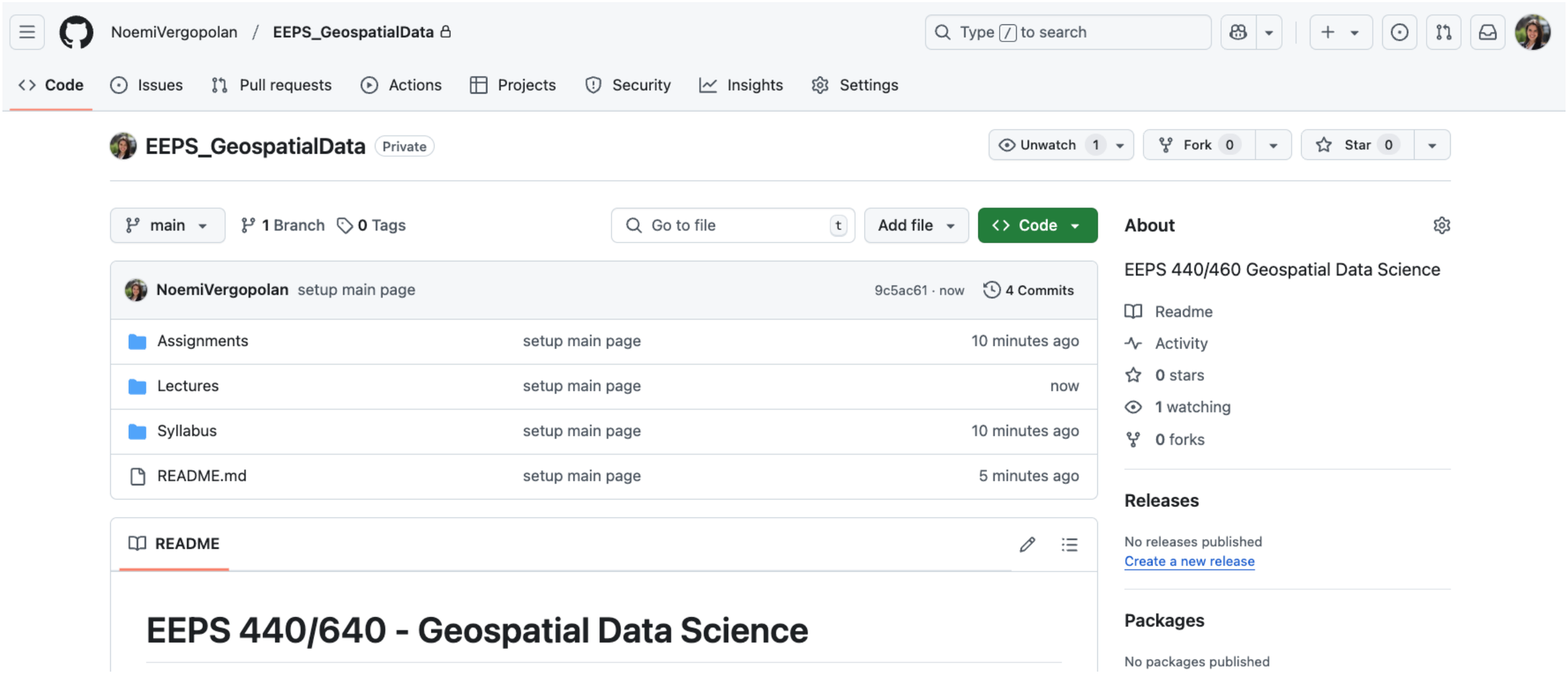
# GitHub: Git on the web


https://www.youtube.com/watch?v=w3jLJU7DT5E

# Version control: Git

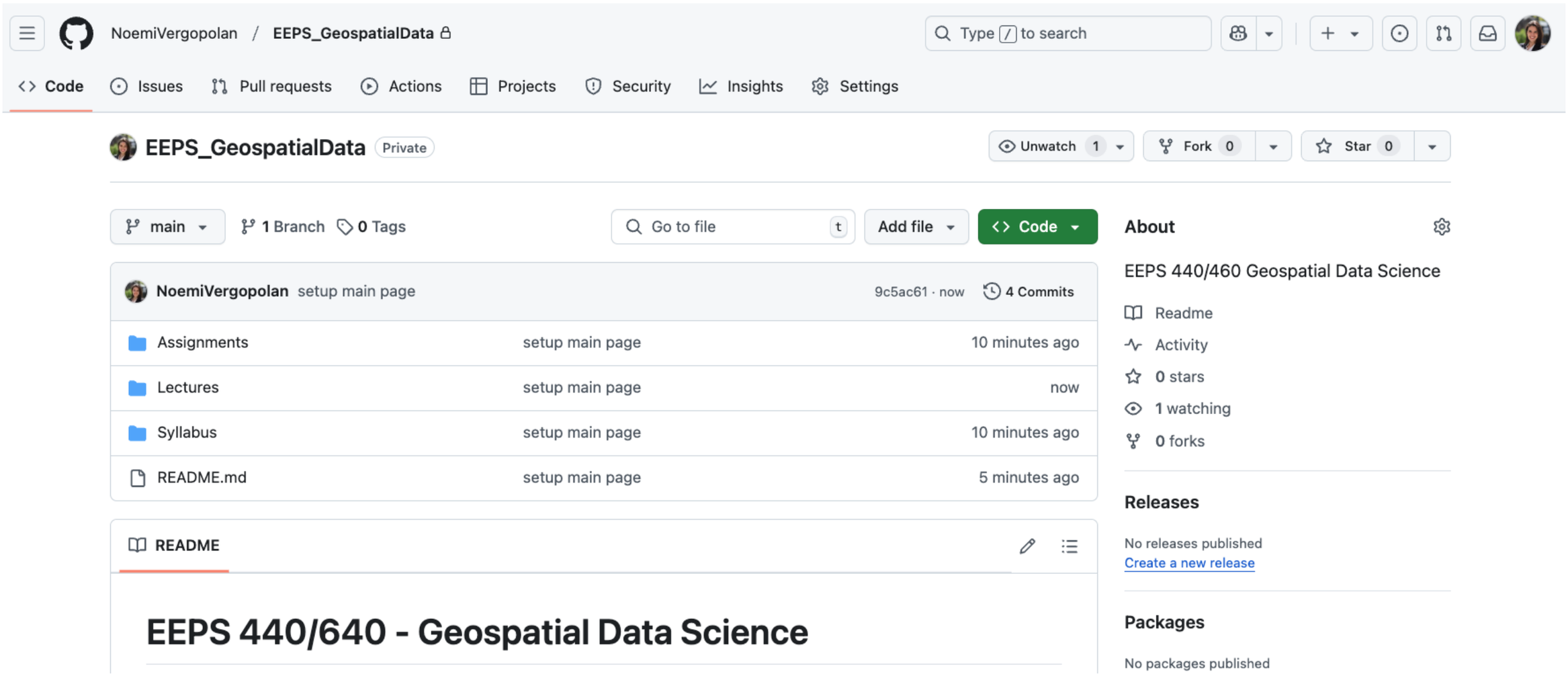

https://vimeo.com/41027679

# The primary class website for this course is on GitHub



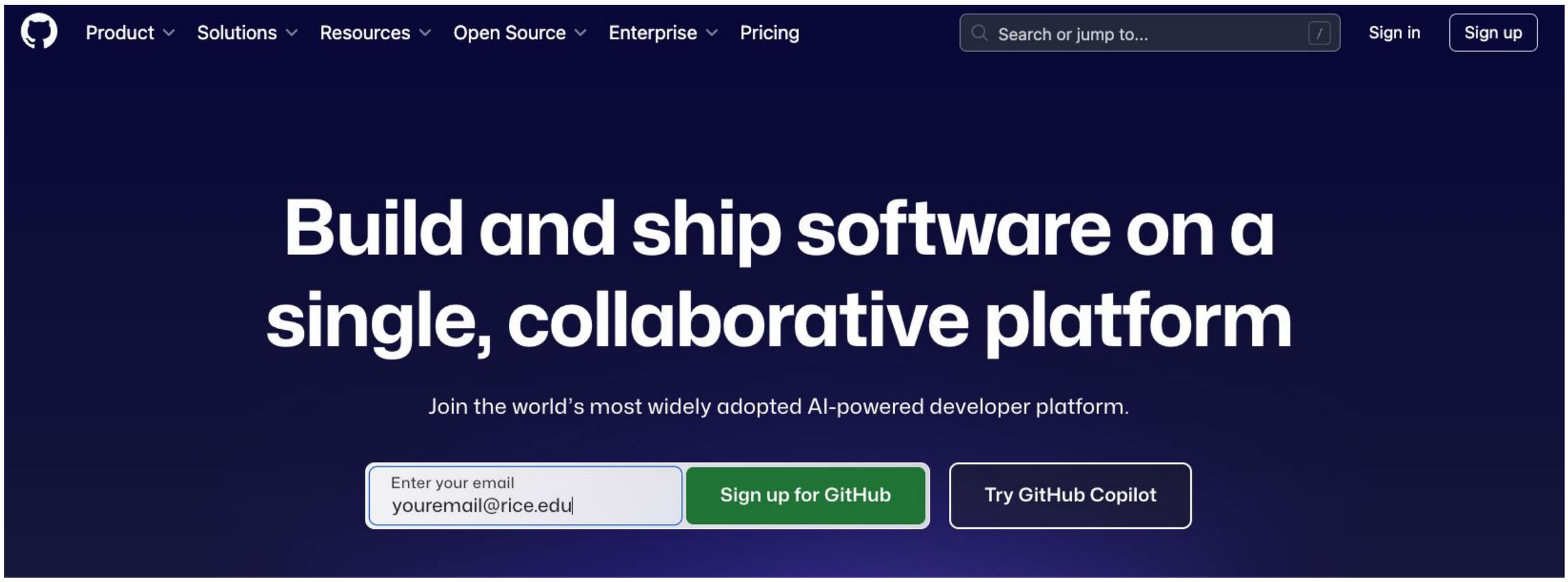

https://github.com/NoemiVergopolan/EEPS_GeospatialData.git

