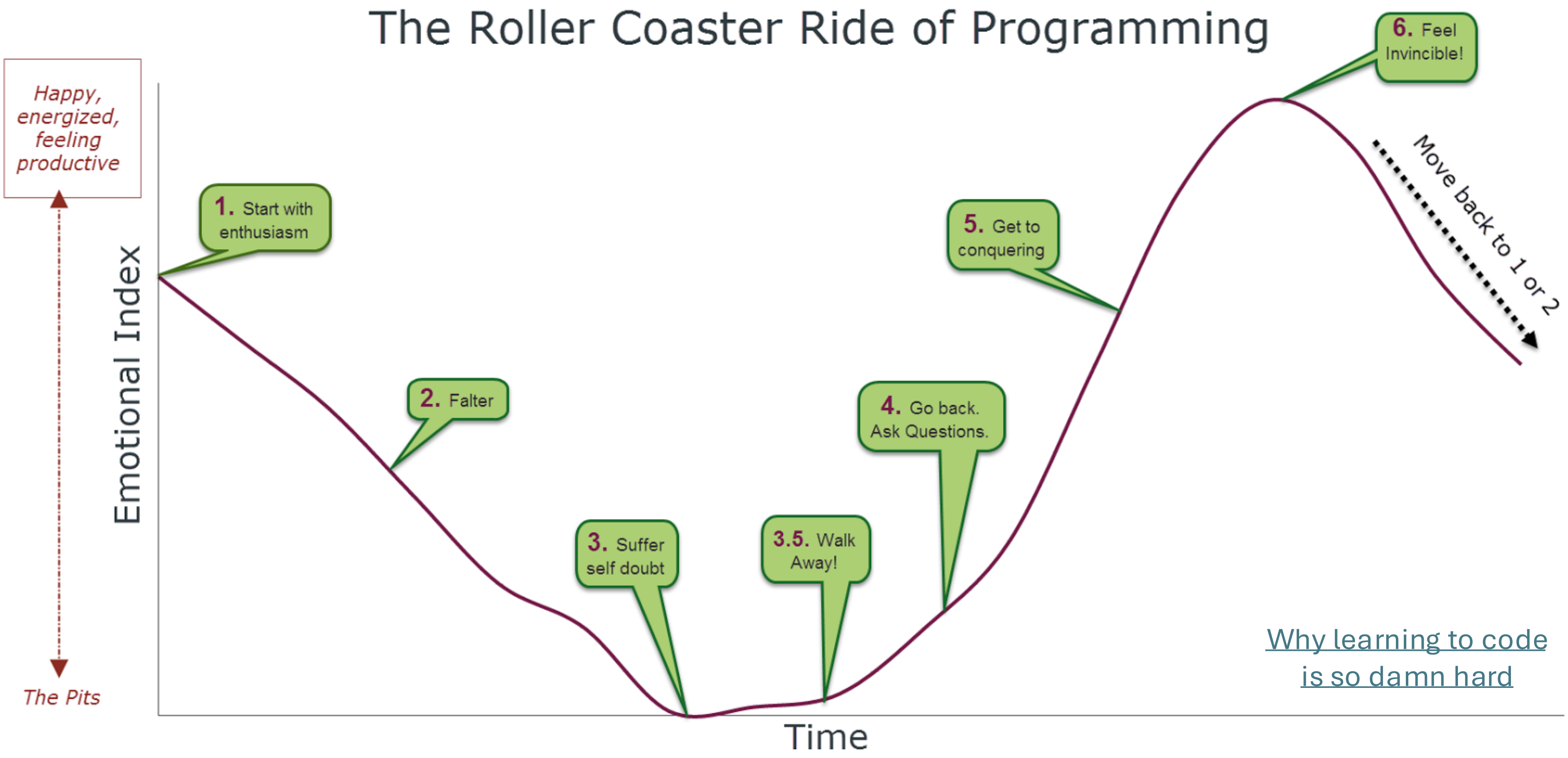
ERT 429/529

Analysis of Geologic Data



The Roller Coaster Ride of Programming

Why learning to code is so damn hard

Instructor: Dr. Yifan Cheng ycheng46@buffalo.edu

Student Hours: Tue. 1-1:30 PM, Th. 1-1:30 PM, or by appointment

January 22nd, 2026

Spring 2026

Today’s Learning Outcomes

1. Know my expectations and pedagogical approach to the course

2. Know the course layout

3. Access GitHub CodeSpaces and open a Jupyter Notebook

Teaching Philosophy

1. My job is to help you understand the subject matter and set you up for success moving forward

2. Process and methods are more important than memorizing facts

3. You are adults and will be treated as such

Goals of the Course

1. Become familiar with generalized statistical techniques

2. Know how to determine what statistical approaches are appropriate (or not) for a particular problem

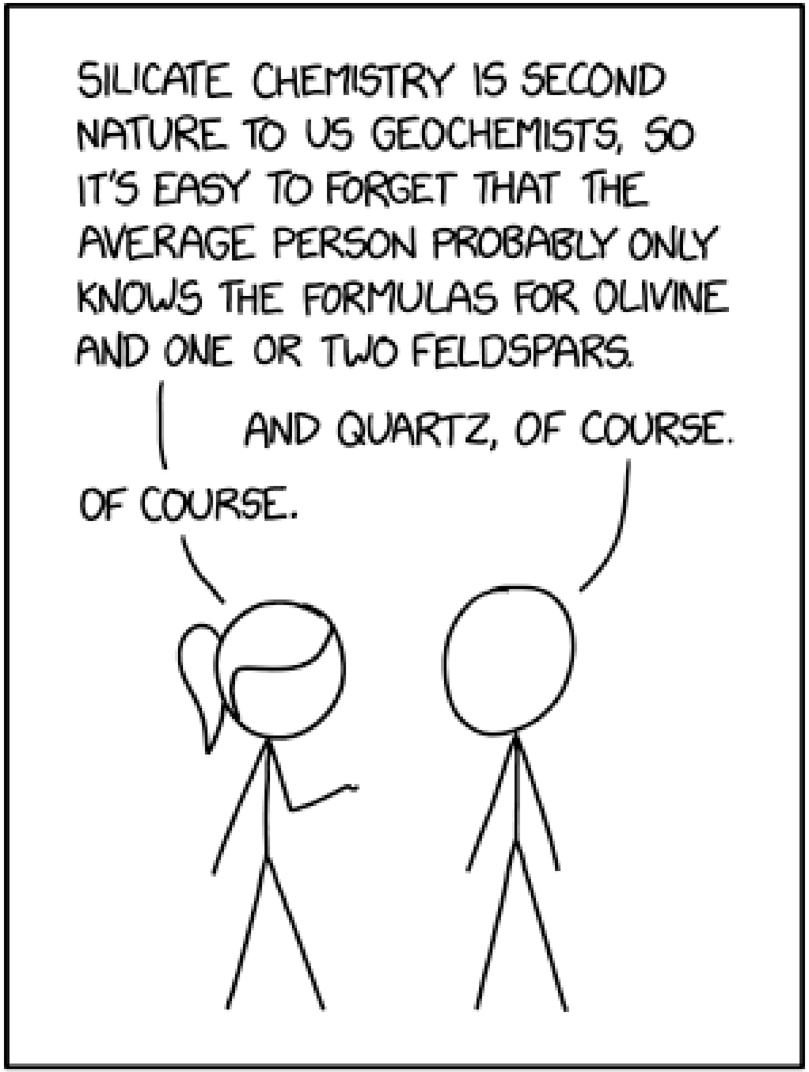
3. Develop fluency in the Python programming language for statistical analyses, but also best practices for producing reproducible code

