

# STA141B

# Data and Web Technologies for Data Analysis

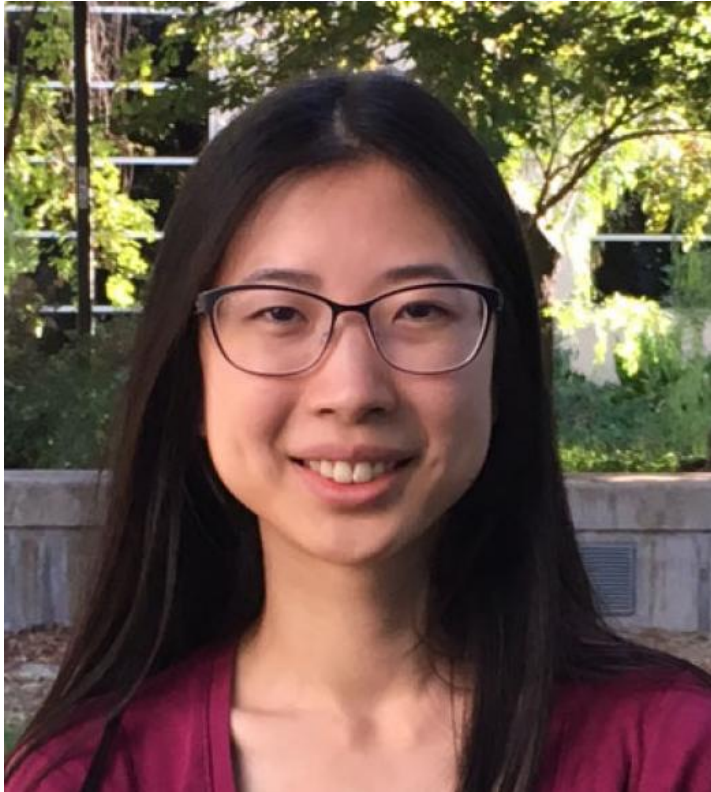
# Spring 2024

Instructor: Duncan Temple Lang

Quick Links: Canvas Piazza

Github: [https://github.com/duncantl/STA141B\\_S24](https://github.com/duncantl/STA141B_S24)

