Week 1: Statistical Reasoning and Data Science

Jupyterhub, Rstudio/Rmd, R, and the Tidyverse

Scott Schwartz

July 21, 2021

Statistical Reasoning and Data Science (DS)

Modern Stats+DS is

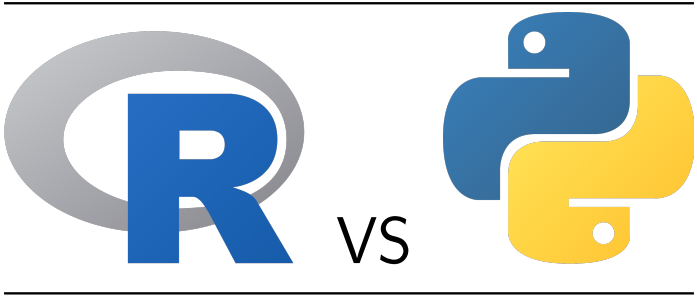
1. software, programming, and computational tools implementing
2. mathematical and algorithmic data analysis methodologies
3. explained and advocated with written and verbal communication
4. to facilitate data-driven and evidence-based decision making

1. software, programming, and computational tools

Modern Stats+DS is based on using computational tools

- Programming IS required for DS and modern applied stats
- It's not an STA130 prerequisite but wanting to learn to code is

Pirates versus Snakes



Modern Stats+DS learning communities propegate knowledge and solutions

- R is ubiquitous in traditional data analysis and statistical communities
  - and increasingly supports machine learning (ML), bioinformatics, etc.
- Python is dominant in nearly all application involving coding, such as DS and ML
  - and increasingly overlaps and supports users traditionally from the R community

Class Check Round 1

https://pollev.com/sta (3 questions)

Community / Mentorship Program!

Statistical Reasoning and Data Science (DS)

Modern Stats+DS is

1. software, programming, and computational tools implementing
2. mathematical and algorithmic data analysis methodologies
3. explained and advocated with written and verbal communication
4. to facilitate data-driven and evidence-based decision making

2. math, algorithms, and data analysis Part I: Structured Learning

The course (on average across students) should take **10 hours a week**

Section	ET	Monday	Thursday	Friday
L0101	9:10 AM	2hr Lec		2hr Tut
L0201	2:10 PM	2hr Lec		2hr Tut
L0101+L0201	5:00 PM		R 2hr HW Due	
L0101+L0201	10:00 PM			+1hr Tut work Due

with 7 hours allocated as indicated above and 3 hours left available for study and review, office hours and piazza discussion boards, and eventual team project work.

Boost your skills with free textbook resources like R for Data Science by Wickham and Grolemund, free learning tools like the DoSS Toolkit from STA130 profs Alexander and Caetano et al., and free online primers, cheatsheets, and courses

Class Check Round 2

https://pollev.com/sta (3 questions)

2. math, algorithms, and data analysis Part II: Unstructured Learning

When first learning, structured course material is good

Eventually it's faster to learn and troubleshoot problems yourself

2. math, algorithms, and data analysis Part II: Structured Learning → Unstructured Learning

The Troubleshooting / Coding Resilience workflow

1. Have I reviewed the course notes well enough?
2. Is this the kind of problem a google search could fix?
3. Could this have already been answered on the course piazza?
4. Am I SURE this hasn't already been answered on the course piazza?
5. Would discussing my problem in an office hour be more helpful than piazza?
6. Would discussing my problem with a peer be more helpful than piazza?
7. Is asking on the course piazza probably the fastest way to get an answer?

Class Check Round 3

https://pollev.com/sta (3 questions)

Statistical Reasoning and Data Science (DS)

Modern Stats+DS is

1. software, programming, and computational tools implementing
2. mathematical and algorithmic data analysis methodologies
3. explained and advocated w/ written and verbal communication
4. to facilitate data-driven and evidence-based decision making

3. written and verbal communication: course focus

2hr Lec	Monday 9:10AM ET and 2:10PM ET	N/A
Coures Project	Concepts, Communication, Coding	**20%**
3hr Tut HW	Friday 9:10PM ET and 2:10PM ET	**14%**
2hr R HW	Due Thursday 5pm	**7%**
Exams	Midterm 20% and Final 34%	**54%**
Study/Review	Slides, Demos, Questions	N/A
Participation	Mentorship and Surveys	**5%**