Disclaimer

- This course has been taught in the R programming language since it was first developed.

- However, I think learning Python would be more beneficial for both research and industry-driven careers.

- Small hiccups might occur during the course! Please be kind ☺


SILICATE CHEMISTRY IS SECOND NATURE TO US GEOCHEMISTS, SO IT'S EASY TO FORGET THAT THE AVERAGE PERSON PROBABLY ONLY KNOWS THE FORMULAS FOR OLIVINE AND ONE OR TWO FELDSPARS.
AND QUARTZ, OF COURSE.
OF COURSE.

EVEN WHEN THEY'RE TRYING TO COMPENSATE FOR IT, EXPERTS IN ANYTHING WILDLY OVERESTIMATE THE AVERAGE PERSON'S FAMILIARITY WITH THEIR FIELD.

Course Materials

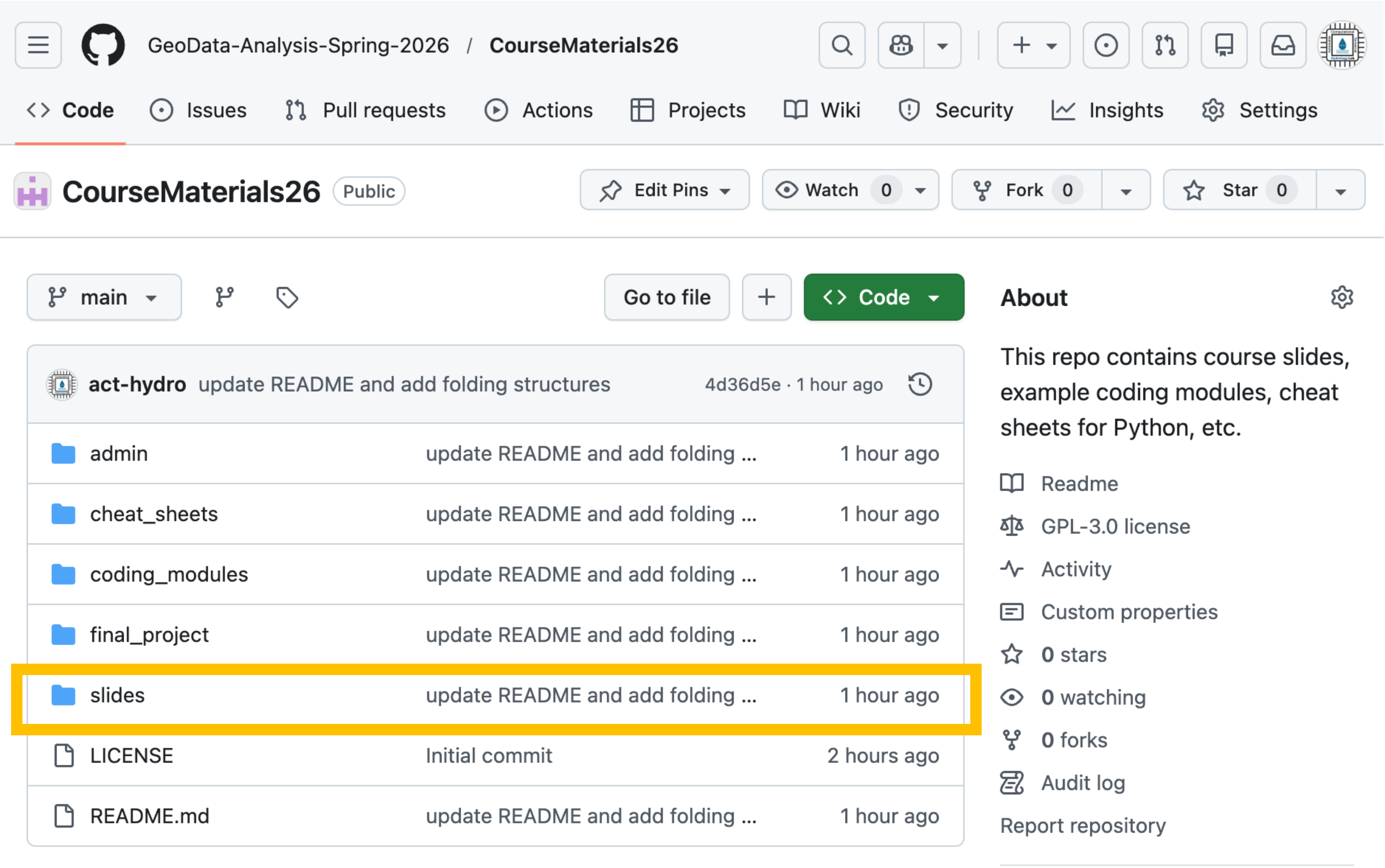
- No required textbook, but there are many resources available for Python and statistics
  - Introduction To Computer Science And Programming in Python
  - Python for Data Science
  - Basics of Python Programming: a Quick Guide for Beginners [available online through UB library]