# All lectures, assignments, and materials will be on the GitHub repository

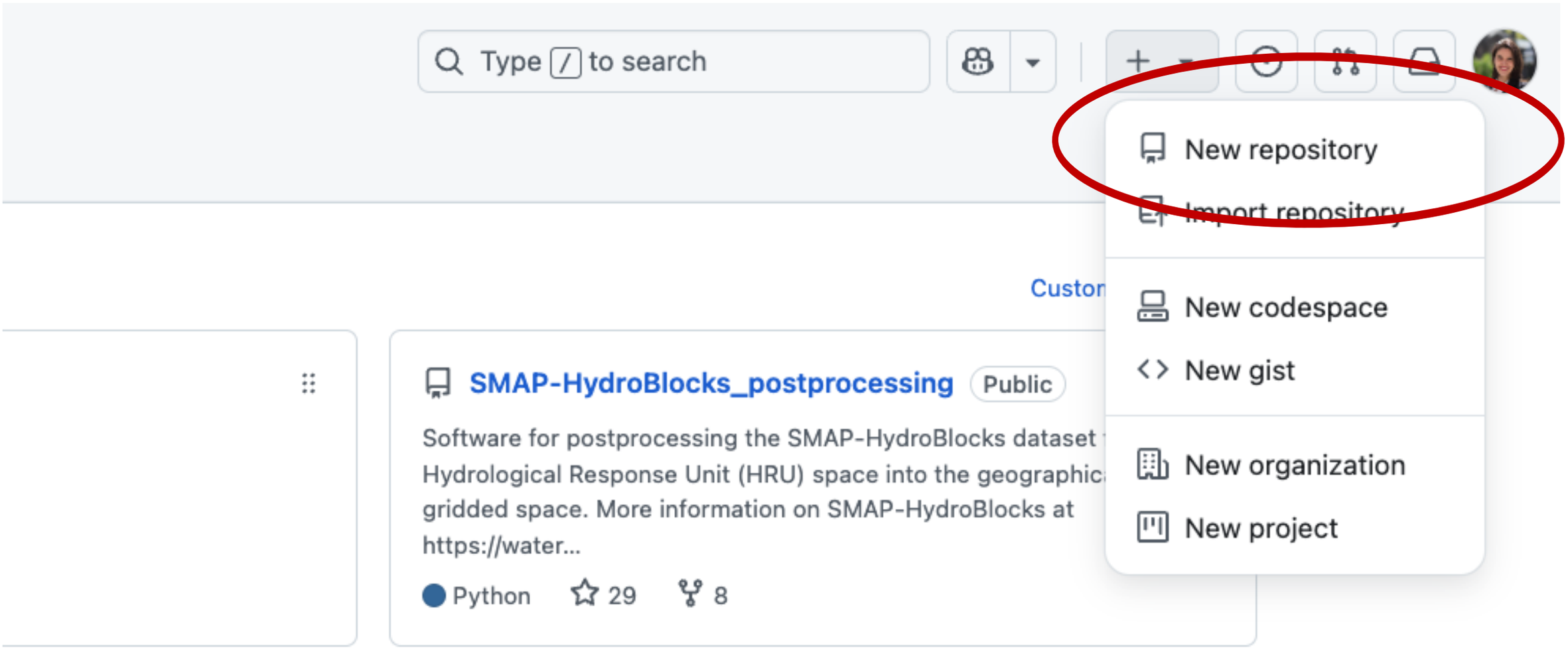



*Solutions to assignments will be posted on Canvas

# Create a personal account on GitHub.com



# Create a new repository from your homepage



# Naming your class repository GDS25_initials

## Create a new repository

A repository contains all project files, including the revision history. Already have a project repository elsewhere? Import a repository.

