

# Introduction to Data Science

Stat 220

March 27 2023

## Something about me

- Second year at Carleton
- Originally from Nepal
- Research in Bayesian computation and machine learning
- Avid learner and traveler



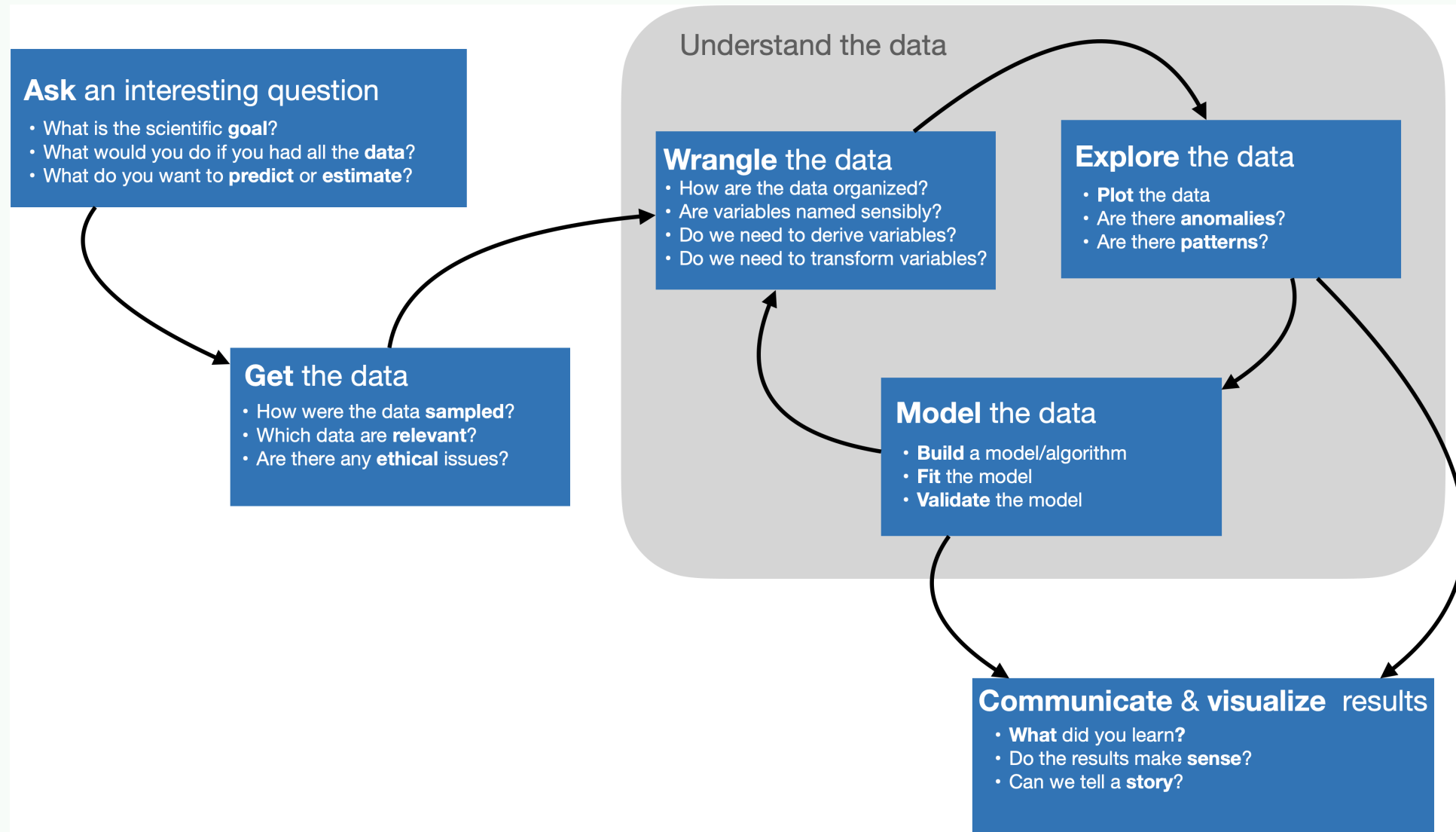
On top of Nordkette in Innsbruck, Austria