- Course organizations
  - **Github**: course materials, in-class coding practices, homework submissions.
  - **UBLearns**: notifications, homework instructions.

- Lectures

GitHub

Course GitHub Organization: https://github.com/GeoData-Analysis-Spring-2026



act-hydro update README and add folding structures 4d36d5e · 1 hour ago

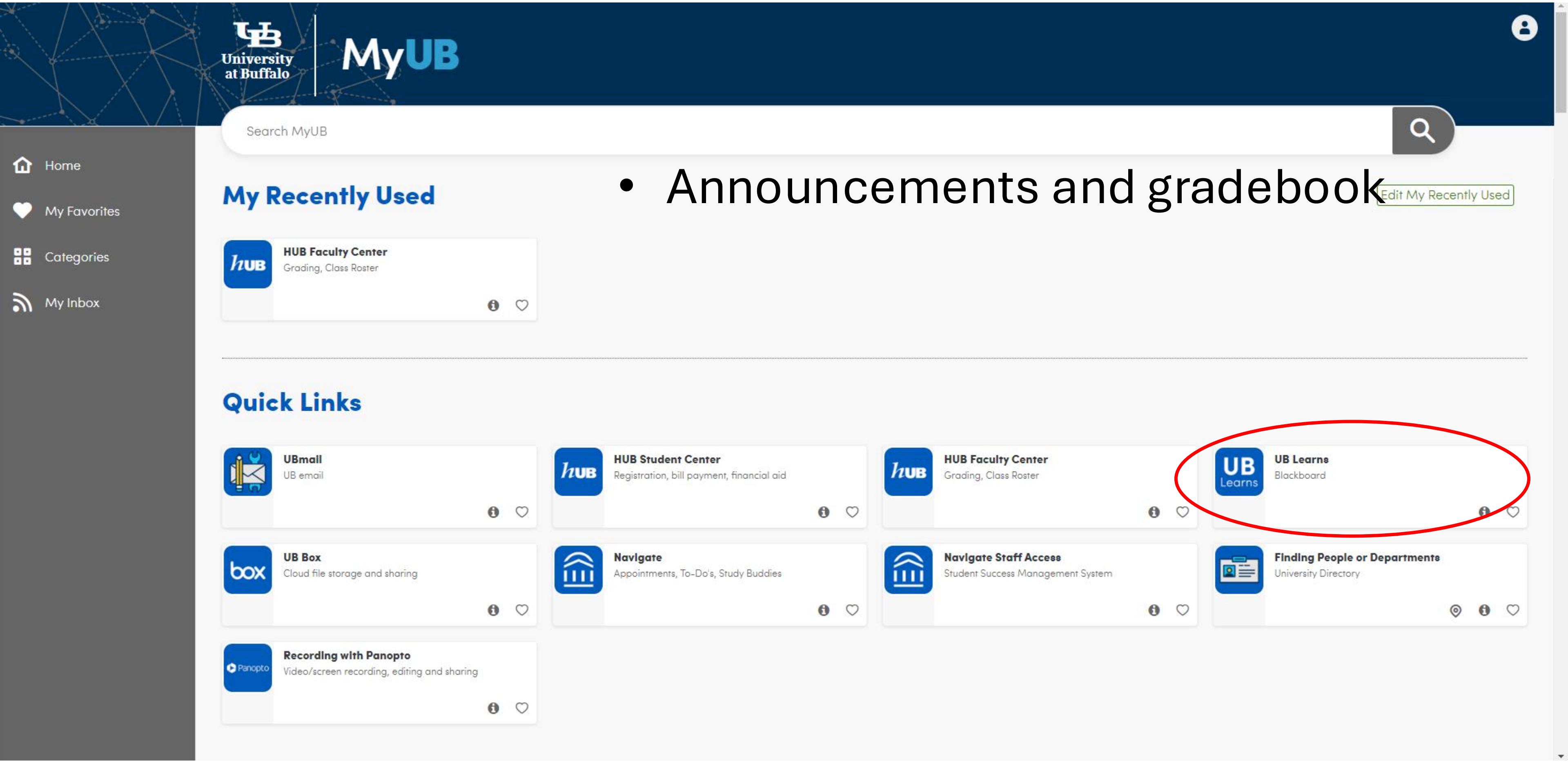
admin	update README and add folding ...	1 hour ago
cheat_sheets	update README and add folding ...	1 hour ago
coding_modules	update README and add folding ...	1 hour ago
final_project	update README and add folding ...	1 hour ago
slides	update README and add folding ...	1 hour ago
LICENSE	Initial commit	2 hours ago
README.md	update README and add folding ...	1 hour ago

This repo contains course slides, example coding modules, cheat sheets for Python, etc.