*Required fields are marked with an asterisk (*).*

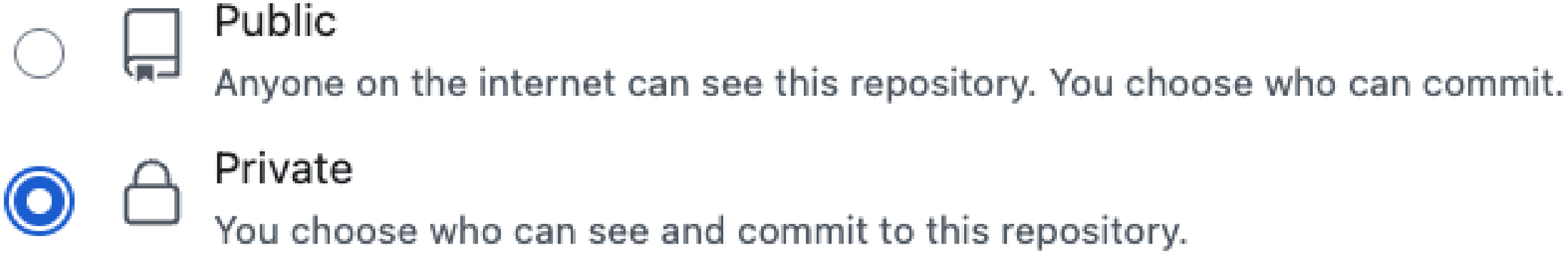
**Owner *** **Repository name ***

NoemiVergopolan / GDS25_nvr

GDS25_nvr is available.

Great repository names are short and memorable. Need inspiration? How about **curly-fortnight** ?

Make private (unless you want everyone to see your completed assignments!)



# Create the repository!

**Initialize this repository with:**

☐ Add a README file

This is where you can write a long description for your project. Learn more about READMEs.

**Add .gitignore**

.gitignore template: None

Choose which files not to track from a list of templates. Learn more about ignoring files.

**Choose a license**

License: None

A license tells others what they can and can't do with your code. Learn more about licenses.