## What is data science?

*Data science is the application of **computational** and **statistical** techniques to gain insight into some problem in the real world*

Data Science = scientific inquiry +  
data collection +  
data processing +  
visualization +  
statistics +  
machine learning +  
communication

# Data Science in a nutshell



## Stat 220: Introduction to Data Science

Focus on the “soup to nuts” approach to problem solving

- *data wrangling*
  - *reshaping, cleaning, gathering*
- *learning from data*
  - *EDA tools*
  - *statistical learning methods*
- *communication*
  - *reproducibility*
  - *effective visualization*

# Gendered language in professor reviews

## Gendered Language in Teacher Reviews

I've had [trouble keeping this site up continuously](#) during COVID. As of March 2021, I'm now trying a [new strategy](#) to cache common queries on the server even when the underlying database is down. If you find that many searches don't change the results, that's why.

This interactive chart lets you explore the words used to describe male and female teachers in about 14 million reviews from RateMyProfessor.com.

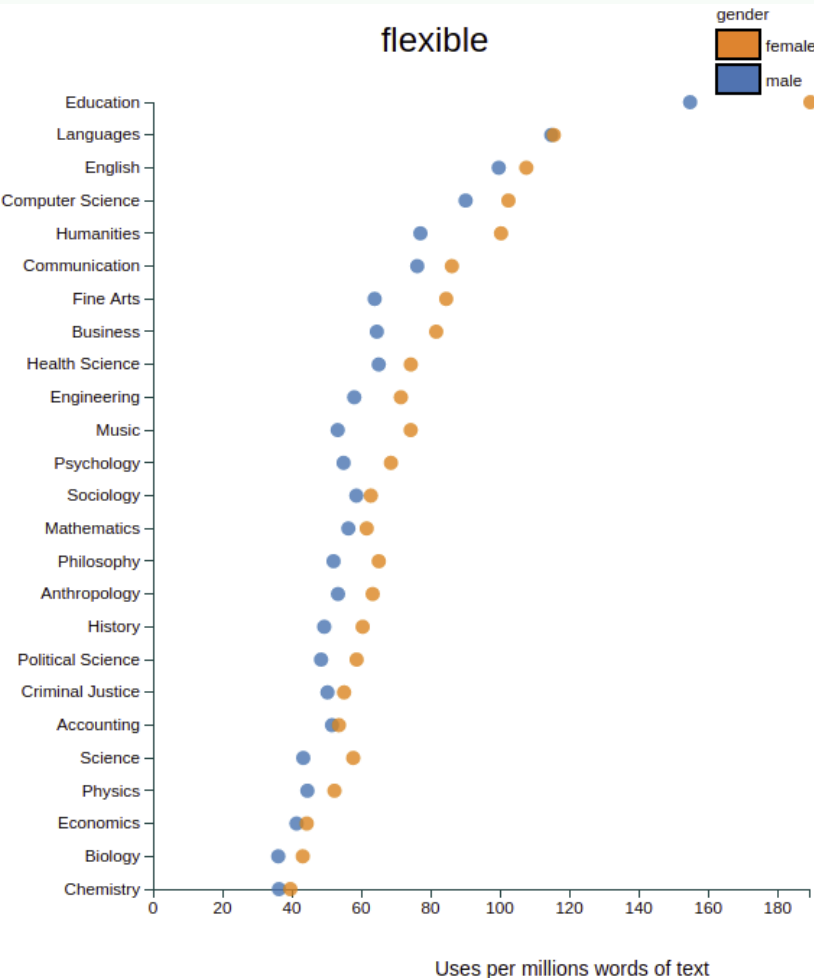
Not all words have gender splits, but a surprising number do. Even things like pronouns are used quite differently by gender.

**Search term(s) (case-insensitive):**  
**use commas to aggregate multiple terms**

flexible

All ratings Only positive Only negative

You can enter any other word (or two-word phrase) into the box above to see how it is split across gender and discipline: the x-axis gives how many times your term is used per million words of text (normalized against gender and field). Or limit to just negative or positive reviews (based on the numeric ratings on the site). For some more background, see [here](#).



## How to make friends and succeed in Data Science?

1. Actively follow along! RMarkdown (.Rmd) documents will be provided for you each class
  - use these to take notes and run code “live” in class.
2. Ask questions!
  - This is new for everyone, no question is a bad question.
3. When you don't know if something will work, try it!
  - experimentation is key in this class.