# TAs



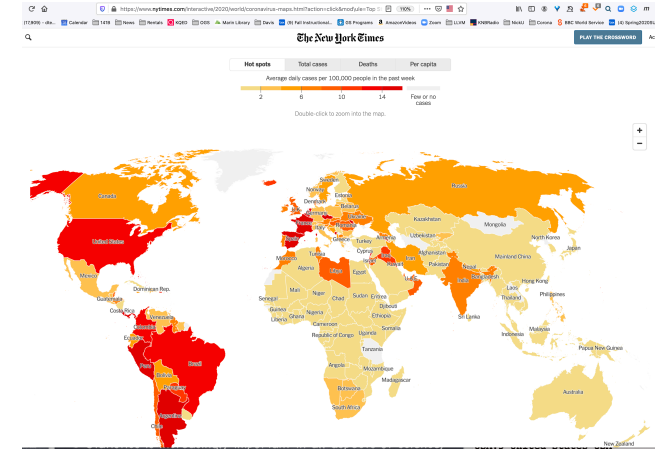
Emily Chang  
PhD Student  
Statistics



Yemisi Obasemo  
PhD Student  
Statistics

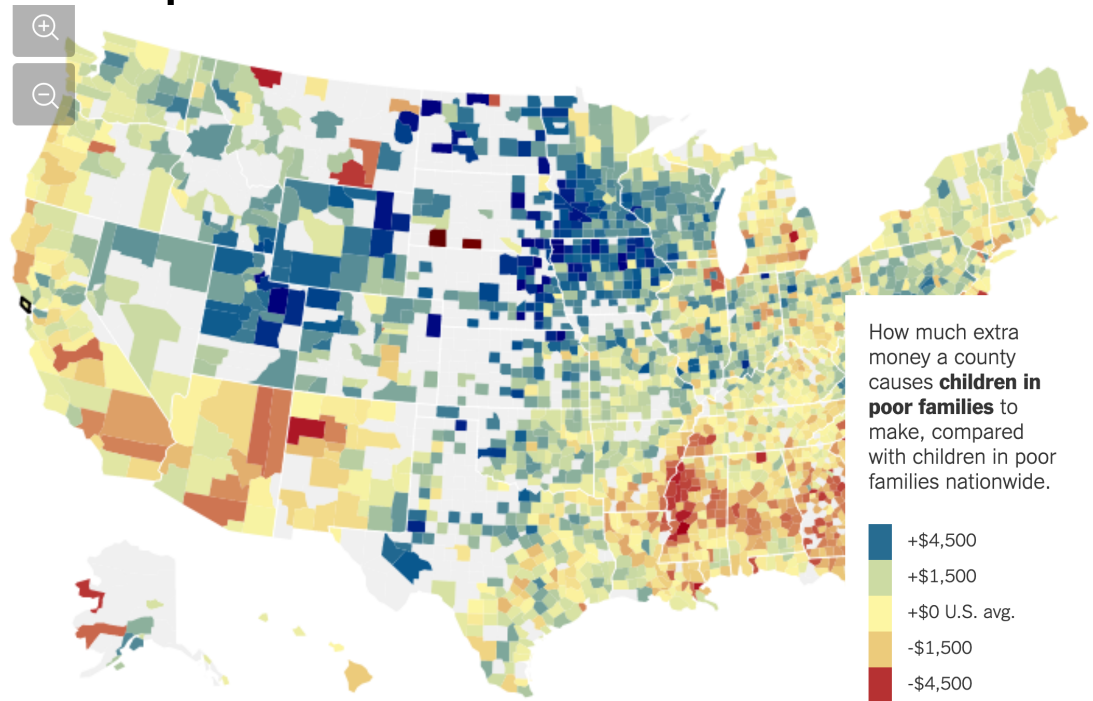
# Motivation

- So much more data available today
  - Online and from digital devices
- We want to be able to integrate datasets to
  - explore new questions
  - verify other people's claims and conclusions
  - provide different views & insights.



# Example

- Income Mobility
- Income disparity by county
- How do we get the data?
- How to integrate other data?
- Demographics
- Education
- Migration
- Employment
- More recent data...



<https://www.nytimes.com/interactive/2015/05/03/upshot/the-best-and-worst-places-to-grow-up-how-your-area-compares.html>

# StackOverflow - R Questions and Answers

<https://stackoverflow.com/questions/tagged/r>

- How do we programatically mine this for interesting information
- Names of popular functions
- People who provide good answers
- How to write good questions
- How do we get the data?

The screenshot shows the Stack Overflow interface for the 'r' tag. The left sidebar contains navigation links: Home, PUBLIC, Stack Overflow (selected), Tags, Users, FIND A JOB (Jobs, Companies), TEAMS, and What's this? (Free 30 Day Trial). The main content area is titled 'Questions tagged [r]' and includes a search bar, a 'Watch tag' button, and a list of 367,171 questions. The questions are sorted by 'Newest' and include filters for 'Active', 'Bountied', 'Unanswered', and 'More'. The first question is 'Return the multiply of a column' with 0 votes, 0 answers, and 10 views. The second question is 'How do I create a grid, extract mean raster values, and plot results?' with 0 votes, 0 answers, and 6 views. The third question is 'Delete incomplete cases in nested dataframe using map function from purrr' with 0 votes, 1 answer, and 6 views. The fourth question is 'What is note\_ind:ncol(dataset) mean in R?' with 0 votes, 0 answers, and 6 views.

stackoverflow Products

What's the deal with Deno? We talk with a major contributor to find out. [Listen now.](#)

Home PUBLIC **Stack Overflow** Tags Users FIND A JOB (Jobs Companies) TEAMS What's this? (Free 30 Day Trial)

## Questions tagged [r]

[Watch tag](#) [Ignore tag](#) [Learn more...](#) [Improve tag info](#) [Top users](#) [Synonyms \(2\)](#) [r jobs](#)

367,171 questions [Newest](#) [Active](#) [Bountied 8](#) [Unanswered](#) [More](#) [Filter](#)

0 votes 0 answers 10 views [Return the multiply of a column](#)  
In the below, the data frame index denotes the value while t1:t2 denotes the number of times that specific value was recorded at a specific point in time. For example index 10 at t1 equals 1 ...  
[r](#) [dataframe](#) asked 3 mins ago [user11418708](#) 328 • 1 • 7

