Data Science

Grayson White

Math 241

Week 1 | Spring 2026

Goals for Today (and next time)

Today:

- Getting started in Math 241
  - Course structure and technologies
  - Where to find resources
  - Course expectations

Wednesday:

- Decomposing graphics

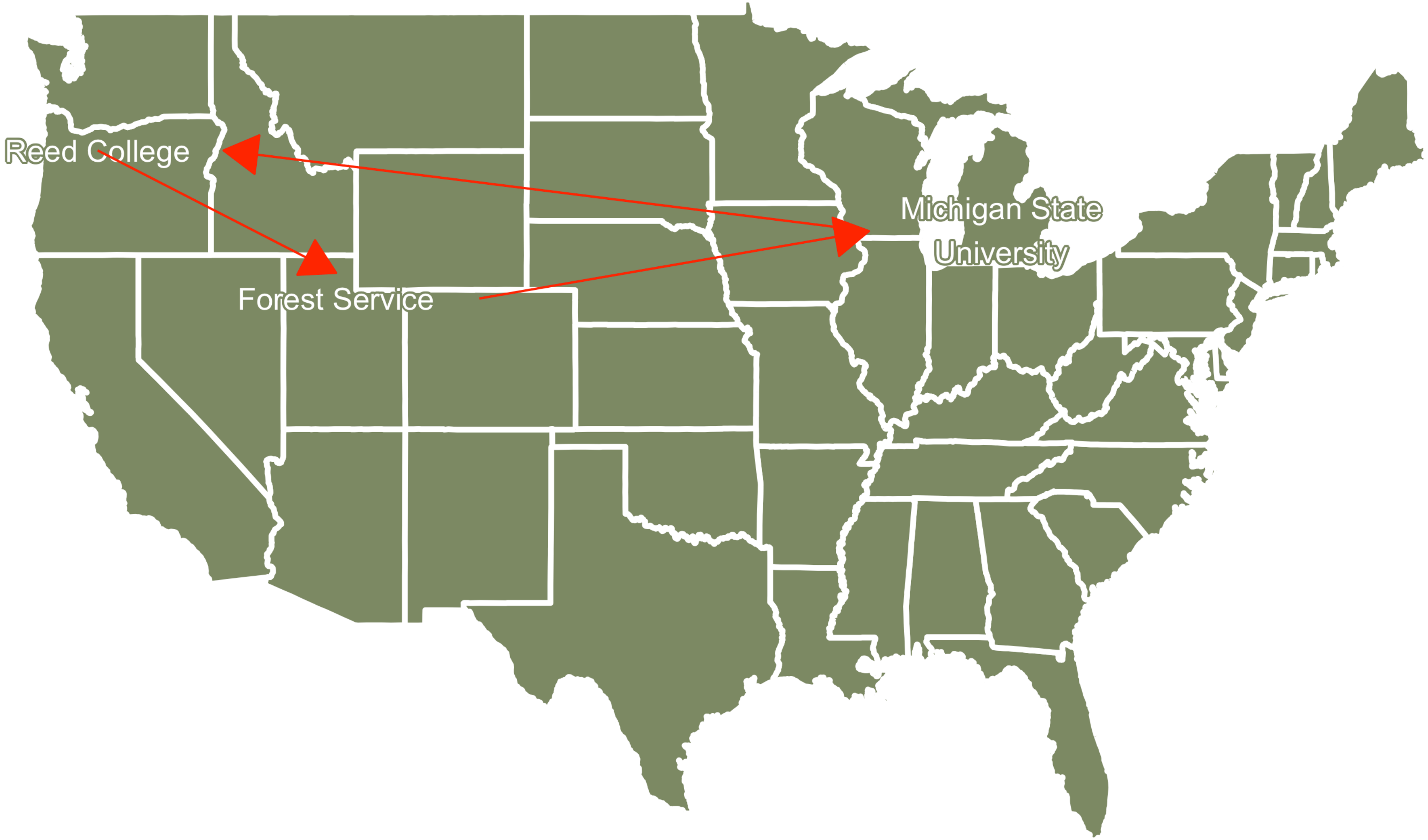
But first, let me quickly introduce myself...

…and after that I’ll have each of you introduce yourselves 😄

My info

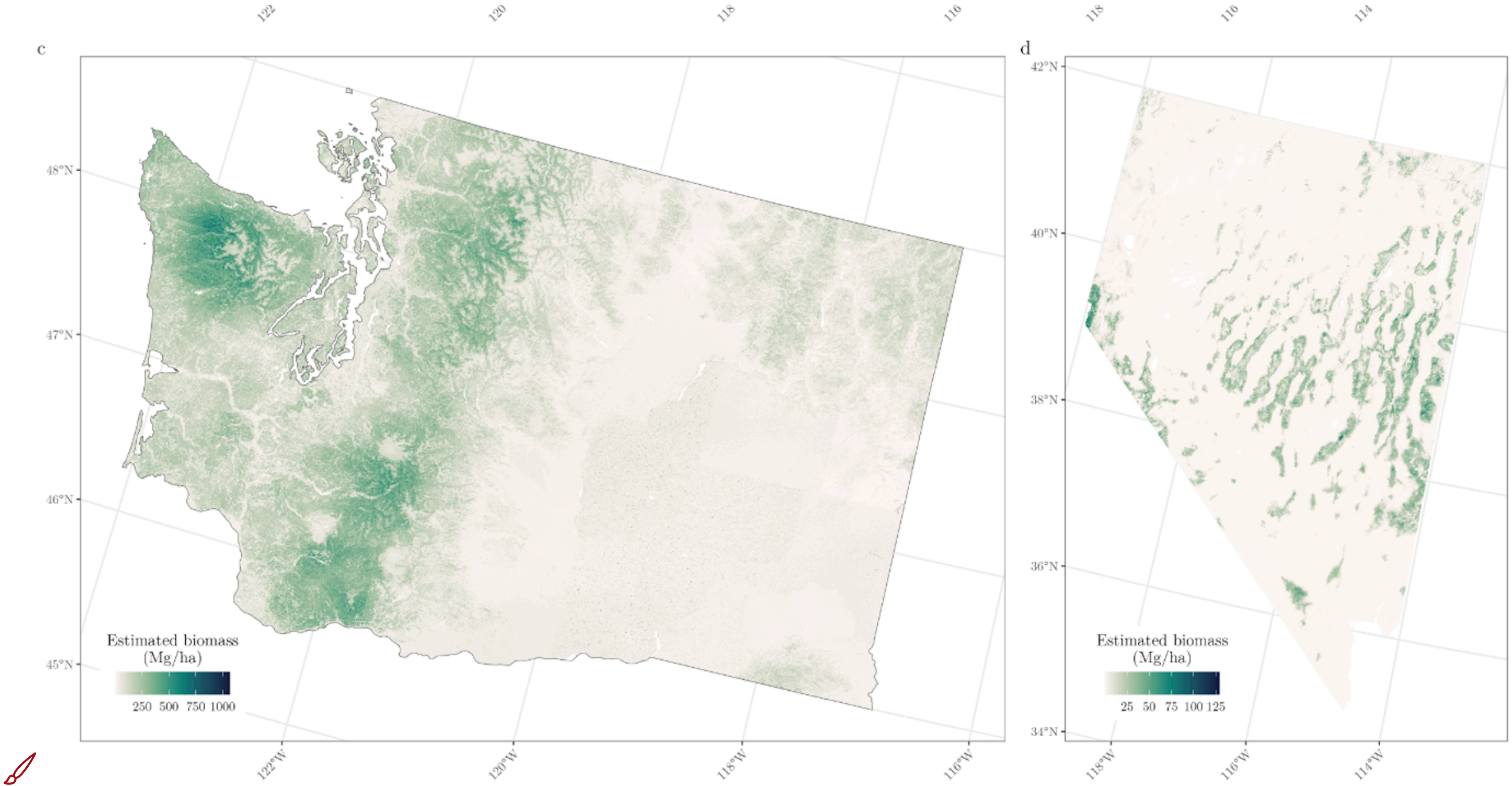
- **Name**: Grayson White (he/him), you can call me ‘Grayson’
- **Email**: gwhite@reed.edu
- **Office**: Library 390
- **Office hours**: TBD, please fill out the office hours survey ASAP for your availability to be considered.
  - This week, I’ll hold office hours on Tuesday 1pm - 3pm and Friday 8am - 10am, or by appointment.

Let's start with my path (back) to Reed...