Boost your skills with free textbook resources like R for Data Science by Wickham and Grolemund, free learning tools like the DoSS Toolkit from STA130 profs Alexander and Caetano et al., and free online primers, cheatsheets, and courses

3. written and verbal communication: course rules

2hr Lec	not mandatory, but also not recorded	N/A
Coures Project	Due Dec 8th: Emergency Absences ONLY	20%
3hr Tut HW	Best 7/9 Policy + College Registrar	14%
2hr R HW	Best 7/9 Policy + College Registrar	7%
Exams	Midterm Reweight + FAS Defer Petition	54%
Study/Review	not mandatory, but good exam practice	N/A
Participation	Due Dec 8th: No Exceptions	5%

For more details on the **Best 7/9 Policy**, **Midterm reweighting**, etc., please review the grading policies and accommodation request protocols. *E.g., "7/9 policy" extensions must be requested within one week of return to UofT.*

Class Check Round 4

https://pollev.com/sta (4 questions)

Statistical Reasoning and Data Science (DS)

Modern Stats+DS is

1. software, programming, and computational tools implementing
2. mathematical and algorithmic data analysis methodologies
3. explained and advocated with written and verbal communication
4. to facilitate data-driven and evidence-based decision making

4. facilitating data-driven evidence-based work: Jupyterhub

Jupyterhub is a cloud-based service allowing you to run R/Rstudio from any web browser

- You can run R/Rstudio locally as well, and you may need to install Rstudio so you can do this for this class in case of temporary Jupyterhub service outage

From Jupyerhub in a gif-y

Logging in reconnecting

UNIVERSITY OF TORONTO JUPYTERHUB

4. facilitating data-driven evidence-based work: Rstudio

Rstudio is just a nice GUI IDE program that wraps up R along with a .Rmd code editor and file and data management capabilities

- GUI means "graphical user interface" while IDE means "integrated developers environment", and .Rmd RMarkdown files (based on Markdown markup language) combine text with R code to create .html and .pdf output files and presentation slides (like the ones I'm using here now).

4. facilitating data-driven evidence-based work: Knitting

Rstudio knitting as illustrated by @allison_horst

Knitting

Knitting

Knitting

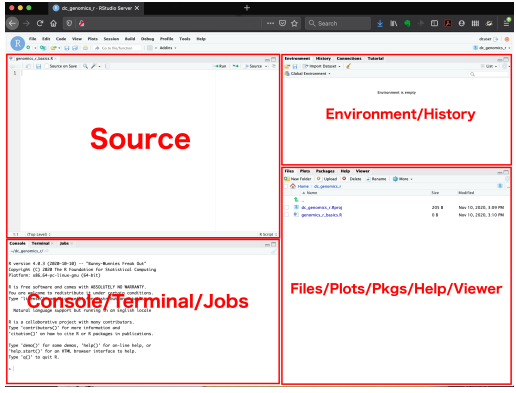
Knitting

Knitting

Knitting

4. facilitating data-driven evidence-based work: Rstudio + Code Chunks

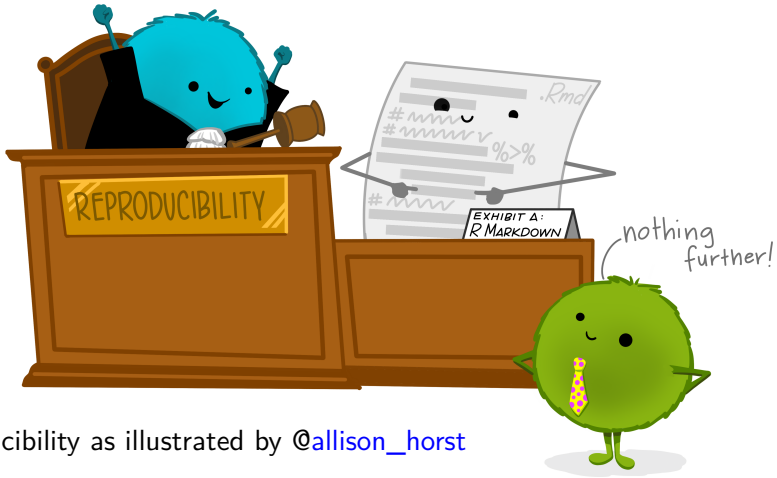
1. ***Source*** [Top Left pane]: Rmarkdown code+text files
2. ***Console*** [Bottom Left pane]: Code+Knitting history
3. ***Environment*** [Top Right pane]: Variables and data sets
4. ***Files*** [Bottom Right pane]: Figure and knitting outputs