0 votes 0 answers 6 views [How do I create a grid, extract mean raster values, and plot results?](#)  
I am trying to create a grid within a state boundary, extract mean elevation values within that grid, and then plot those average values within said grid. I have no clue if I am doing this right ...  
[r](#) [raster](#) asked 8 mins ago [fin](#) 1 • 1

0 votes 1 answer 6 views [Delete incomplete cases in nested dataframe using map function from purrr](#)  
I would like to delete incomplete cases from each dataframes of a nested tibble. I did try to use the map function (purrr package), but I received the following error message "Error in parent.env(..."  
[r](#) [tidyverse](#) [purrr](#) asked 14 mins ago [AmelV](#) 1

0 votes 0 answers 6 views [What is note\\_ind:ncol\(dataset\) mean in R?](#)  
I have this line of code but I don't know what it means especially the note\_ind part. apply(mydat[,c(1,2,3,note\_ind:ncol(dataset))],c(1,2),as.numeric)  
[r](#) asked 18 mins ago [123 456](#)

# Job Postings

<https://www.indeed.com/jobs?q=data+scientist&l=san+francisco>

The screenshot displays the Indeed job search interface. At the top, the search bar shows 'What data scientist' and 'Where san francisco'. Below the search bar, there are filters for 'Date Posted', 'Remote', 'within 25 miles', 'Salary Estimate', 'Job Type', 'Location', 'Company', 'Experience Level', and 'Education'. The search results are sorted by 'relevance - date' and show 'Page 1 of 1,719 jobs'.

The job listings include:

- Senior Data Scientist** at CyberCoders, San Francisco, CA 94103. Salary: \$150,000 - \$170,000 a year. Full-time. Just posted.
- Data Analyst** at Netomi, San Francisco, CA. Salary: \$80,000 - \$100,000 a year. Full-time. 14 days ago.
- Remote Data Scientist** at CyberCoders, Remote in San Francisco, CA 94104. Salary: \$150,000 - \$200,000 a year. Full-time. Just posted.
- Lead Data Scientist** at CyberCoders, San Francisco, CA. Salary: \$160,000 - \$190,000 a year. Full-time.

The detailed view of the 'Senior Data Scientist' position at CyberCoders shows the following details:

- Job details:** Salary: \$150,000 - \$170,000 a year. Job Type: Full-time.
- Benefits:** Disability insurance, Health insurance, Dental insurance, Paid time off, Vision insurance, Life insurance.
- Full Job Description:** Senior Data Scientist. If you are a remote Senior Data Scientist with experience, please read on! We are one of the fastest growing subscription based companies for health, wellness, and lifestyle products. We provide our members with a personally curated box delivered to their doorstep, streaming outlets for lifestyle tips, exclusive deals, and a community for members to connect and grow together! We currently have over one million members and we're looking to add a Senior Data Scientist to help with our growth!
- What You Will Be Doing:** As our Senior Data Scientist, you'll be tasked with utilizing your Machine Learning capabilities to help us better personalize our curated products and outreach for current and prospective members. This includes analyzing big data sets, creating data models, predictive analysis, and more! If you're an experience Senior Data Scientist, apply now!
- What You Need for this Position:**
  - Master's Degree or higher
  - 5+ years of Data Science experience
  - Machine Learning
  - Python
  - SQL
  - R
  - Databricks

- Look at job postings for different types of jobs
- Salary distribution
- What are typical required skills
- Level of education.
- Integrate with cost of living of cities.
- How do we get the data?

# Non-standard Data Formats

# timestamp=2006-02-11 08:31:58

# usec=250

# minReadings=110

t=1139643118358;id=00:02:2D:21:0F:33;pos=0.0,0.0,0.0;degree=0.0;00:14:bf:b1:97:8a=-38,2437000000,3;00:14:bf:b1:97:90=-56,2427000000,3;00:0f:a3:39:e1:c0=-53,2462000000,3;00:14:bf:b1:97:8d=-65,2442000000,3;00:14:bf:b1:97:81=-65,2422000000,3;00:14:bf:3b:c7:c6=-66,2432000000,3;00:0f:a3:39:dd:cd=-75,2412000000,3;00:0f:a3:39:e0:4b=-78,2462000000,3;00:0f:a3:39:e2:10=-87,2437000000,3;02:64:fb:68:52:e6=-88,2447000000,1;02:00:42:55:31:00=-84,2457000000,1

t=1139643118744;id=00:02:2D:21:0F:33;pos=0.0,0.0,0.0;degree=0.0;00:14:bf:b1:97:8a=-38,2437000000,3;00:0f:a3:39:e1:c0=-54,2462000000,3;00:14:bf:b1:97:90=-56,2427000000,3;00:14:bf:3b:c7:c6=-67,2432000000,3;00:14:bf:b1:97:81=-66,2422000000,3;00:14:bf:b1:97:8d=-70,2442000000,3;00:0f:a3:39:e0:4b=-79,2462000000,3;00:0f:a3:39:dd:cd=-73,2412000000,3;00:0f:a3:39:e2:10=-83,2437000000,3;02:00:42:55:31:00=-85,2457000000,1