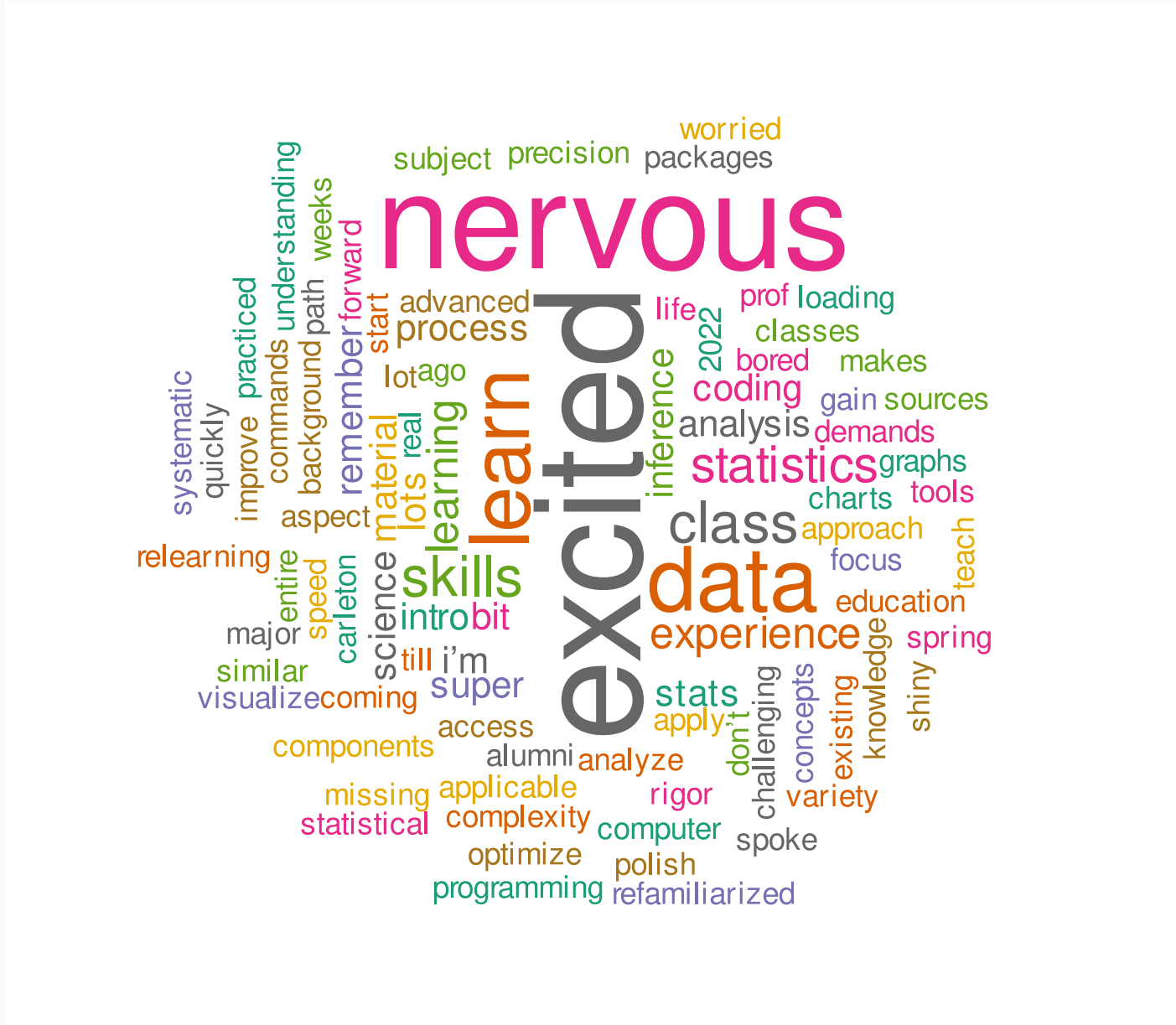
**Tell me something about yourself!**

- **Your name?**
- **Gender Pronouns?**
- **Why are you interested in data science?**
- **Your favorite media personality?**