To run code-chunks use (PC) or (Mac)

- This is just one of many keyboard (key-binding) shortcuts in Rstudio

4. facilitating data-driven evidence-based work: Reproducibility

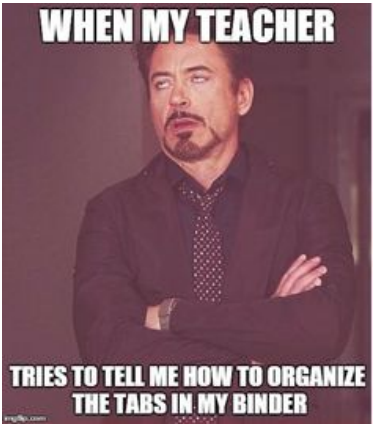


Rmd reproducibility as illustrated by @allison_horst

4. facilitating data-driven evidence-based work: File Management

From Jupyerhub in a gif-y

folders relocations uploading linking projects files exporting



4. facilitating data-driven evidence-based work: R

R is a data analysis programming console to support the work of statisticians, data scientists, and other data-centric professions

- R is a programming language, but its methods and algorithms are usually built-in or loaded from packages, so most R users don't build algorithms or data types like in C++ or java

R as Illustrated by @allison_horst

4. facilitating data-driven evidence-based work: R Packages

Newly developed statistical methods are often made available as R package

This makes the publication manuscripts stronger and increases a methods use

Comprehensive R Archive Network (CRAN) currently has 18403 packages

For those new to R and arriving with interest in DS, the tidyverse is the key set of packages

The tidyverse includes R packages that help facilitate modern stats+DS

4. facilitating data-driven evidence-based work: Common Errors with Packages and Libraries

Installing Packages, Loading Libraries, and Common Errors

Installing Packages, Loading Libraries, and Common Errors

- You can't run R code unless its package is loaded as `library(packagename)`
- You can't load packages unless `install.packages("packagename")` is run first
- Packages installed on Jupyterhub must be re-install for each new session...
- Some pre-installed packages, e.g., `library(tidyverse)`, will always load
- `There is no package called` errors mean `install.packages()` is needed
- But you can't "knit" .Rmd files `install.packages("packagename")` code
  - So you have to frequently comment and uncomment `install.packages()`

Class Check Round 5

https://pollev.com/sta (5 questions)

And some Self Quiz questions you should be asking yourself...

Can you log into JupyterHub and open and knit an RMarkdown document in RStudio?
Can you navigate the Code, R Console, and Files panes in the RStudio interface?
Did you see the Environment pane and the Help, Packages, and Tutorial tabs?

Statistical Reasoning and Data Science (DS)

Modern Stats+DS is

1. software, programming, and computational tools implementing
2. mathematical and algorithmic data analysis methodologies
3. explained and advocated with written and verbal communication
4. to facilitate data-driven and evidence-based decision making

Statistical Reasoning and Data Science (DS)
Course Learning Objectives

1. Implement the computational steps involved in the management and statistical analysis of data using R.
2. Carry out a variety of statistical analyses in R and interpret the results of the analyses.
3. Clearly communicate the results of statistical analyses to technical and non-technical audiences.
4. Identify appropriate uses of statistical methods to answer questions, including their strengths and weaknesses.
5. Describe how statistical methods can be used to learn from data, including methods for description, explanation, and prediction.


R MARKDOWN
we're getting the BAND BACK Together.
TEXT
OUTPUTS!
CODE
HORST '19

Rstudio Demo

1. Click this jupyterhub repo launcher link
   - If this fails (1) log into RStudio on Jupyterhub, (2) delete the folder, (3) retry
2. Navigate to the intended directory in the (bottom right) files pane
3. Open the intended .Rmd file
4. Follow along / complete the demo
5. Understand Rmd+pdf submission requirements
6. Determine and commit to a file organization scheme