---

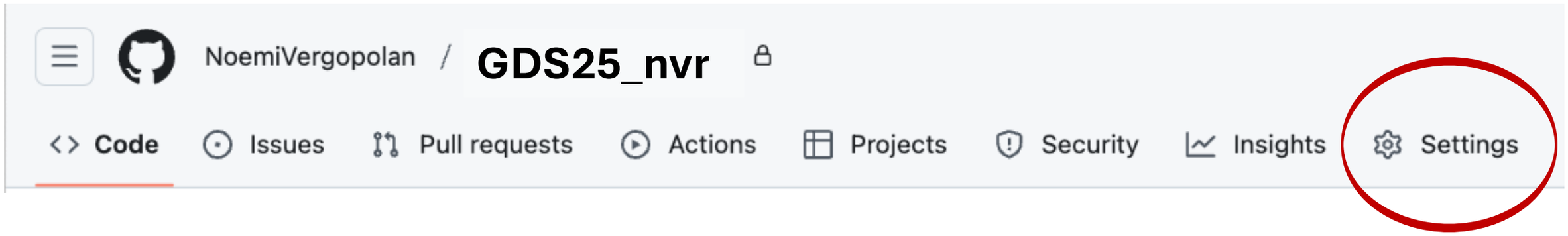
ⓘ You are creating a private repository in your personal account.

---

Create repository

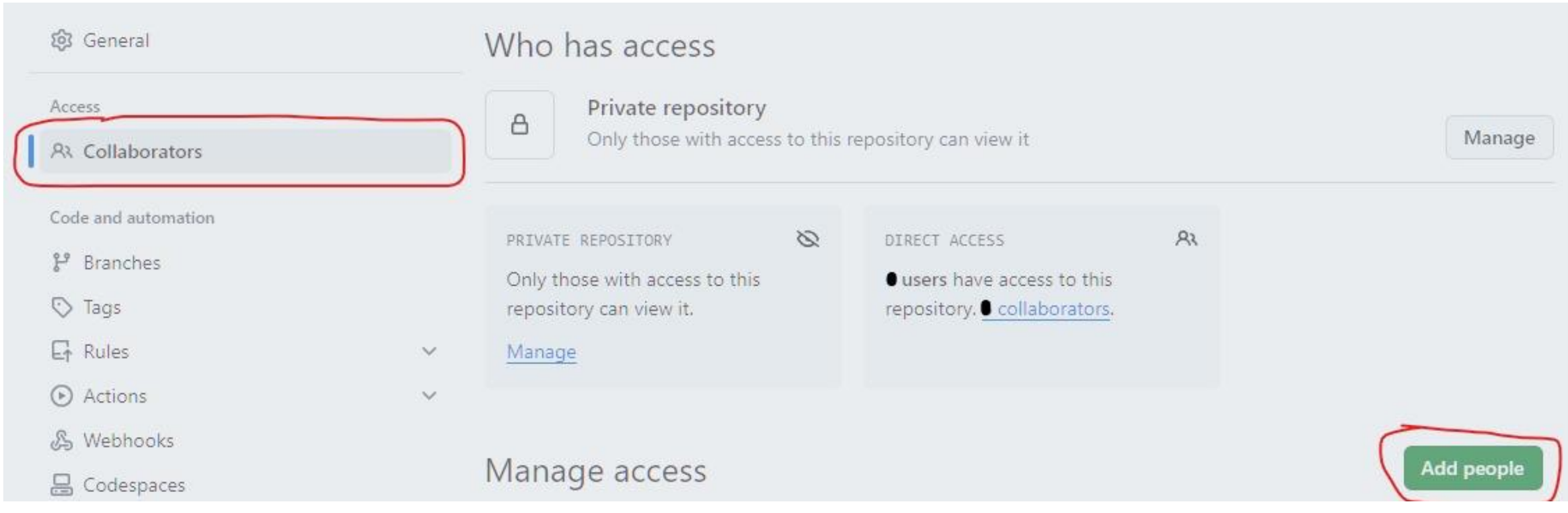
# Last thing you need to do is share your private repository with myself

On your repository, go to:
"Settings" then "Collaborators" then "Add People"