# Survey Responses: One thing you are excited and nervous about?



## Class Pipelines

<https://stat220-spring23.netlify.app>

- *Please bookmark this page: should be checked multiple times a day*
- *Most of the course information and schedule will be posted in [moodle](#)*
- *Use moodle for submitting class activity and seeing grades*
- *Use Github for homework and projects*

## Necessary skills to be mastered

- programming with data
- statistical modeling
- domain knowledge
- communication

## What will a typical day/week look like?

### *Before class:*

- *Some reading/video to introduce some topics*
- *Work on homework/projects, come with questions*

### *During class:*

- *Mini lectures*
- *Class activities*

What you need to do next . ...

- read the [Rstudio for Stat220](#) page
- read the [GitHub for Stat220](#) page
- read the [Software for Stat220](#) page

## R Vs Python for data science

"R is written by statisticians, for statisticians,"

Norm Matloff, Author of The Art of R Programming, Prof. of Computer Science, UC Davis

### Advantages of R over Python:

- Not so steep learning curve as Python
- R has many generic functions that are universal, e.g. `print()`, `plot()`, `summary()`
- R Comprehensive R Archive Network (CRAN) has many user-friendly packages
- R's basic `help()` and `example()` functions are much more informative than Python's counterparts

### Script file

Write code here

To run code put your cursor on the line and click the run button

Edit to correct errors

⇒ record of commands that worked

Save scripts with the .R extension

⇒ syntax will be highlighted

⇒ good practice

<- is the assignment operator

⇒ puts what is on the right in to the object on the left

⇒ Assign results if you want to use them again

### Console

When you click run, code is sent to the console and executed

> is the prompt

⇒ do not type it

⇒ appears when R is ready for next command

Command output goes here by default

⇒ output is in a different colour

⇒ [1] indicates 3.4 is the first element of the output

⇒ many commands will not have output, the prompt just reappears

Script: where you write code

Console: where output goes

### Environment

Name objects by assignment to use them again

All the objects you created in your session

Saving the environment saves all the objects, but not the code with a .RData extension

### History

A history of every command you sent to the console, mistakes included.

File can be saved but usually you just need the script

Environment: where saved output goes

### Packages

Many functions come with R

A huge amount of extra functionality is available in packages

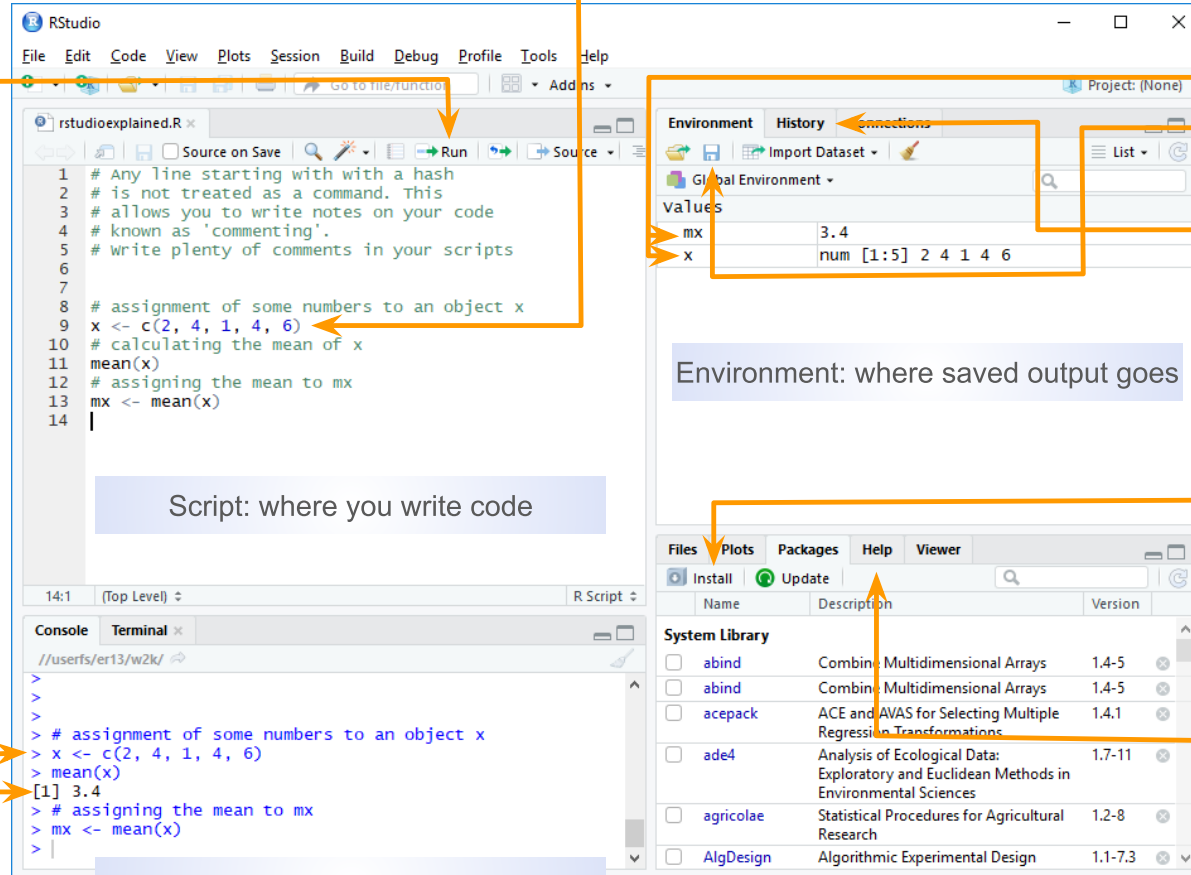
Packages can be installed by clicking the Install button