Research Interests

Statistical modeling with environmental applications



Research Interests

Advising Undergraduate Forestry Data Science Research

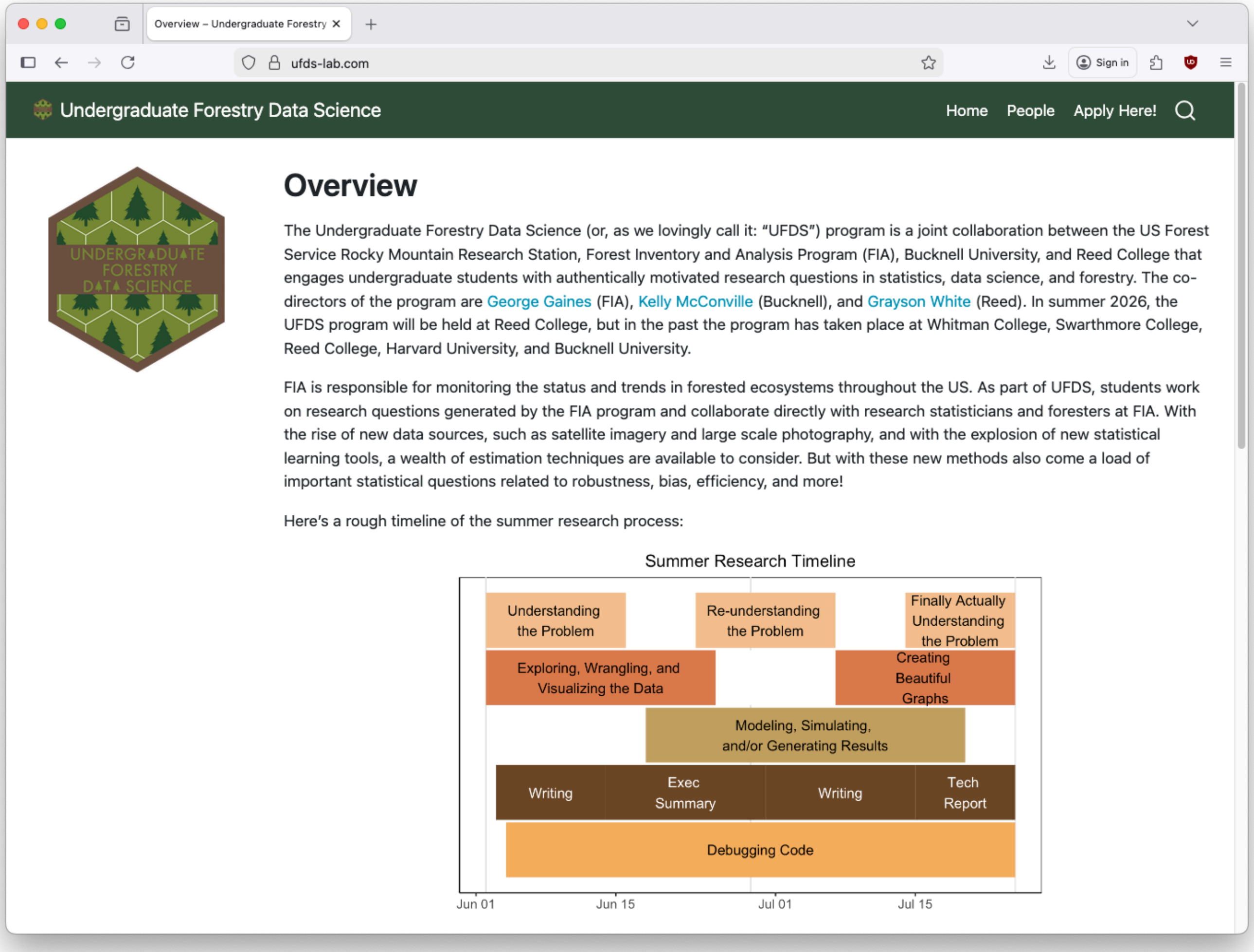
Sidebar: Undergraduate Forestry Data Science Research

This summer, I'll be hiring Reed students in my undergraduate forestry data science (UFDS) program!

- UFDS is a long-running collaboration with the US Forest Service where we answer statistical and data science research questions for the US Forest Service.
- Students will work on problems generated by the Forest Service and will collaborate directly with Research Statisticians and Foresters at the Forest Service.
- Co-directed by myself and Kelly McConville (Bucknell University), this summer's research will occur on Reed's campus and will include students from both colleges!

Sidebar: Undergraduate Forestry Data Science Research

- For more info, tentative projects, FAQs, and to apply see the website ufds-lab.com


Overview – Undergraduate Forestry
ufds-lab.com
Sign in
Undergraduate Forestry Data Science
Home
People
Apply Here!
UNDERGRADUATE FORESTRY DATA SCIENCE
Summer Research Timeline
Understanding the Problem
Re-understanding the Problem
Finally Actually Understanding the Problem
Exploring, Wrangling, and Visualizing the Data
Creating Beautiful Graphs
Modeling, Simulating, and/or Generating Results
Writing
Exec Summary
Writing
Tech Report
Debugging Code
Jun 01
Jun 15
Jul 01
Jul 15

Overview

The Undergraduate Forestry Data Science (or, as we lovingly call it: "UFDS") program is a joint collaboration between the US Forest Service Rocky Mountain Research Station, Forest Inventory and Analysis Program (FIA), Bucknell University, and Reed College that engages undergraduate students with authentically motivated research questions in statistics, data science, and forestry. The co-directors of the program are George Gaines (FIA), Kelly McConville (Bucknell), and Grayson White (Reed). In summer 2026, the UFDS program will be held at Reed College, but in the past the program has taken place at Whitman College, Swarthmore College, Reed College, Harvard University, and Bucknell University.

FIA is responsible for monitoring the status and trends in forested ecosystems throughout the US. As part of UFDS, students work on research questions generated by the FIA program and collaborate directly with research statisticians and foresters at FIA. With the rise of new data sources, such as satellite imagery and large scale photography, and with the explosion of new statistical learning tools, a wealth of estimation techniques are available to consider. But with these new methods also come a load of important statistical questions related to robustness, bias, efficiency, and more!

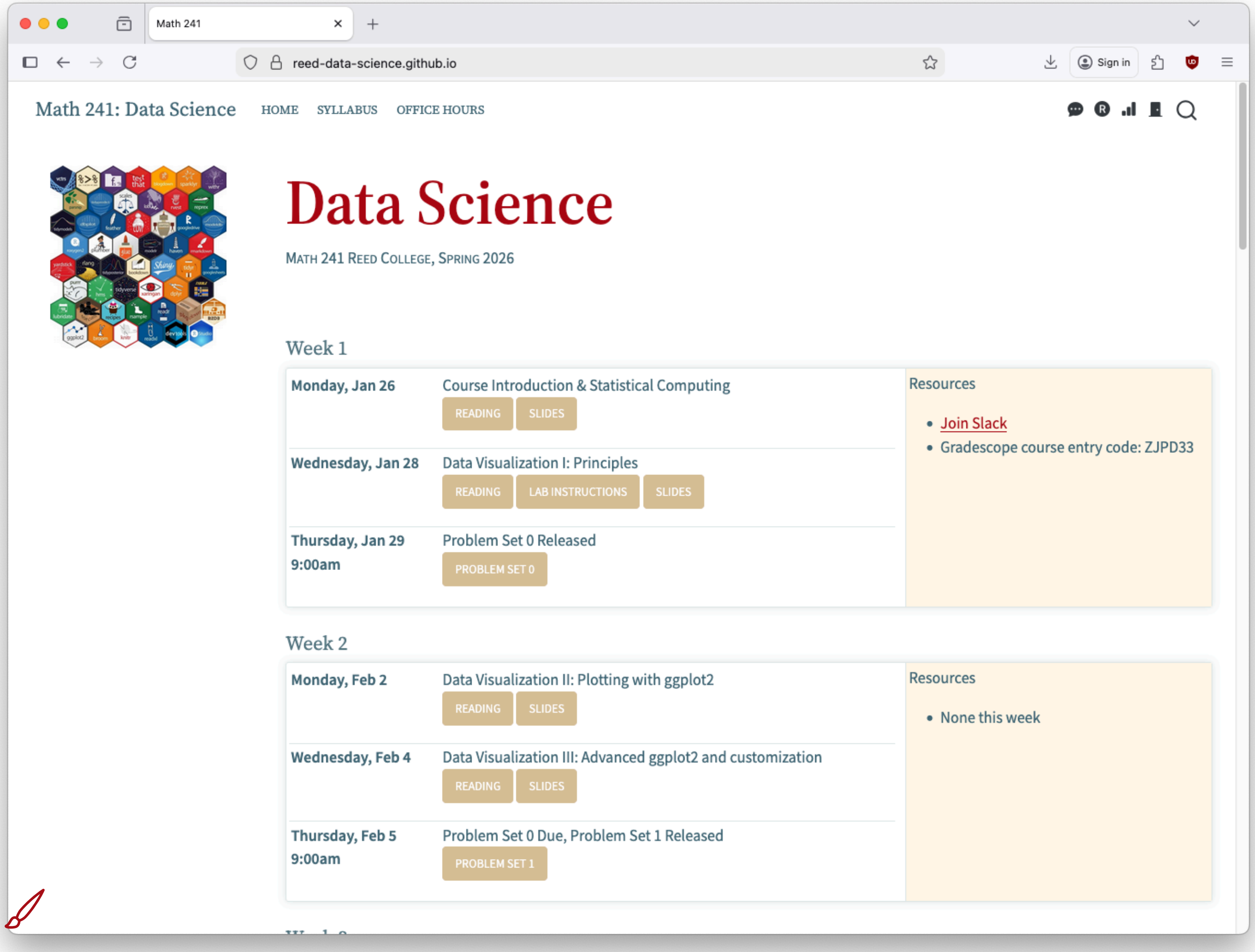
Here's a rough timeline of the summer research process:

Introductions!