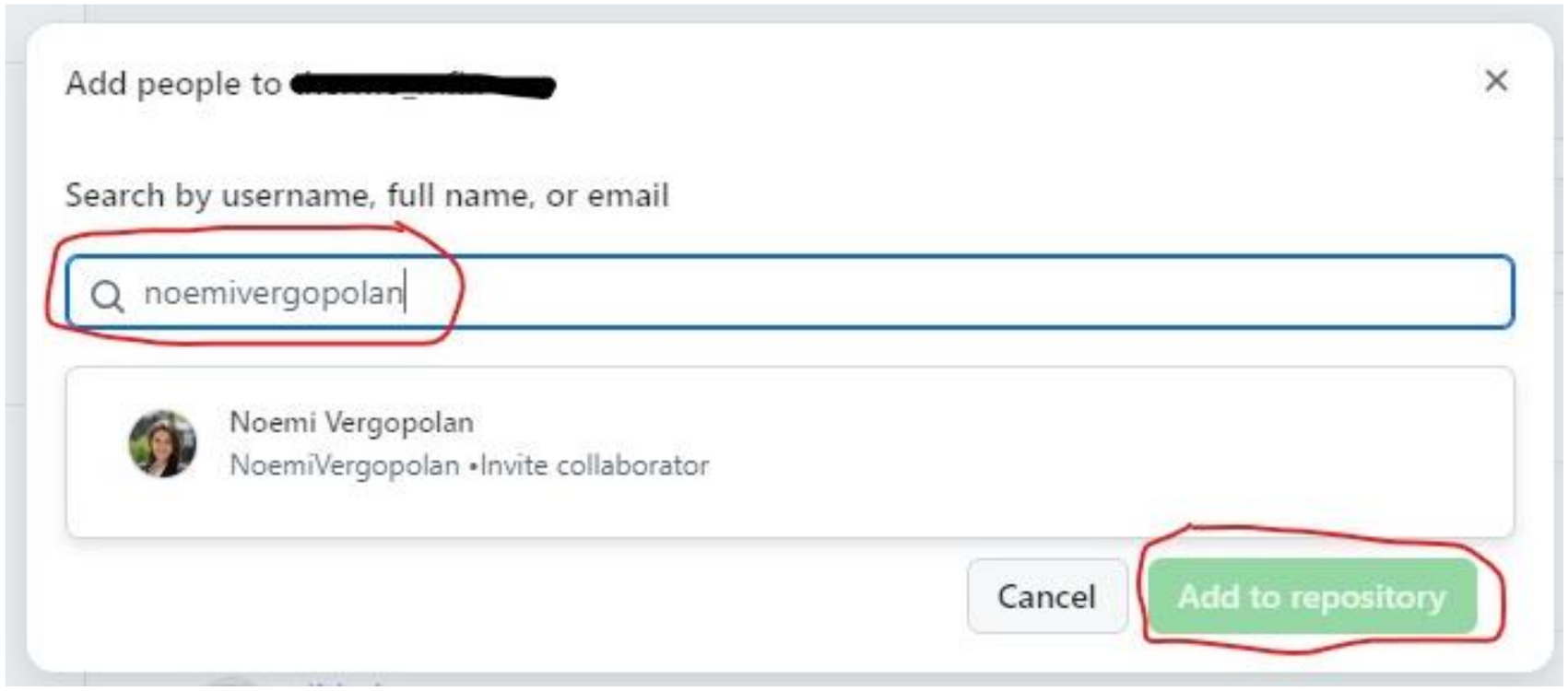
# Last thing you need to do is share your private repository with myself

On your repository, go to:
"Settings" then "Collaborators" then "Add People"