# More Complex Formats

## Player version 2.1.3

## File version 0.3.0

## Format:

## - Messages are newline-separated

## - Common header to each message is:

## time host robot interface index type subtype

## (double) (uint) (uint) (string) (uint) (uint) (uint)

## - Following the common header is the message payload

0000000000.100 16777343 6668 laser 00 004 001 +0.000 +0.000 0.000 0.156 0.155

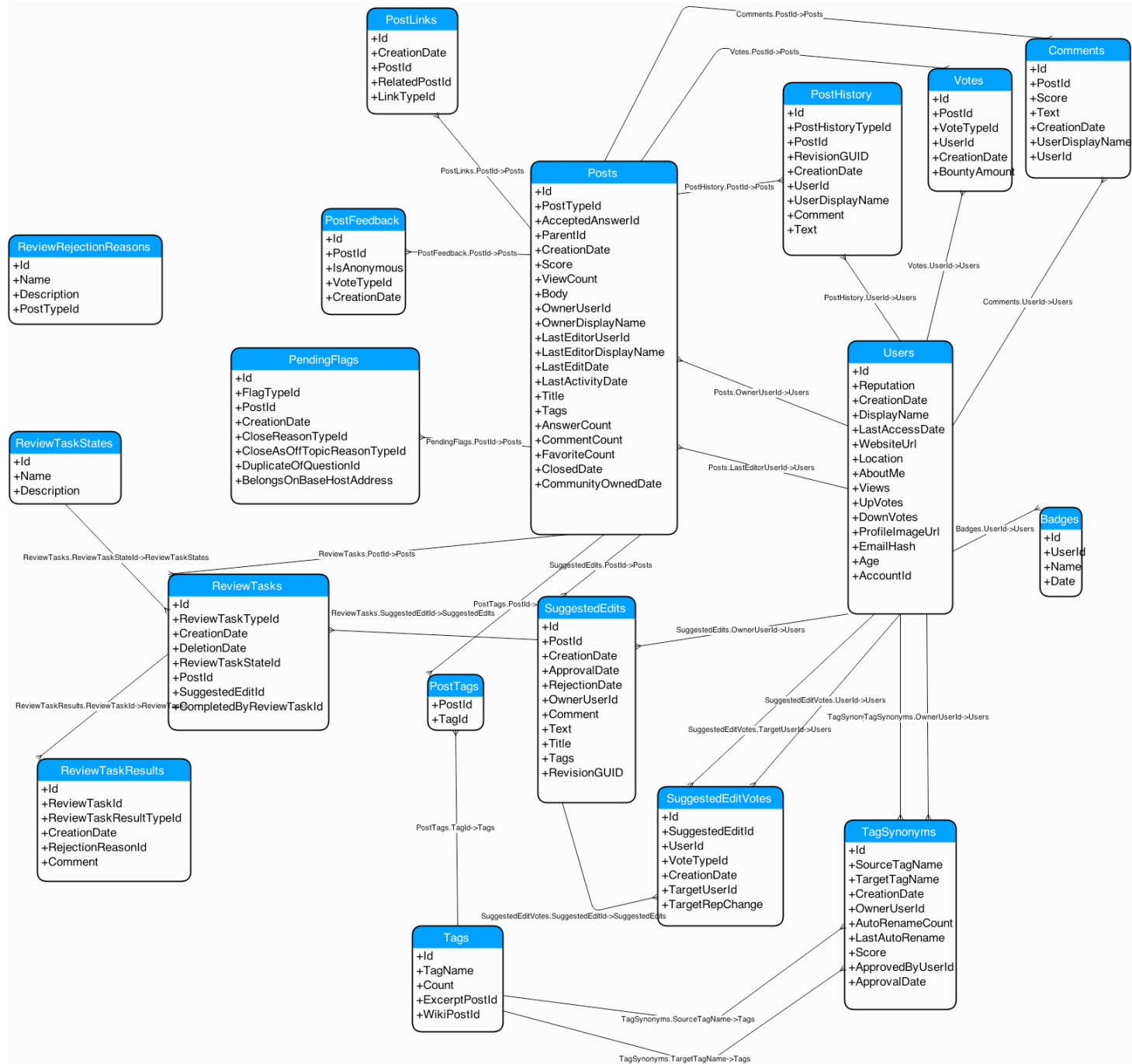
0000000000.200 16777343 6668 position2d 00 004 001 -00.040 +00.000 +0.000 +00.440 +00.380