Would love to hear: your name, pronouns, major, what you’re excited about in this class, and a fun fact!

Getting Started in Math 241

Course website



Math 241: Data Science HOME SYLLABUS OFFICE HOURS

Data Science

MATH 241 REED COLLEGE, SPRING 2026

Week 1

Monday, Jan 26	Course Introduction & Statistical Computing READING SLIDES	Resources - Join Slack - Gradescope course entry code: ZJPD33
Wednesday, Jan 28	Data Visualization I: Principles READING LAB INSTRUCTIONS SLIDES	
Thursday, Jan 29 9:00am	Problem Set 0 Released PROBLEM SET 0	

Week 2

Monday, Feb 2	Data Visualization II: Plotting with ggplot2 READING SLIDES	Resources - None this week
Wednesday, Feb 4	Data Visualization III: Advanced ggplot2 and customization READING SLIDES	
Thursday, Feb 5 9:00am	Problem Set 0 Due, Problem Set 1 Released PROBLEM SET 1	

- The course website, reed-data-science.github.io, will be the central location for all our course materials.
- We'll also use some other technologies and resources for collaboration, dissemination, and communication.

Getting Started in Math 241

Other Resources

Reed's **RStudio Server** or a **local installation of RStudio**, for completing coursework,

A course-wide **Slack** workspace, for course communication,

Gradescope, for turning in assignments and exams, and

Our course GitHub organization, for collaboration, portfolio-building, and dissemination of work.

Let’s take a look at the course website…

reed-data-science.github.io

What’s this course about?

And what even is *data science*?

What is data science?

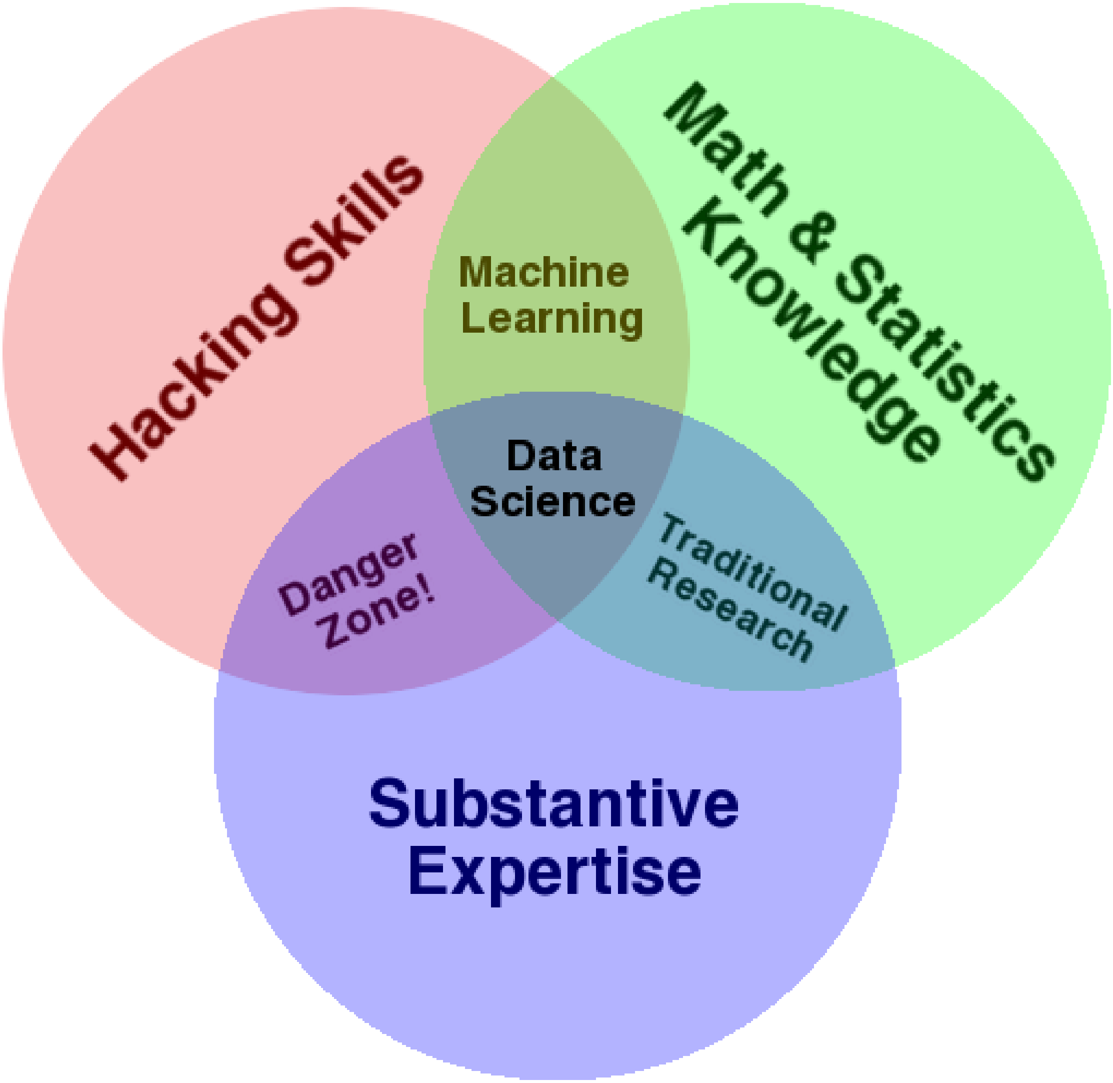
- How is **data science** not the same thing as **statistics**?

- Wikipedia says:

> “Data science is an inter-disciplinary field that uses scientific methods, processes, algorithms and systems to **extract knowledge** and insights from structured and unstructured **data**.”

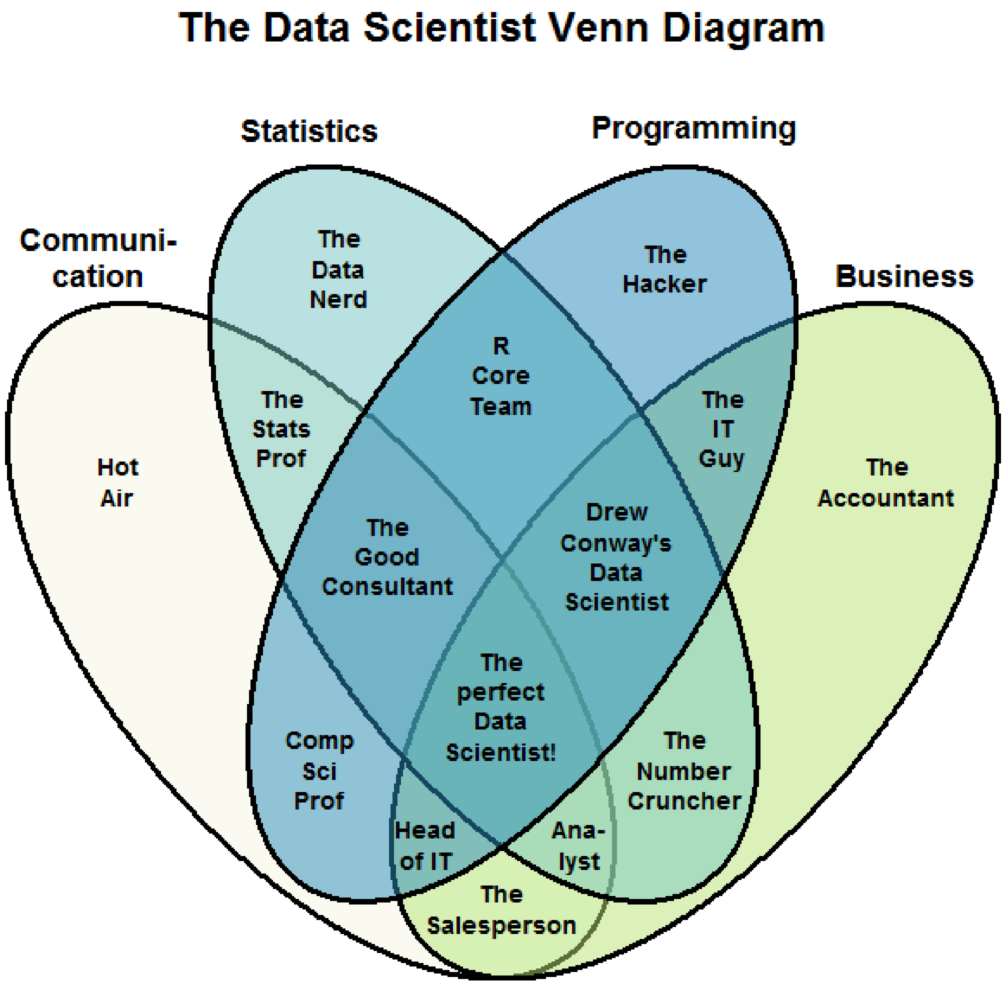
- But isn’t statistics the field of extracting knowledge from data?