Slides (in PDF format) will be available on **GitHub** for each week in the “slides” folder

UBLearns

- Announcements and gradebook



My Recently Used

HUB Faculty Center
Grading, Class Roster

Quick Links

UBmail
UB email

HUB Student Center
Registration, bill payment, financial aid

HUB Faculty Center
Grading, Class Roster

UB Learns
Blackboard

UB Box
Cloud file storage and sharing

Navigate
Appointments, To-Do's, Study Buddies

Navigate Staff Access
Student Success Management System

Finding People or Departments
University Directory

Recording with Panopto
Video/screen recording, editing and sharing

Email Etiquette

- Email is the only way I have to contact the whole class outside of lectures themselves

- You **need** to check your email regularly at UB
  - Yes, 95+ percent can probably be ignored but the other 5% can be critical!

- Any emails I send out will have the course # at the start of the subject line (“ERT 429 or ERT 529: …”)

- If you are emailing me please also use the course number
  - Referring to “the exam” is not helpful to tell me which of my exams across classes you might be talking about

Accommodations

- If you have any accommodations from the Accessibility Resources Office, or believe that you are eligible for them, please get them to me **as soon as possible**

- Some of the most common accommodations are extended time on timed assignments and the option to take exams in quiet spaces (usually the Accessibility Resources Office)

Course Grade Structure

	Grade Percent
Attendance	10%
Weekly Problem Sets	35%
Exams	30%
Final Project	25%

Total Grade % = (attendance % * 0.1) + (problem sets % * 0.35) + (exams % * 0.3) + (final project % * 0.25)

Attendance

- I will take attendance with a minute paper at the end of each lecture
    - In this minute paper, you will need to list 1) what you like about this lecture, and 2) what you do not fully understand. I will revisit the minute paper the first thing in the next lecture
    - This is to **1)** maximize instant feedback, **2)** make sure students are all receiving the same set of information, and **3)** provide a dedicated time to work with the material

- Life happens, if you attend 24 lectures (out of all lectures) you get the full 10% for attendance

Weekly Problem Set Details

- Most weeks (except exam weeks) you will be asked to write your code and perform analyses on a dataset applying the course material from that week

- Prompts will start off as step-by-step instructions but… become progressively less detailed in how to accomplish the task(s) as the semester goes on

Weekly Problem Set Details

- Problem sets will be released after class on Thursdays and due at the start of the following Thursday's class

- Submission will be via GitHub and will usually contain both the code you have written as well as output in the form of graphs or statistical results
  - You are expected to submit a **Jupyter Notebook** per homework, which will include all the code and the written part.
  - Detailed instructions concerning how to submit homework will be provided by the end of this lecture.
  - Graduate students will have a few additional questions each week

Exam Details

- The are take-home exam will be longer-form version of problem sets where the focus in on whether you can integrate statistical techniques together

- Despite being take-home exams, we will have one full class period (February 26th and April 9th) dedicated to starting the exams where students can ask questions
  - The complete exams are due a week later (March 5th and April 16th) through UBLearns

Final Project Details

- The last three weeks of the course will be for an application of the techniques you have developed over the semester

- Students will be given a choice of datasets (with associated tasks) to process, analyze, and write a report of the results
  - Graduate Students may propose their own project if they wish

- In addition to the report graduate students will also each give a short presentation (~15 minutes) on their final project on the last day of class (May 5th)

Final Project Details

- There will be several checkpoints for the project as shown in the table below

Checkpoint	Due Date	Project Grade	Outcome
Pick topic	March 26, 11:59PM	5%	Approval
Draft Section 1	April 23, 11:59PM	15%	Draft comments
Presentation (grads only)	May 5 Class	15%	Awesome Presentations
Project Due	May 12, 11:59PM	80%/65%	Final Project Grade

Bonus Exercises

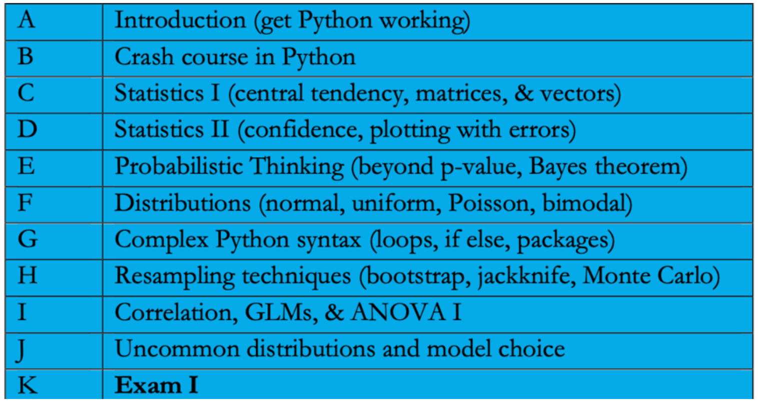
- This is a stats course, but to do so we need to use a programming environment (like Python)