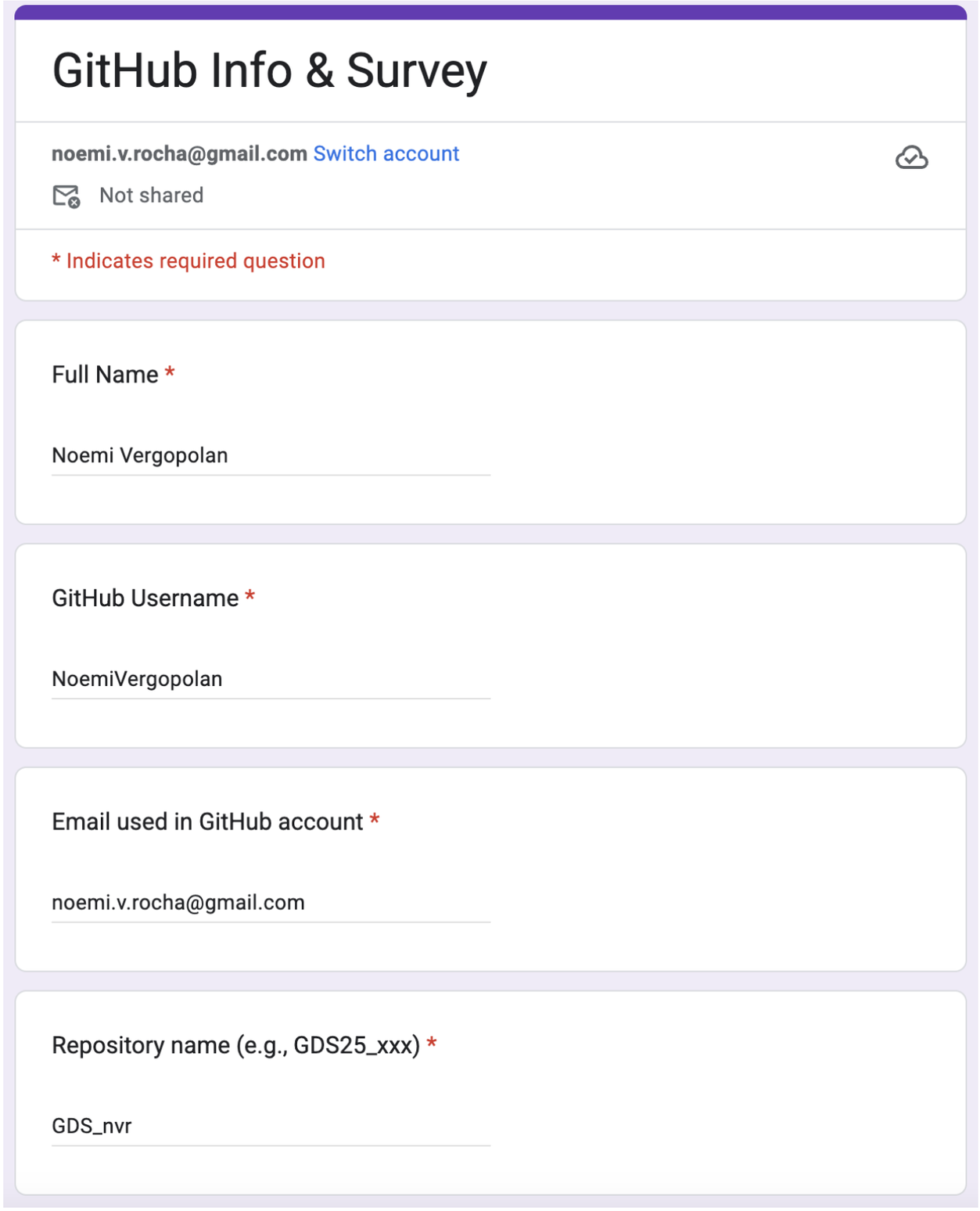
# Last thing you need to do is share your private repository with myself

On your repository, go to:
"Settings" then "Collaborators" then "Add People"
**noemivergopolan**



# https://tinyurl.com/rs2r2xss

**Share with me your GitHub info & fill survey**



GitHub Info & Survey

noemi.v.rocha@gmail.com Switch account

Not shared

* Indicates required question

Full Name *

Noemi Vergopolan

GitHub Username *

NoemiVergopolan

Email used in GitHub account *

noemi.v.rocha@gmail.com

Repository name (e.g., GDS25_xxx) *

GDS_nvr

## Issue with class starting 15 or 30min earlier?

**Thursday:**

- Intro to Python Programming

**Next week:**

- How to access the Docker, Github, and Homeworks
- Numpy and Scipy

Deadlines for the homework will be updated!