Data Science: The Original Venn Diagram Definition



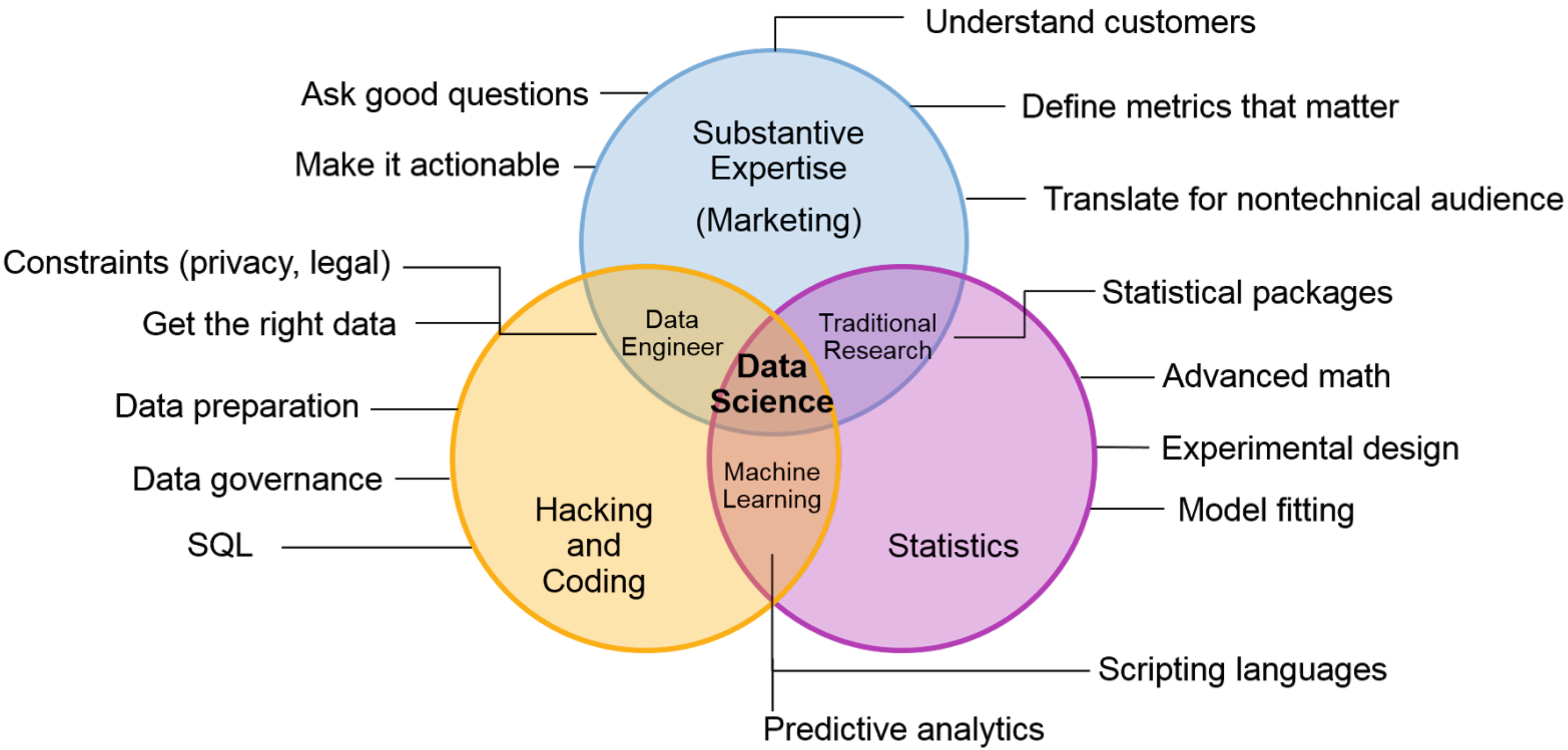
Drew Conway (2012)

Data Science: And the Chorus of Follow-up Venn Diagrams



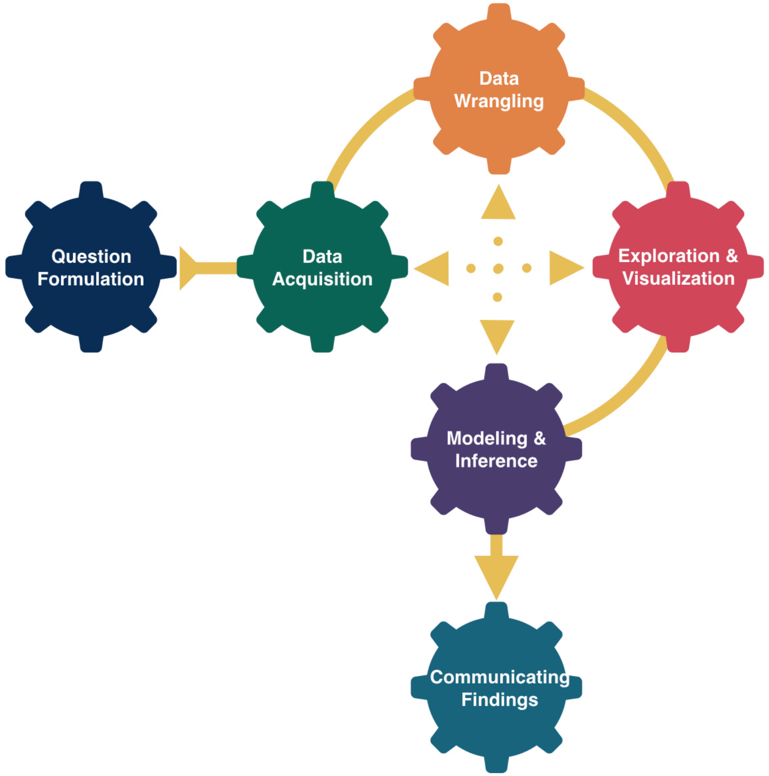
Stephan Kolassa (2015)

Data Science: And the Chorus of Follow-up Venn Diagrams



Gartner (2016)

Defining Data Science through the Data Analysis Workflow



- Is the data analysis workflow different for data scientists than it is for statisticians?

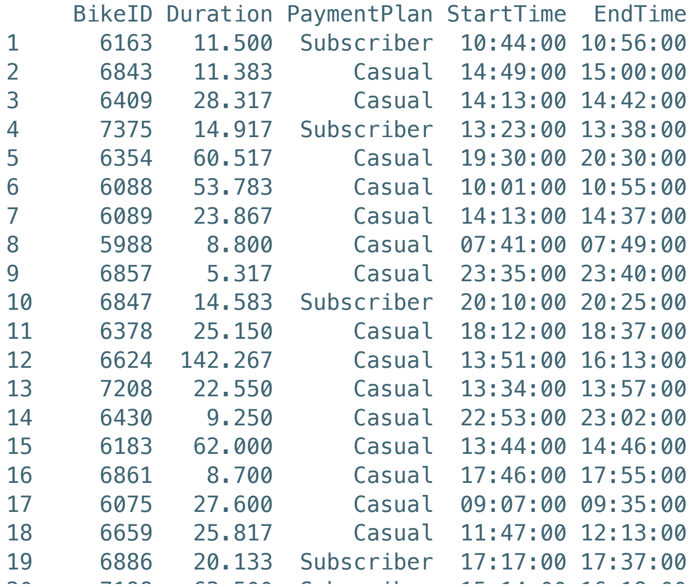
> *“A data scientist is a statistician with a MacBook Pro.”*

Learning Goals

Goal: Wrangle and interact with a variety of data types

Requires expanding our understanding of **data structures** in R.

- Most common data structure = `data.frame`/data set/spreadsheet
- Example: Portland Biketown bike rental data.



```
     BikeID Duration PaymentPlan StartTime  EndTime
1      6163   11.500  Subscriber  10:44:00 10:56:00
2      6843   11.383      Casual  14:49:00 15:00:00
3      6409   28.317      Casual  14:13:00 14:42:00
4      7375   14.917  Subscriber  13:23:00 13:38:00
5      6354   60.517      Casual  19:30:00 20:30:00
6      6088   53.783      Casual  10:01:00 10:55:00
7      6089   23.867      Casual  14:13:00 14:37:00
8      5988    8.800      Casual  07:41:00 07:49:00
9      6857    5.317      Casual  23:35:00 23:40:00
10     6847   14.583  Subscriber  20:10:00 20:25:00
11     6378   25.150      Casual  18:12:00 18:37:00
12     6624  142.267      Casual  13:51:00 16:13:00
13     7208   22.550      Casual  13:34:00 13:57:00
14     6430    9.250      Casual  22:53:00 23:02:00
15     6183   62.000      Casual  13:44:00 14:46:00
16     6861    8.700      Casual  17:46:00 17:55:00
17     6075   27.600      Casual  09:07:00 09:35:00
18     6659   25.817      Casual  11:47:00 12:13:00
19     6886   20.133  Subscriber  17:17:00 17:37:00
```

- Rows = observations
- Columns = variables
  - Two types: categorical and quantitative

- What variables **don't** fit neatly into these two types?
- What other **data structures** are there?