0000000000.200 16777343 6668 position2d 00 001 001 -14.000 -07.000 +0.785 +00.000 +00.000 +00.000 0

0000000000.200 16777343 6668 laser 00 001 001 0001 -3.1416 +3.1416 +0.01740495 +2.0000 0361 1.838 0 1.807 0 1.778 0 1.749 0 1.723 0  
1.697 0 1.673 0 1.650 0 1.628 0 1.607 0 1.587 0 1.568 0 1.550 0 1.533 0 1.517 0 1.501 0 1.486 0 1.472 0 1.459 0 1.446 0 1.434 0 1.423 0  
1.412 0 1.402 0 1.392 0 1.383 0 1.375 0 1.367 0 1.359 0 1.352 0 1.346 0 1.340 0 1.334 0 1.329 0 1.324 0 1.320 0 1.316 0 1.313 0 1.310 0  
1.307 0 1.305 0 1.303 0 1.302 0 1.301 0 1.300 0 1.300 0 1.300 0 1.301 0 1.302 0 1.303 0 1.305 0 1.307 0 1.310 0 1.313 0 1.316 0 1.320 0  
1.324 0 1.329 0 1.334 0 1.340 0 1.346 0 1.352 0 1.359 0 1.367 0 1.375 0 1.383 0 1.392 0 1.402 0 1.412 0 1.423 0 1.434 0 1.446 0 1.459 0  
1.472 0 1.486 0 1.501 0 1.517 0 1.533 0 1.550 0 1.568 0 1.587 0 1.607 0 1.628 0 1.650 0 1.673 0 1.697 0 1.723 0 1.749 0 1.778 0 1.807 0  
1.838 0 1.843 0 1.905 0 1.941 0 2.000 0 2.000 0 2.000 0 2.000 0 2.000 0 2.000 0 2.000 0 2.000 0 2.000 0 2.000 0 2.000 0  
2.000 0 2.000 0 2.000 0 2.000 0 2.000 0 2.000 0 2.000 0 2.000 0 2.000 0 2.000 0 2.000 0 2.000 0 2.000 0 2.000 0  
2.000 0 2.000 0 2.000 0 2.000 0 2.000 0 2.000 0 2.000 0 2.000 0 2.000 0 2.000 0 2.000 0 2.000 0 2.000 0 2.000 0  
2.000 0 2.000 0 2.000 0 2.000 0 2.000 0 2.000 0 2.000 0 2.000 0 2.000 0 2.000 0 2.000 0 2.000 0 2.000 0 2.000 0  
0 1.669 0 1.683 0 1.701 0 1.716 0 1.738 0 1.765 0 1.774 0 1.811 0 1.817 0 1.852 0 1.864 0 1.885 0 1.921 0 1.941 0 1.978 0 2.000 0 1.944 0  
1.889 0 1.835 0 1.838 0



# Relational Databases



# Complexity

- Data from many different types of sources
  - Web pages, Web services, databases, FTP servers, github, ...
- in many different formats
  - CSV, Excel, XML, JSON, YAML, PDF, ...
  - Structured, semi-structured and free-form text.
- Different Domain Specific Languages to manipulate data - SQL, XPath, Regular Expressions, CSS selectors

# Learning Goals

- Learn important and commonly used technologies to access and manipulate data.
- Learn fundamental concepts and data technology architectures so you can quickly embrace new technologies
- Be able to get data from many different sources and formats.
- Strengthen and Master programming knowledge
- Deeper experience with R
- Computational problem solving.