- The best way to become proficient at coding is to do it and just become more and more familiar

- So every week I will have a set of simple practice exercises using Python and/or stats
  - Each fully completed exercise is worth 0.5% toward your total grade (up to a maximum of +5%)

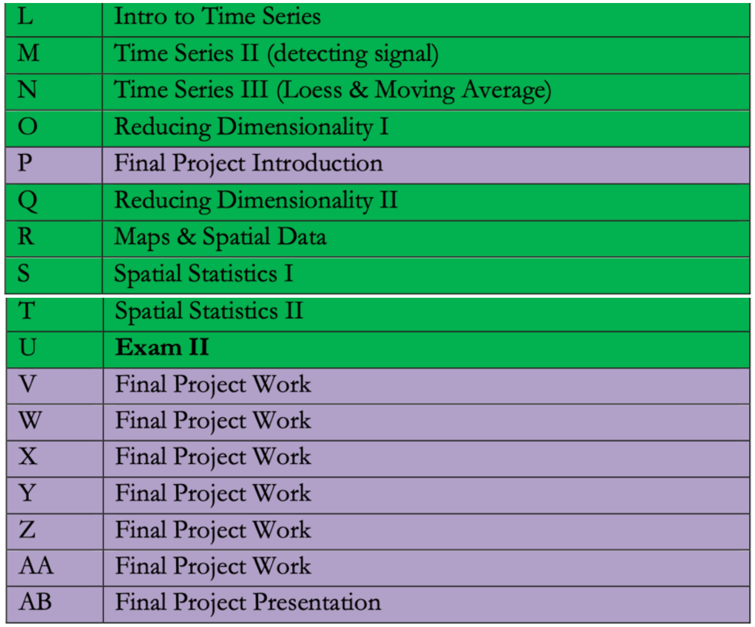
Lecture Schedule

Programming in Python and “universal” statistics



A	Introduction (get Python working)
B	Crash course in Python
C	Statistics I (central tendency, matrices, & vectors)
D	Statistics II (confidence, plotting with errors)
E	Probabilistic Thinking (beyond p-value, Bayes theorem)
F	Distributions (normal, uniform, Poisson, bimodal)
G	Complex Python syntax (loops, if else, packages)
H	Resampling techniques (bootstrap, jackknife, Monte Carlo)
I	Correlation, GLMs, & ANOVA I
J	Uncommon distributions and model choice
K	**Exam I**

Geology-specific statistics and application



L	Intro to Time Series
M	Time Series II (detecting signal)
N	Time Series III (Loess & Moving Average)
O	Reducing Dimensionality I
P	Final Project Introduction
Q	Reducing Dimensionality II
R	Maps & Spatial Data
S	Spatial Statistics I
T	Spatial Statistics II
U	**Exam II**
V	Final Project Work
W	Final Project Work
X	Final Project Work
Y	Final Project Work
Z	Final Project Work
AA	Final Project Work
AB	Final Project Presentation

AI and LLMs

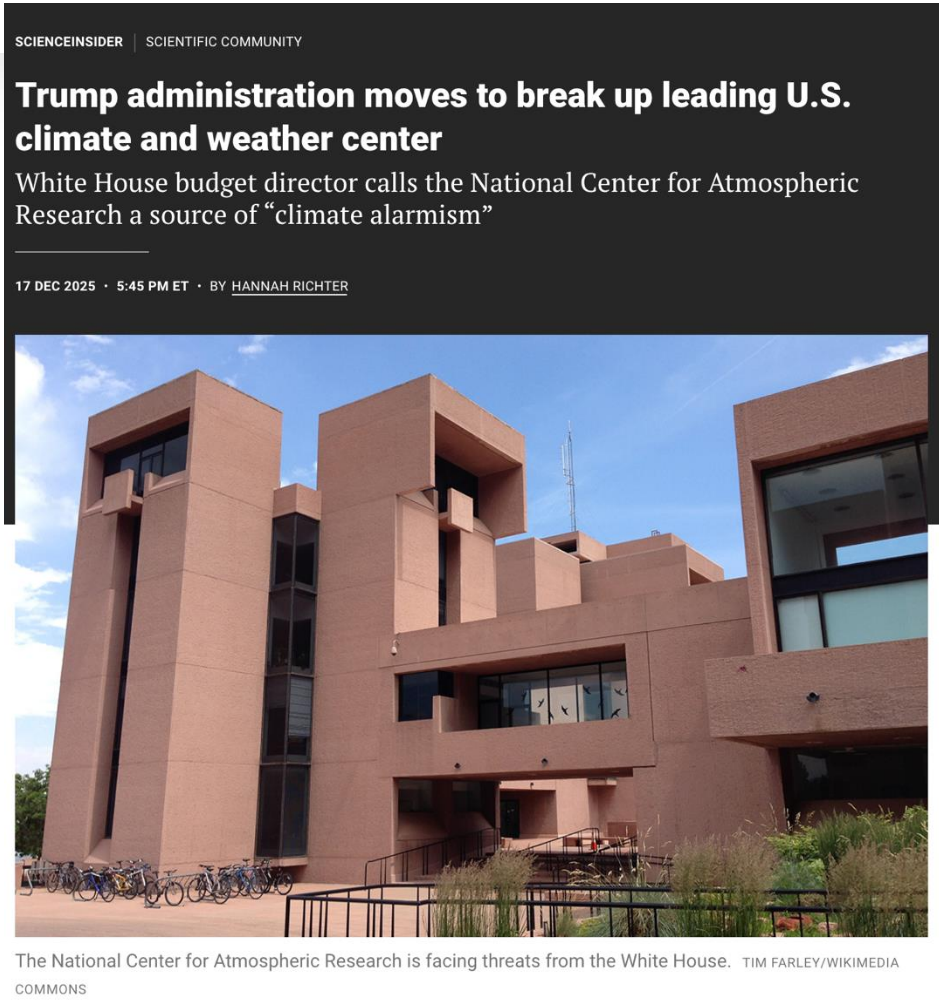
- I have a lot of misgivings about AI/LLM (see here, here, and here) as a tool for academic pursuits and specifically in learning where it is often used as a replacement for engaging with material and actually learning