Variable types: Dates and Times

- Are dates and times categorical or quantitative?
- What makes dates and times different?

```r
biketown %>%
  select(StartTime, EndTime) %>%
  as.data.frame()
```

```
    StartTime  EndTime
1    10:44:00 10:56:00
2    14:49:00 15:00:00
3    14:13:00 14:42:00
4    13:23:00 13:38:00
5    19:30:00 20:30:00
6    10:01:00 10:55:00
7    14:13:00 14:37:00
8    07:41:00 07:49:00
9    23:35:00 23:40:00
10   20:10:00 20:25:00
11   18:12:00 18:37:00
12   13:51:00 16:13:00
13   13:34:00 13:57:00
14   22:53:00 23:02:00
15   13:44:00 14:46:00
16   17:46:00 17:55:00
17   09:07:00 09:35:00
18   11:47:00 12:13:00
19   17:17:00 17:37:00
```

- Will learn to use the `lubridate` package to wrangle dates and times.

Variable types: Factors and Characters

- R typically stores categorical variables as one of two types: `factor` or `character`.

```r
head(biketown$PaymentPlan)
```

```
[1] "Subscriber" "Casual"     "Casual"     "Subscriber" "Casual"
[6] "Casual"
```

```r
class(biketown$PaymentPlan)
```

```
[1] "character"
```

- Why do we need two different types??

- Will learn to use the `forcats` package to wrangle `factor`s.

- Will learn to use the `stringr` package to wrangle `character`s/strings/text.
  - Note: A character vector stores multiple strings/text.

What other data structures are there beyond `data.frames`?

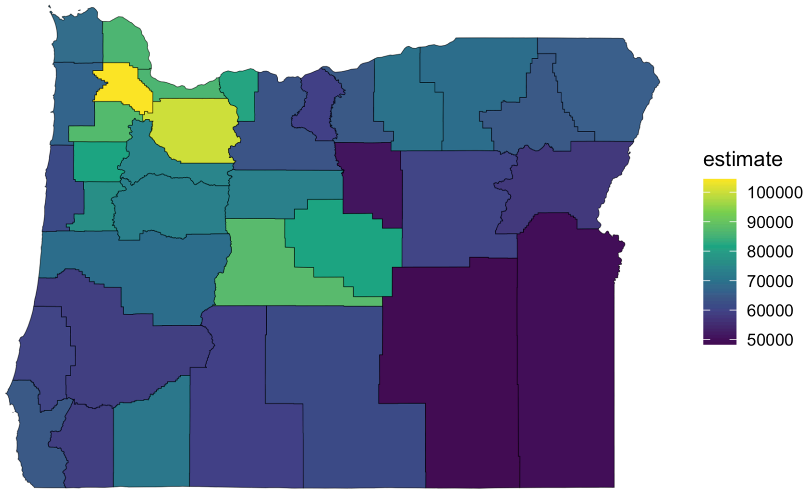
Data Structures: Spatial Data Frames

- Example: American Community Survey Data on median household income for counties in Oregon

```
Simple feature collection with 6 features and 5 fields
Geometry type: MULTIPOLYGON
Dimension:     XY
Bounding box:  xmin: -124.5662 ymin: 41.99562 xmax: -116.7837 ymax: 46.29083
Geodetic CRS:  NAD83
  GEOID                     NAME   variable estimate  moe
1 41015     Curry County, Oregon B19013_001    64769 5572
2 41045   Malheur County, Oregon B19013_001    49902 4655
3 41001     Baker County, Oregon B19013_001    57844 4135
4 41017 Deschutes County, Oregon B19013_001    87640 2429
5 41047    Marion County, Oregon B19013_001    74624 1593
6 41007   Clatsop County, Oregon B19013_001    68705 3596
                        geometry
1 MULTIPOLYGON (((-124.3239 4...
```

Data Structures: Spatial Data Frames

- Example: American Community Survey Data



- Will learn to use the `sf` package for wrangling spatial data

Data Structures: Lists

Lists are a more flexible structure for storing data!

```r
got
```

```
[[1]]
[[1]]$name
[1] "Theon Greyjoy"

[[1]]$gender
[1] "Male"

[[1]]$culture
[1] "Ironborn"


[[2]]
[[2]]$name
[1] "Tyrion Lannister"

[[2]]$gender
[1] "Male"

[[2]]$culture
```

- From a data analysis perspective, they are hard to work with!

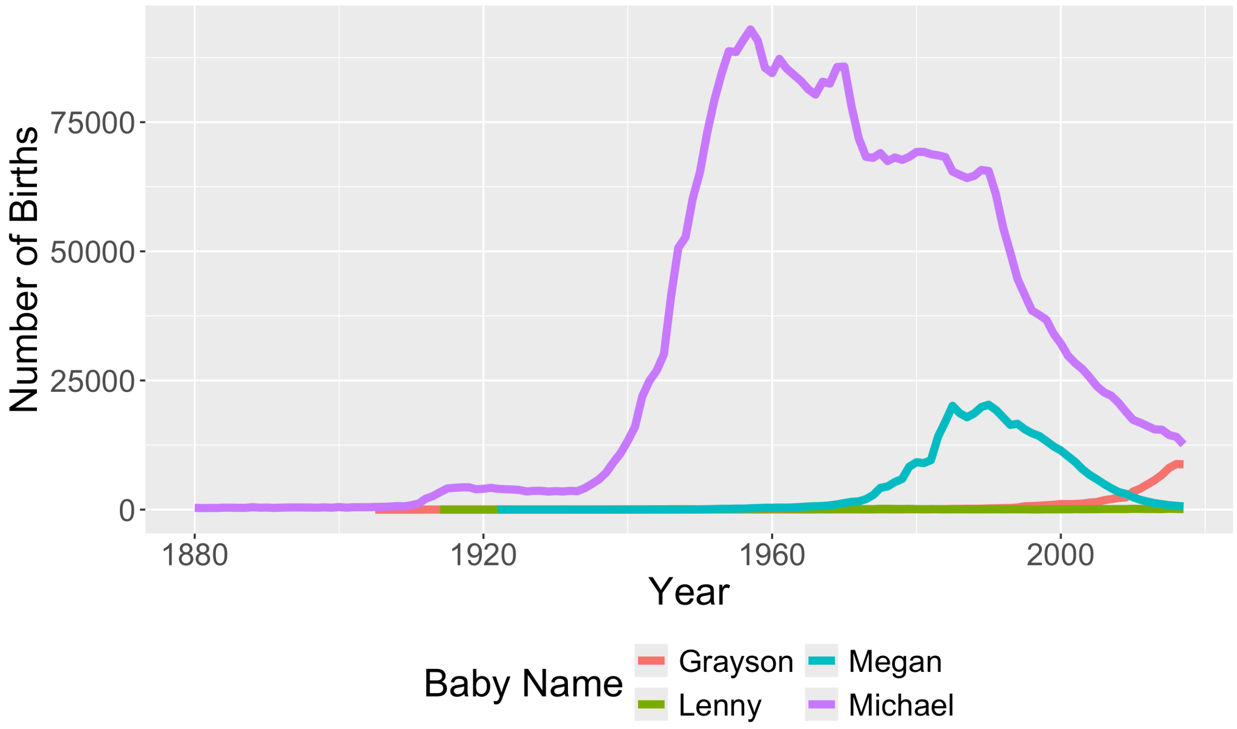
- Will learn to use the purrr package to converting nested lists into data.frames.

```r
map_dfr(got, `[`, c("name", "gender", "culture"))
```

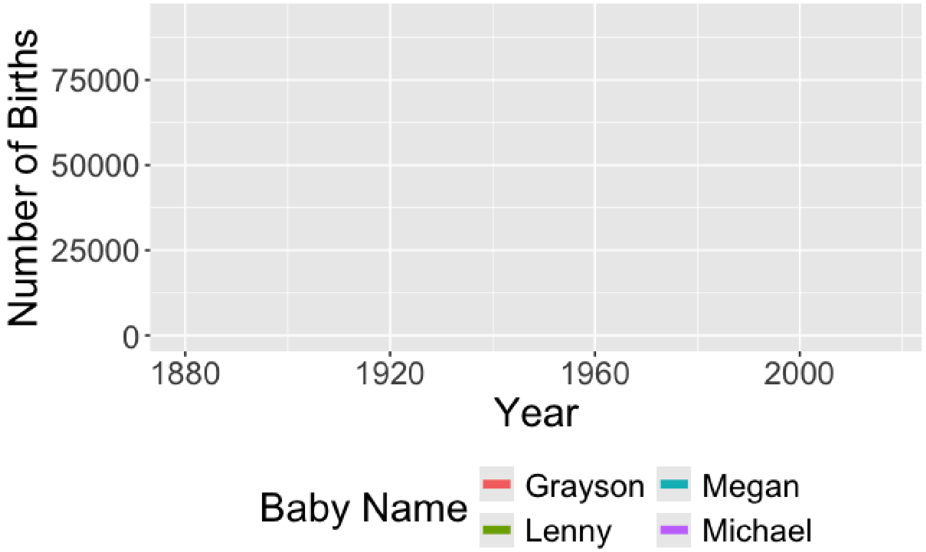
```
# A tibble: 2 × 3
  name             gender culture
  <chr>            <chr>  <chr>
1 Theon Greyjoy    Male   "Ironborn"
2 Tyrion Lannister Male   ""
```