# Syllabus/Topics

- R fundamentals - reading and restructuring non-standard data.
- Data Extraction and Text Manipulation via Regular Expressions
- Relational Databases & SQL
- Web Scraping & APIs (Application Programming Interfaces) - HTTP requests, JSON, XML
- Interactive Data Visualization - Web-based, HTML, JavaScript, SVG, CSS from R.

# Additional/Optional Topics

- If we have time and you are interested
  - Advanced aspects of R
  - UNIX shell basics
  - Version Control (git)

# Making sense of the Data

- The focus is on computing, but we also take this opportunity to explore real data and to do common sense data analysis.
- This does not necessarily mean using complex, sophisticated statistical methodology (unless it is appropriate).
- Focus is more about summarizing the data and finding evidence within data and illustrating your conclusions,
- Or identifying conjectures/hypotheses and exploring with real data.

# Textbook

- No single text book
  - I'll point you to chapters of different books
  - But much of the material will be online Web sites & resources
- You have to use the Web and different resources to find information you need.
- This is a very important but highly non-trivial skill
  - composing the question/goal
  - finding resources
  - honing queries (Web or human) to get the relevant information
  - knowing when to detour and when not to

# Communicating

- Feel free to call me Duncan.
  - My last name is “Temple Lang”
- Ask all questions about the course (content, logistics, ...) on Piazza.
  - Make them public, not private.
- Send private/personal emails to me at [dtemplelang@ucdavis.edu](mailto:dtemplelang@ucdavis.edu)



# Lectures

- I want you to raise questions and discuss problems, questions, concepts in class.
- The beginning of every class, I ask for questions. I expect there to be some. If not, then you are not working on the assignments.
- If I say something you don't understand, but you have tried to follow, ask me to explain it a different way.
- If I speak too quickly, ask me to slow down.

# Assignments as Labs

- ④ *Lecture courses offer the opportunity to show whether you can get the right answer. Labs on the other hand offer an essential opportunity for students to learn about the practice of science and this practice includes presenting one's work in a clear and compelling fashion. – Moskovitz & Kellogg. Science, 29 July, 2011.*
- ④ This is a lab class with guidance/instructions in lectures.
- ④ You need to start working on each lab when I post it
  - ④ They take time
  - ④ Don't wait until a day or two before the due date.
  - ④ Way too much stress and you won't learn much.

# Grading

- 5 assignments - 90%. (18% each, all count)
- Each about 2 weeks.
- 10% participation
  - Piazza, office hours, lecture
  - asking and answering questions

# Exams, Quizzes

- Hands-on computing is difficult to assess via written exams & quizzes
  - So prefer to evaluate by assignments.
- However, if there is significant copying in assignments or cheating, I'll conduct exams.
  - Please don't copy or cheat
  - No benefit to you in medium- and long-term

# Copying & Cheating

- Read and understand the Code of Academic Conduct
- You ***can*** use
  - Ideas and code you find on the Web
  - Code snippets posted on Piazza
  - YOU MUST NOTE IN YOUR CODE AND REPORT WHERE YOU GOT THESE!

# Piazza

- We'll use Piazza for all questions and answers about assignments, code, lecture, ...
- Page: <https://piazza.com/ucdavis/spring2024/sta141b>
- Should already be registered
  - login ASAP
- Change your settings to get notifications of posts
- Please post using your login (not anonymously)
  - Helps for participation grade.

# Before Posting a Question

- Make certain to read all the posts - regularly
  - Part of the course learning.
  - You will learn a lot and find suggestions about how to approach aspects of the assignments.
- Don't repeat a question, i.e., that was asked previously.
- Try to find the answer or some background information before simply expecting others to look it up for you.

# Posting Questions

- Much more likely to get a good answer quickly if you pose a question well
- Focus on what information is necessary for a reader who isn't sitting in front of your computer
  - Enough to give all the relevant details
  - Not too much so reader can't see the relevant element
    - And doesn't bother to reply.



# Posing a Good Question

- Minimal reproducible example
  - Take the time to create a much simpler version of only the part of your code that exhibits the problem.
  - Doing this will often be enough to solve the problem yourself.
    - That's where you learn for the future.
- In the question, state
  - what you expect the code to do and what results you expect it to produce
  - what it actually produces
  - In what ways these are different.
  - Include any error or warning message
  - And the output from `sessionInfo()`

# Bad Question

- “I tried to read the data and it didn’t work”
  - What data - a file, a URL? What format? Does the file actually exist?....
  - What’s “it”?
  - In what way didn’t it “work”?