### Help

Access to manual pages for all installed packages

### Plots

Figure output appears here



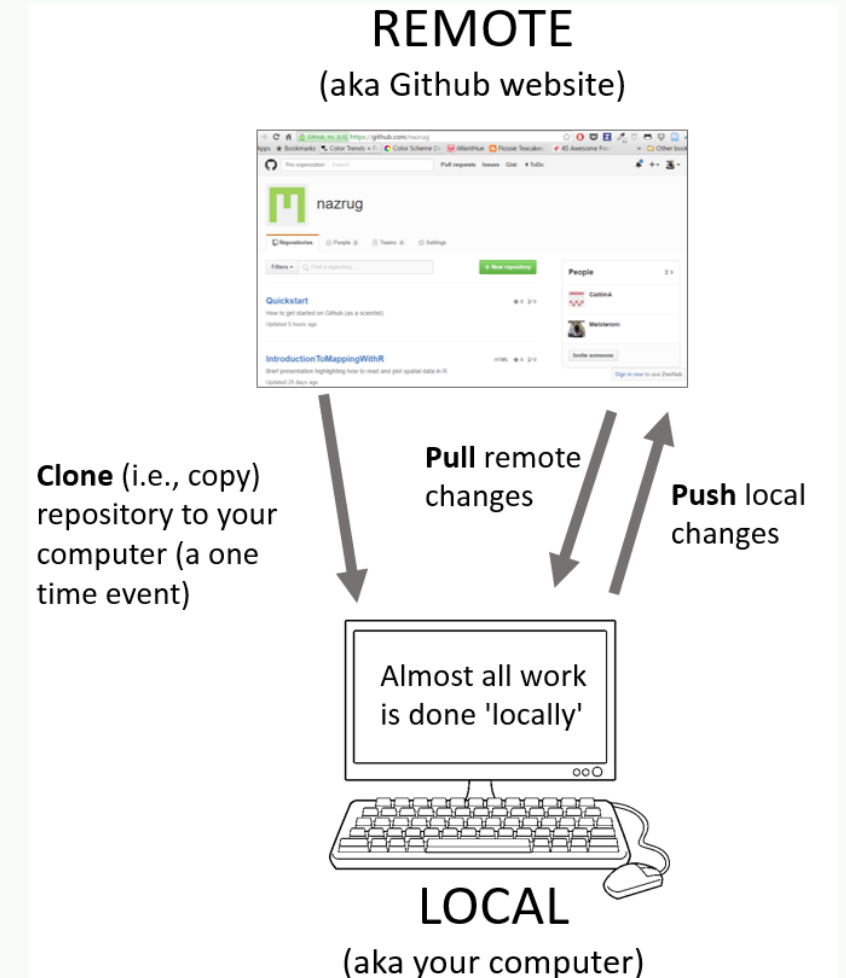
## Using R Markdown for data science

- You will use R Markdown for all work in this class
- A Markdown (.Rmd) file contains
  - R code
  - answers, description of results, report, etc.
- The Markdown file is knit to generate an output document
  - pdf, html, word
  - presentations (html, beamer pdf)
  - dashboards, interactive graphics (html)
- Markdown is designed for reproducibility!
- The slides I produce for this class are made using R Markdown's [Xaringan](#) presentation