Learning Goals: Data Viz

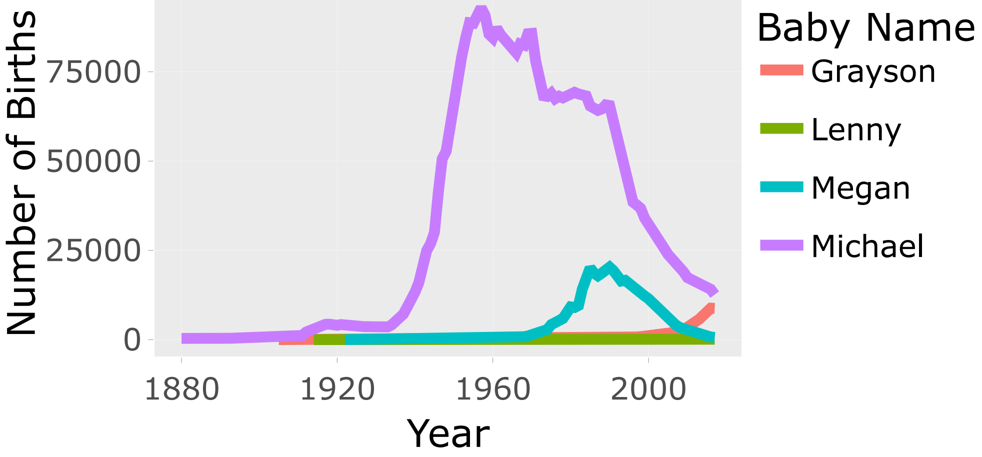
Goal: Create static data visualizations of multivariate data with `ggplot2`



Goal: Create animated data visualizations of multivariate data with `gganimate`



Goal: Create interactive data visualizations of multivariate data with `plotly` and `shiny`

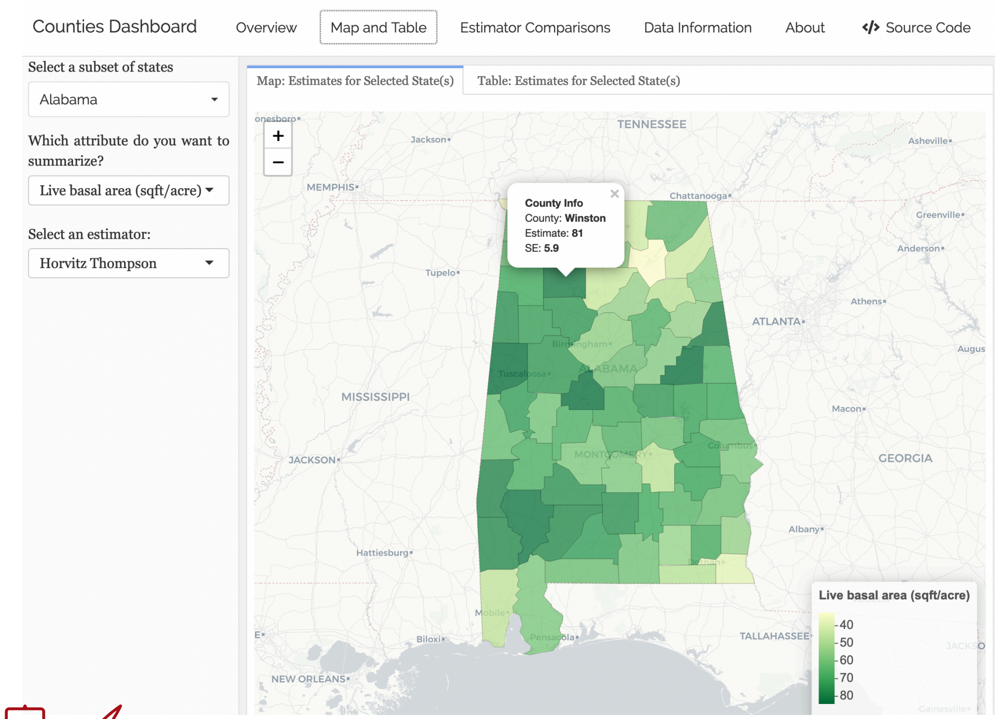


- Also follow **best practices** in data viz, which is Wednesday’s lecture topic!

Learning Goals: Disseminating your work and reproducible workflow

Example: US Forest Inventory and Analysis Program

> Mission: “Make and keep current a comprehensive inventory and analysis of the present and prospective conditions of and requirements for the renewable resources of the forest and rangelands of the US.”



Select a subset of states: Alabama

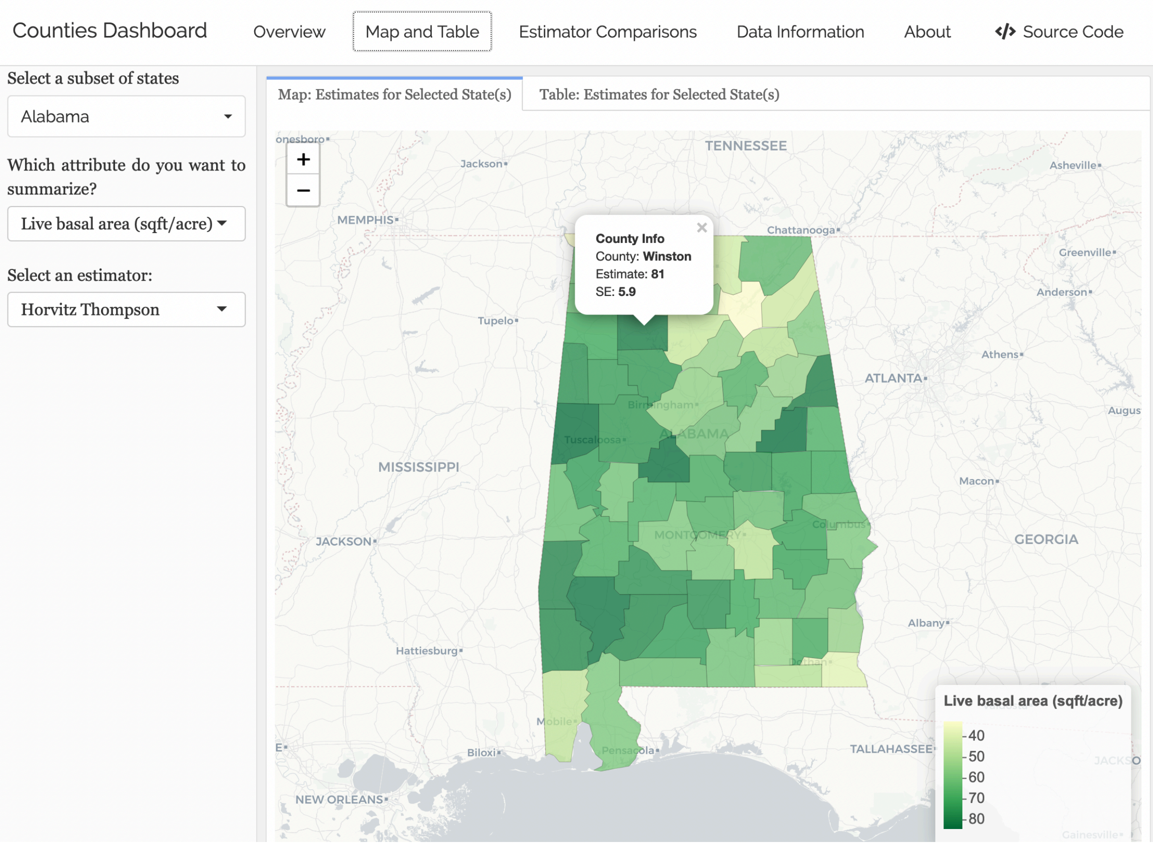
Which attribute do you want to summarize? Live basal area (sqft/acre)

Select an estimator: Horvitz Thompson

County Info
County: **Winston**
Estimate: **81**
SE: **5.9**

- Very common now that any FIA work I do also includes a corresponding dashboard

Goal: Create interactive dashboards with `shinydashboard` and `flexdashboard` for Project 1



Select a subset of states: Alabama

Which attribute do you want to summarize?: Live basal area (sqft/acre)