- However, coding is one of the places where AI/LLM seems to be more useful/reliable
  - We may explore this later in the course but I ask you to refrain from using it until we make it past week 5 in the semester (building Python competency)

Who is this guy?

- 2020 PhD from University of Washington Civil and Environmental Engineering

- Second year at UB as faculty
    - Previously a Scientist at the National Center for Atmospheric Research (NCAR) – The one that Trump threatens to dismantle in December 2025



SCIENCEINSIDER | SCIENTIFIC COMMUNITY

Trump administration moves to break up leading U.S. climate and weather center

White House budget director calls the National Center for Atmospheric Research a source of "climate alarmism"

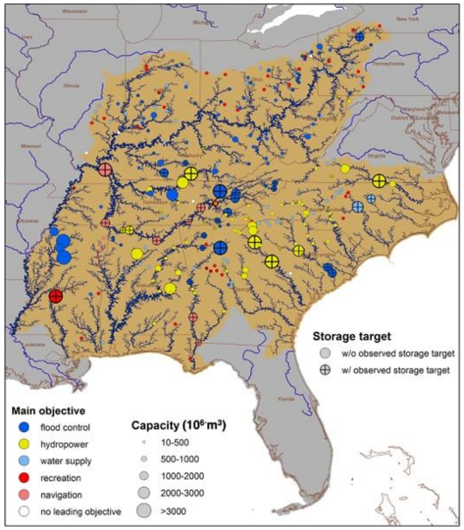
17 DEC 2025 • 5:45 PM ET • BY HANNAH RICHTER

The National Center for Atmospheric Research is facing threats from the White House. TIM FARLEY/WIKIMEDIA COMMONS

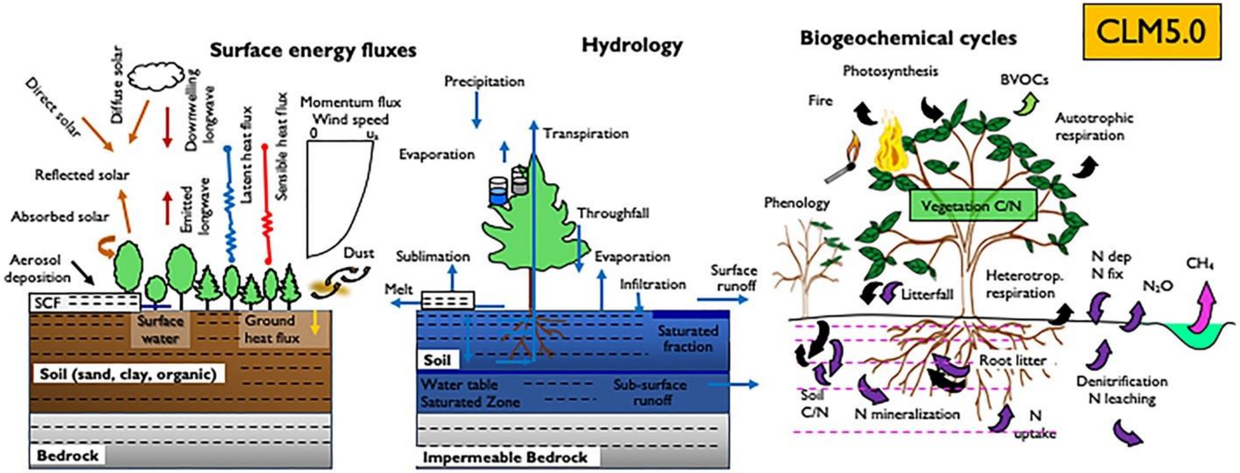
Research Interests

- I am a computational hydrologist and interested in building and improving large-scale numerical models
  - Large-scale land surface and hydrologic models
  - Coupled land-atmospheric models
  - Complex river-reservoir systems
  - High-performance computing (HPC)

River-reservoir systems

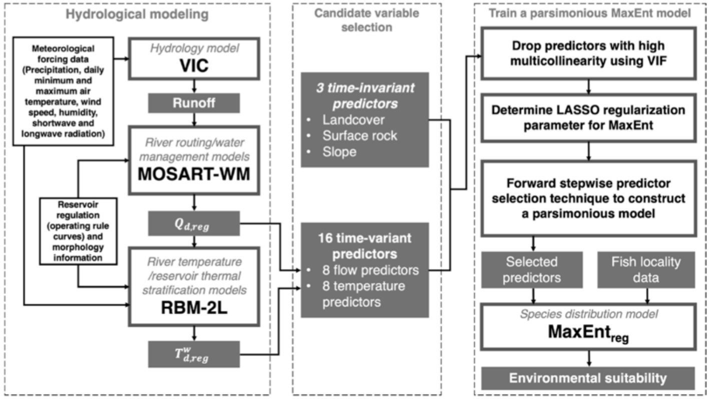


Community Terrestrial Systems Model (CTSM)