## Version Control using Git and GitHub

- **User:** A Github account for you (e.g., deepbas).
- **Organization:** The Github account for one or more user (e.g., DataScienceSpring23).
- **Repository:** A folder within the organization that includes files dedicated to a project.
- **Local Github:** Copies of Github files located your computer.
- **Remote Github:** Github files located on the <https://github.com> website.
- **Clone:** Process of making a local copy of a remote Github repository.
- **Pull:** Copy changes on the remote Github repository to your local Github repository.
- **Push:** Save local changes to remote Github



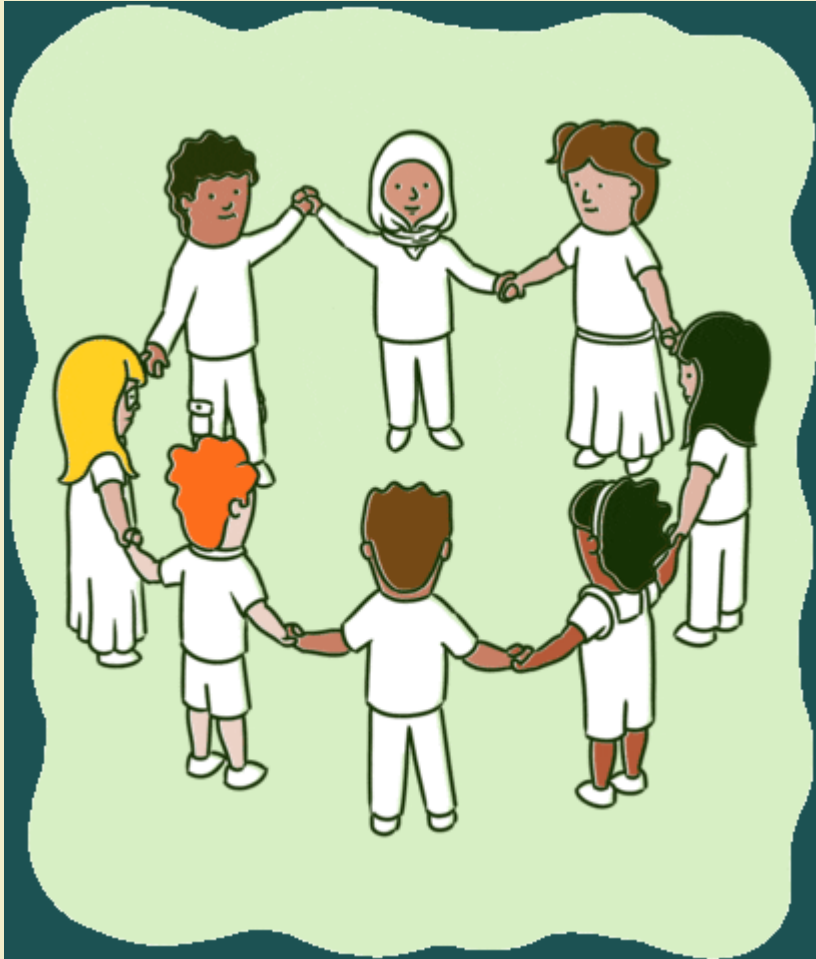
Git Cycle

## Using GitHub and Rstudio for data science

Rstudio lets you create git controlled projects

- create a GitHub repo
- make a Rstudio project using your cloned repo
- edit/create files (.rmd, .r, .csv, . . . )
- commit changes to your local computer using git
- push changes to the GitHub repo (online)
- pull changes made by others to your computer

# Group Activity 1



- Work on the `class_activity_1.Rmd` file
- You can get the file from [moodle](#)
- Talk to your neighbor
- Submit your compiled `.pdf` file to moodle