Select an estimator: Horvitz Thompson

County Info
County: **Winston**
Estimate: **81**
SE: **5.9**

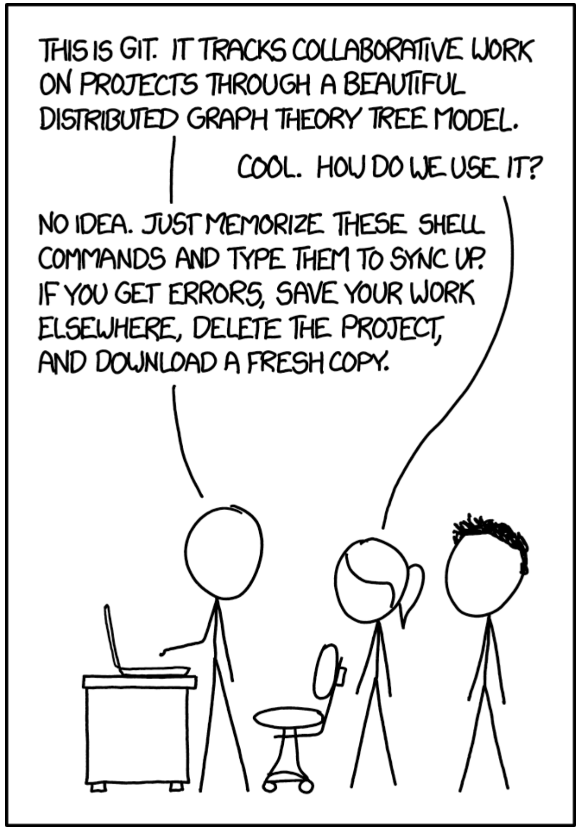
Goal: Learn to create an `R` package with `devtools` for Project 2

I have a few R packages on the Comprehensive R Archival Network:

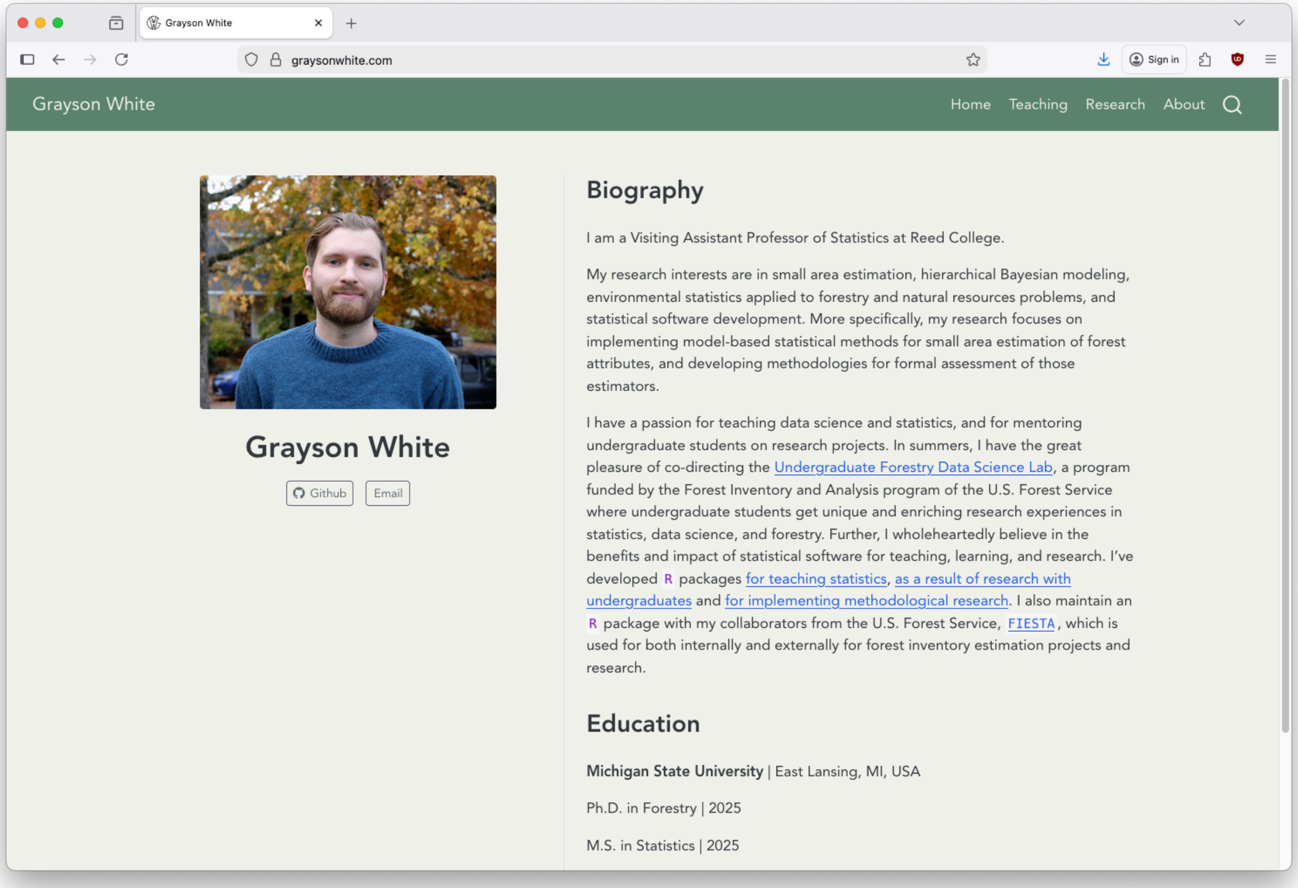
- `gglm`: Produces diagnostic plots for linear models with the `ggplot2` package as its back-end.
- `forested`: A data package that contains US Forest Service data in Washington state.

Goal: Develop a collaborative, version controlled workflow using git and GitHub

- Will use git and GitHub for both of the projects.
- Will also be given a personal GitHub repository for practice and storing course materials.
- GitHub is a great place to share code, R packages, and data work!


THIS IS GIT. IT TRACKS COLLABORATIVE WORK ON PROJECTS THROUGH A BEAUTIFUL DISTRIBUTED GRAPH THEORY TREE MODEL.
COOL. HOW DO WE USE IT?
NO IDEA. JUST MEMORIZE THESE SHELL COMMANDS AND TYPE THEM TO SYNC UP. IF YOU GET ERRORS, SAVE YOUR WORK ELSEWHERE, DELETE THE PROJECT, AND DOWNLOAD A FRESH COPY.

Goal: Learn how to create a website with `quarto`



Grayson White

Github Email

Biography

I am a Visiting Assistant Professor of Statistics at Reed College.

My research interests are in small area estimation, hierarchical Bayesian modeling, environmental statistics applied to forestry and natural resources problems, and statistical software development. More specifically, my research focuses on implementing model-based statistical methods for small area estimation of forest attributes, and developing methodologies for formal assessment of those estimators.

I have a passion for teaching data science and statistics, and for mentoring undergraduate students on research projects. In summers, I have the great pleasure of co-directing the Undergraduate Forestry Data Science Lab, a program funded by the Forest Inventory and Analysis program of the U.S. Forest Service where undergraduate students get unique and enriching research experiences in statistics, data science, and forestry. Further, I wholeheartedly believe in the benefits and impact of statistical software for teaching, learning, and research. I've developed `R` packages for teaching statistics, as a result of research with undergraduates and for implementing methodological research. I also maintain an `R` package with my collaborators from the U.S. Forest Service, `FIESTA`, which is used for both internally and externally for forest inventory estimation projects and research.

Education

Michigan State University | East Lansing, MI, USA

Ph.D. in Forestry | 2025

M.S. in Statistics | 2025

Goal: Utilize a reproducible workflow for all R work!

- Store R work in R scripts and `Quarto` documents.