Research Interests

Also care about how hydrologic changes impact water-related sectors



Hydrological modeling

Meteorological forcing data (Precipitation, daily minimum and maximum air temperature, wind speed, humidity, shortwave and longwave radiation)

Hydrology model VIC

Runoff

River routing/water management models MOSART-WM

Reservoir regulation (operating rule curves) and morphology information

$Q_{d,reg}$

River temperature /reservoir thermal stratification models RBM-2L

$T^{w}_{d,reg}$

Candidate variable selection

3 time-invariant predictors
- Landcover
- Surface rock
- Slope

16 time-variant predictors
- 8 flow predictors
- 8 temperature predictors

Train a parsimonious MaxEnt model

Drop predictors with high multicollinearity using VIF

Determine LASSO regularization parameter for MaxEnt

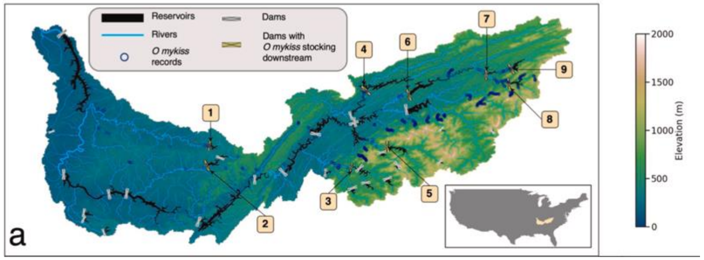
Forward stepwise predictor selection technique to construct a parsimonious model

Selected predictors

Fish locality data

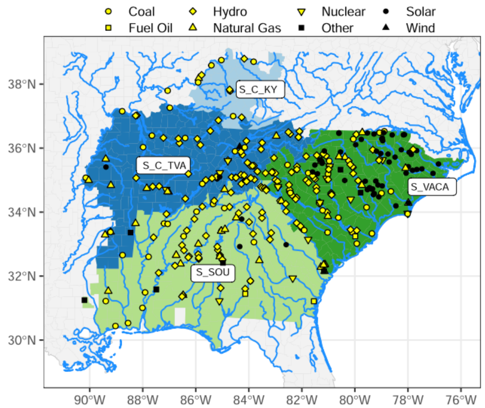
Species distribution model MaxEnt_{reg}

Environmental suitability

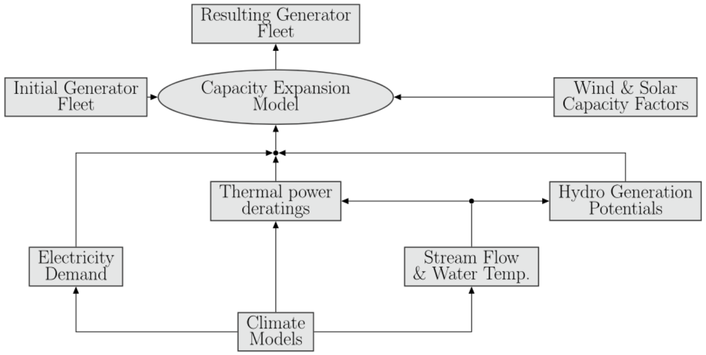


Aquatic Ecosystem: How do reservoir dynamics affect aquatic ecosystems?

Cheng, Y., et al., 2022, *Journal of Hydrology*



Fonseca, et al., 2021, *Environmental Science and Technology*



Resulting Generator Fleet

Initial Generator Fleet

Capacity Expansion Model

Wind & Solar Capacity Factors

Thermal power deratings

Hydro Generation Potentials

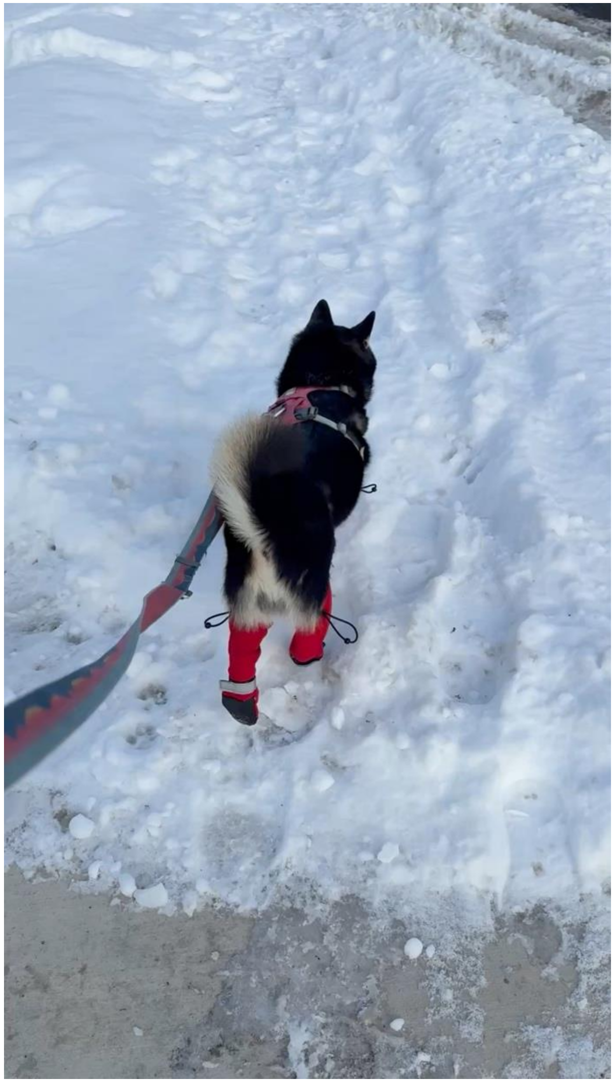
Electricity Demand

Stream Flow & Water Temp.