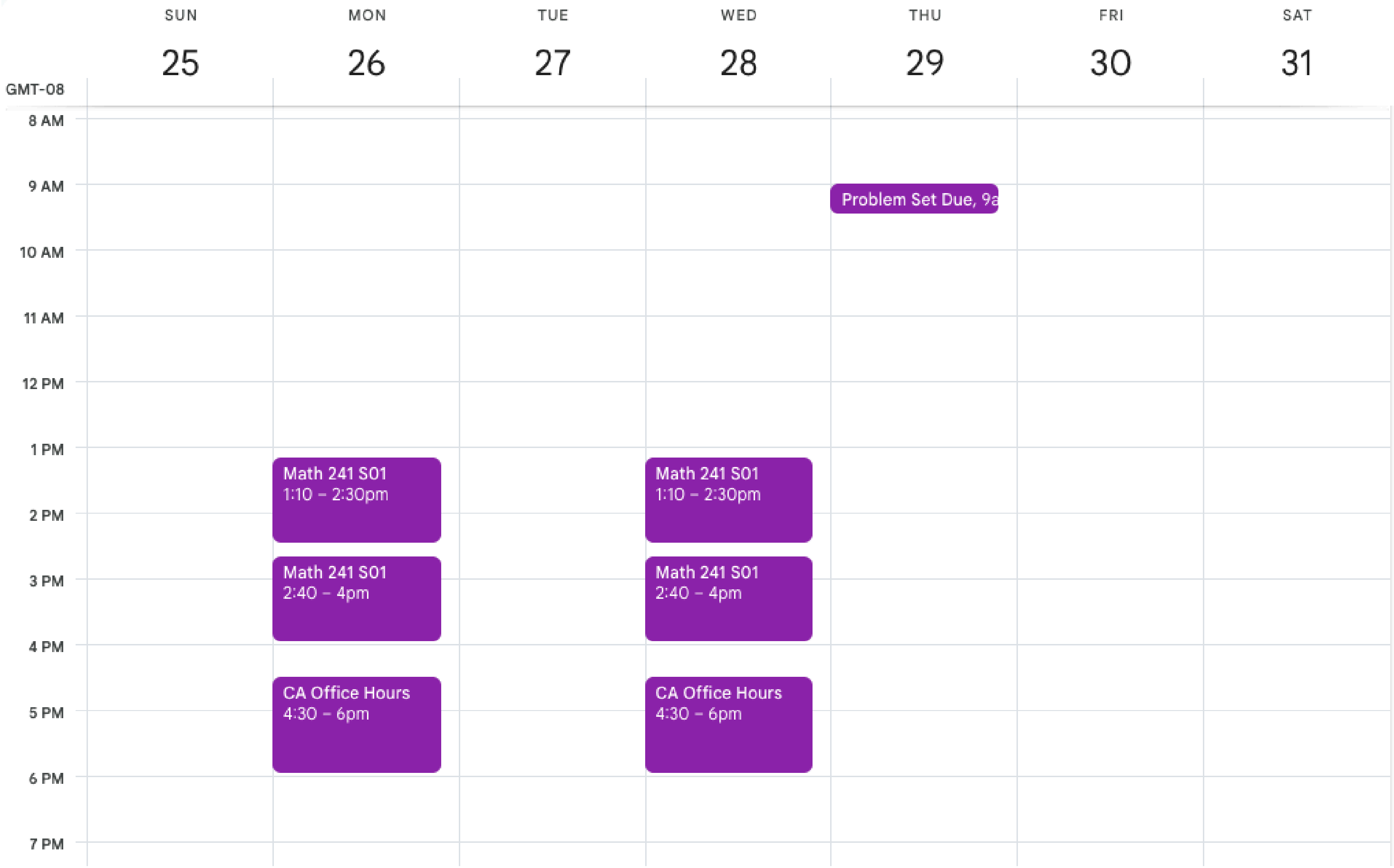
- Create reproducible coding examples so it is easy to ask and answer coding questions.

Now, let's discuss course structure and policies

Course assistant info

- **Name**: Chris Li (they/them)
- **Email**: yiyuanli@reed.edu
- **Office**: ETC 105B
- **Office hours**: Monday and Wednesday, 4:30pm - 6:00pm

A typical week in Math 241



GMT-08	SUN 25	MON 26	TUE 27	WED 28	THU 29	FRI 30	SAT 31
8 AM							
9 AM					Problem Set Due, 9a		
10 AM							
11 AM							
12 PM							
1 PM		Math 241 S01 1:10 – 2:30pm		Math 241 S01 1:10 – 2:30pm			
2 PM		Math 241 S01 2:40 – 4pm		Math 241 S01 2:40 – 4pm			
3 PM							
4 PM		CA Office Hours 4:30 – 6pm		CA Office Hours 4:30 – 6pm			
5 PM							
6 PM							
7 PM							

Assignments and projects

- **Problem sets** (*weekly-ish*)
  - Assigned Thursday by 9am.
  - Due the following Thursday at 9am.
  - Assigned on weeks where you are not actively working on projects.
  - Planning on 7 problem sets, including problem set 0. Subject to change by ± 1 problem set.
- **Projects** (*two of them*)
  - To be completed in groups!
  - Project 1 will be assigned during Week 6 and due during Week 8.
  - Project 1 presentation on the Wednesday before spring break.
  - Project 2 will be assigned during Week 11 and due during Finals Week (i.e. "Week 14").
- **In-class activities** (*semi-regular*)
  - Individual activities.
  - Group activities.
- **Attendance and active participation**
  - Is key to your learning!
  - Please participate in class and in Slack.
  - And will be considered in your final grade.
  - Let me know if you cannot make it to class for health reasons or otherwise! If you are experiencing extenuating circumstances that inhibit your ability to attend I will work with you to help you succeed in the class. Communication is key.

Late Work

- **Problem sets**: up to 4 extensions days can be used throughout the semester.
  - e.g., 1 additional day for 4 labs, 4 additional days for 1 lab, …
  - rounding days up

- **Projects**: no late projects are accepted.

- **In-class assignments**: cannot be made up.

Course Climate

> *We expect everyone in this class to strive to foster a learning environment that is equitable, inclusive, and welcoming. If you experience any barriers to learning, please come to Professor Grayson White or a college administrator with your concerns.*
>
> *Code of Conduct*:
>
> *We expect all members of Math 241 to make participation a harassment-free experience for everyone, regardless of age, body size, visible or invisible disability, ethnicity, sex characteristics, gender identity and expression, level of experience, education, socio-economic status, nationality, personal appearance, race, religion, or sexual identity and orientation.*
>
> *We expect everyone to act and interact in ways that contribute to an open, welcoming, inclusive, and healthy community of learners. You can contribute to a positive learning environment by demonstrating empathy and kindness, being respectful of differing viewpoints and experiences, and giving and gracefully accepting constructive feedback.*[1]

01:00

[1] This Code of Conduct is adapted from the Contributor Covenant, version 2.0.

Engagement

- During lecture and lab, remove distractions.
  - When we are on our computers, close email, social media, news, etc.
  - Hide your phone.
- I have high expectations but know that all of you (regardless of your stats, math, or computing background) have the ability to meet them.

- Being **actively present** is key.

Artificial Intelligence (AI) Policy

> *Artificial intelligence (AI) tools, such as ChatGPT, Claude, Co-Pilot, Gemini, and others are being used to generate code, analyze data, write, and much more (and they are getting quite good at many of these tasks!). On the other hand, learning to think critically about a problem at hand, and engaging with your peers, tutors, and instructors when not understanding a concept or question are integral components of a liberal arts education. This dichotomy puts Math 241 in a interesting position as a course which aims to provide a cutting edge data science education at a liberal arts college. One of my main goals as the instructor of this course is to turn each of you into high quality data scientists who have a deep understanding of the R programming language and the ability to communicate insights about data on your own. Because of this, we will take a nuanced approach to engaging with AI tools in this course.*

- Well, what's the policy??

01:00

Artificial Intelligence (AI) Policy

- **For all content delivered or assigned before Spring Break**: The use of generative AI tools, such as ChatGPT and others, are strictly prohibited in any stage of the work process for this course.
- **After we return from Spring Break**: Along with other material, I will teach about how to effectively engage with AI tools as a data scientist. We will learn how to effectively debug and write code with AI tools. Assignments after Spring Break will have clear instructions regarding the ways that you are (and are not) allowed to engage with AI on them. Note that you may **never** use AI tools in the writing process for this course.
- Always, if you have questions about whether a tool is allowed for this course or if the way you plan to engage with a tool is allowed, ask the instructor before using it.

Why this policy?

On one hand…

- In today’s job market, understanding generative AI tools and being able to engage with them is a key skill for data scientists and software engineers.

On the other hand…

- Relying too heavily on AI tools to write your code and think for you prohibits understanding of the core concepts that make you a good programmer and data scientist.

My goal: Make you a strong programmer with a substantial base of understanding early in the course so that you can use that base to…

- independently engage with data and do data science,
- utilize and critique generative AI output, and
- continue to embrace and engage in the very human aspects of data science.

Questions?

About the syllabus, AI, the course, etc.?

Learning to be a Coder

Many of the assignments will provide opportunities to stretch yourself and learn code **not** covered directly in class. Why?

(Another) Goal: Develop our abilities to effectively search for and evaluate potential solutions and then adapt the code to our situation.

RStudio Community
All things RStudio

stackoverflow

Potential Erroneous Side Effect: Concluding that you are bad at coding because you can't solve the problem "on your own" or because you find the answers on StackOverflow confusing/unhelpful.

Why aren't I allowing you to use AI for this? (before spring break): AI is generally much too quick to provide an answer and takes much of the learning out of the process. Also, it is often the case that generative AI will produce a much too complex solution to a problem that can be solved elegantly. Once you develop your strong coding base, you'll be able to use AI and critique its output.

(Another) Goal: Develop our abilities to effectively search for and evaluate potential solutions and then adapt the code to our situation

I encourage employing the following strategy for solving coding questions:

→ Try the problem.

→ If and when you get stuck, look to the internet for potential solutions.

→ If you find some promising ideas, try them out.

→ For the code that seems most helpful, spend time figuring out what each line does.

→ If the code doesn't exactly solve your problem, adapt it. Even if it does, still consider whether or not modifications should be made.

→ If still stuck, post your question on our class Slack and/or come to office hours. **Don't just stay stuck.**

Key: Try to find the right balance between independent learning and supported learning.

- And, get help **before** the frustration sets in!

Next time

We'll start **decomposing** graphics and discussing **best practices** for creating graphics!