Climate Models

Electric power systems: How will climate change affect capacity expansion decisions?

Non-Academic Life



Mika (born in Seattle, 2018) is trying to get used to the cold weather in Buffalo

Non-Academic Life

- Board games
- Museums
- Musicals
- Skiing
- Travel
- Playing fetch with Mika
- …

Why Are You Here?

- Probably because you need credits to finish your degree…

- You don’t have to “like” stats to recognize that it’s incredibly powerful
  - How most scientific findings are communicated and validated
  - Also easily manipulated/misused to give false impressions to the general public

More Pragmatism

- Both statistics and programming are general tools

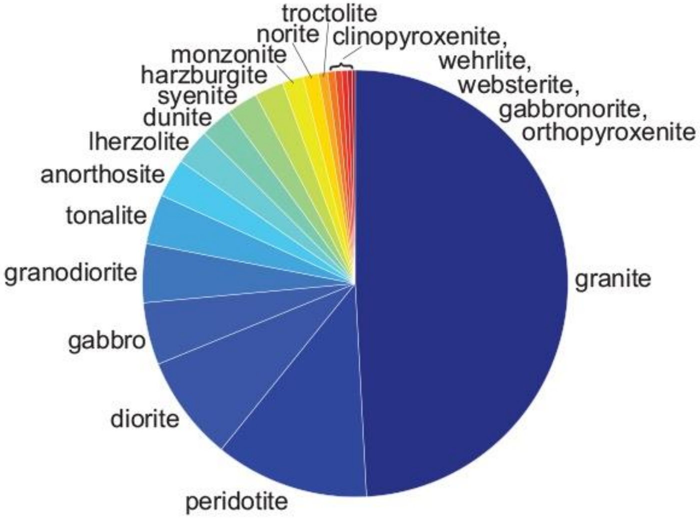
- Every industry uses those skills for some part of its operation!
  - Financial risk analysis
  - Data analyst
  - Research and development

- These are transferrable skills that will serve you well anywhere

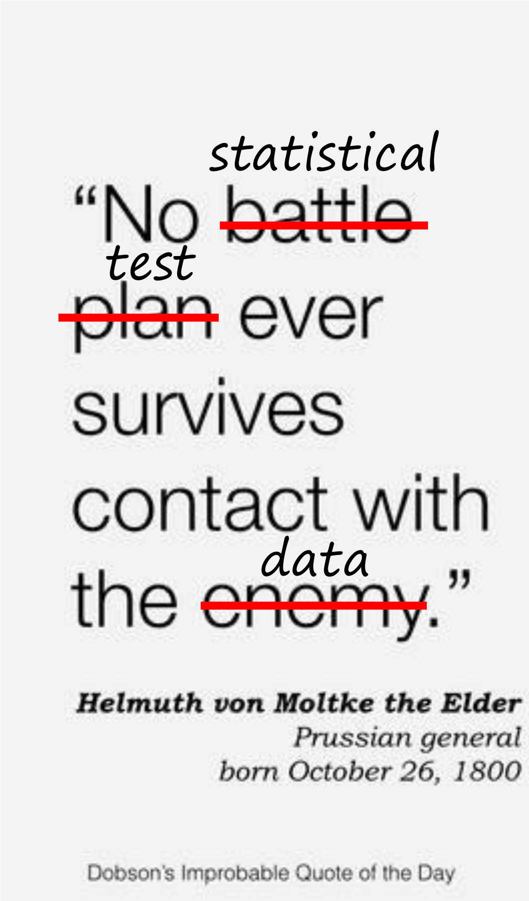
Statistics and Real Data

- Standard statistics make lots of assumptions about data that in the real world are almost always violated

- This is true in geology as well



Frequency distribution of documents using the given rock names in the GeoRef database, 1970-2018. The ten most common names account for more than 90% of the citations.



“No ~~battle~~ statistical ~~plan~~ test ever survives contact with the ~~enemy.~~ data”

Helmuth von Moltke the Elder
Prussian general
born October 26, 1800

Dobson's Improbable Quote of the Day

Statistical Issues in Geology

- Geology covers a wide variety of subjects and some of them have inherent difficulties from a statistical point of view

- We have only one Earth and one timeline (historical)
  - Cannot run another Earth to collect more samples, have to build models to simulate data

- Processes and events are sequential and related both in space and time
  - Autocorrelation, events are not independent

- There is a finite amount of data that can be collected for some problems (ex. rocks >3 Ga)
  - Sample size is inherently limited

Python and Jupyter Notebook

- We will use the Python programming language in this course for all our analyses

- Python is the most widely used high-productivity language in Scientific Computing. Its very **simple syntax** and broad library support make it ideal for quickly building scalable applications.

- A Jupyter Notebook is an open-source, web-based interactive environment for creating and sharing documents containing live code, equations, visualizations, and narrative text

python™

To GitHub!

- Please register for a GitHub account if you do not have one

- We’ll turn over to a brief look at GitHub and Jupyter Notebook for the rest of the lecture
  - Following the instruction
  - https://github.com/GeoData-Analysis-Spring-2026/CourseMaterials26/blob/main/coding_modules/GitHub_intro.pdf

Today’s Learning Outcomes

1. Know my expectations and pedagogical approach to the course
2. Know the course layout
3. Access GitHub CodeSpace and open a Jupyter Notebook