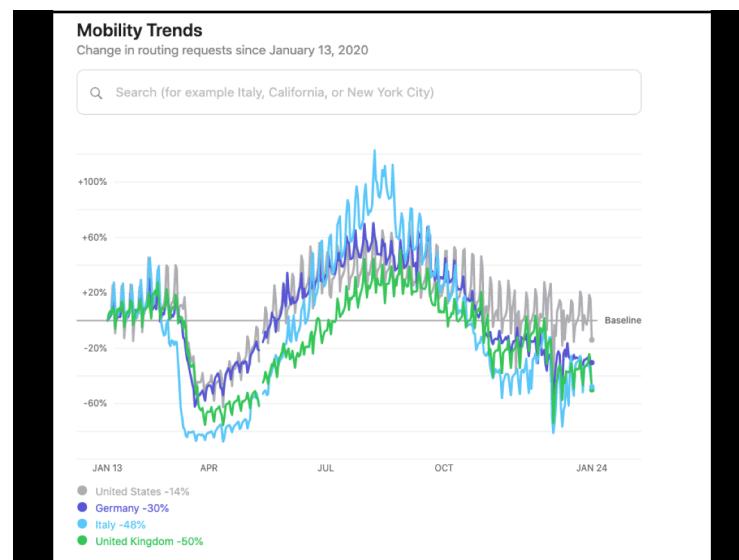


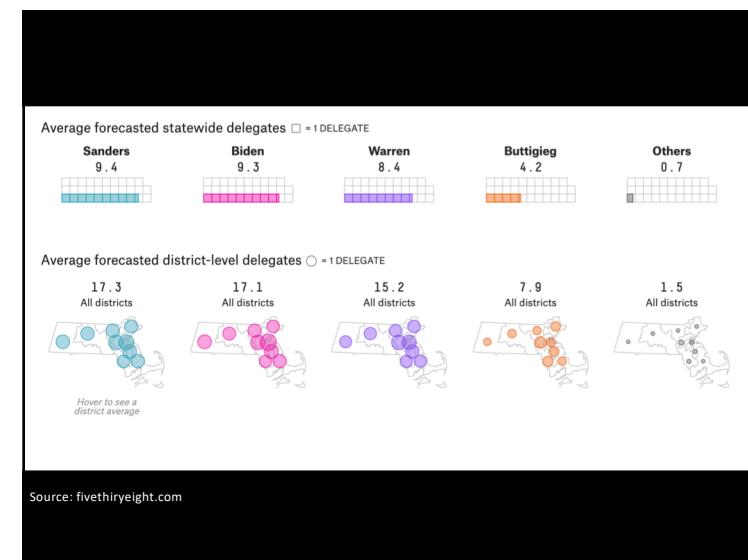
# Introduction to Data Science for Biology

UMass Boston

Spring 2021



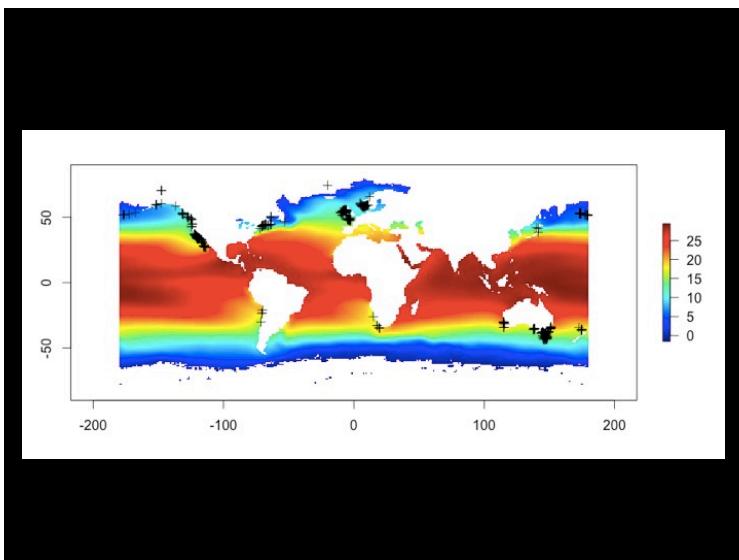
## We Are Awash in Data





SalemSound\_Swath\_DATA\_2013.xlsx

YEAR	MONTH	DAY	DATE	SITE	TRANSECT	SP_CODE	0-2 IN	20-40 IN	40-20 OFF	20-0 OFF
2	2013	7	24	7/24/2013 BAKER	1	HOAM	1	1	2	1 LY
3	2013	7	24	7/24/2013 BAKER	1	CABO	1	1	1	0 LY
4	2013	7	24	7/24/2013 BAKER	2	ASFO	0	6	7	1 LY
5	2013	7	24	7/24/2013 BAKER	2	HOAM	1	3	2	2 LY
6	2013	7	24	7/24/2013 BAKER	2	CABO	2	8	11	2 LY
7	2013	7	24	7/24/2013 BAKER	2	CABO	0	1	2	0 LY
8	2013	7	24	7/24/2013 BAKER	2	CABO	0	1	2	0 LY
9	2013	7	24	7/24/2013 BAKER	2	CAMA	1	9	5	2 LY
10	2013	7	24	7/24/2013 BAKER	2	ASFO	0	4	6	0 LY
11	2013	7	24	7/24/2013 BAKER	2	ASRU	0	0	1	0 LY
12	2013	7	30	7/30/2013 BAKER	3	HOAM	13	3	6	2 LY
13	2013	7	30	7/30/2013 BAKER	3	CABO	3	3	0	3 LY
14	2013	7	30	7/30/2013 BAKER	3	CABO	8	1	0	2 LY
15	2013	7	30	7/30/2013 BAKER	3	HESA	1	0	1	0 LY
16	2013	7	30	7/30/2013 BAKER	3	HOAM	3	1	4	2 BY
17	2013	7	30	7/30/2013 BAKER	4	CAIR	1	1	1	1 BY
18	2013	7	30	7/30/2013 BAKER	4	CABO	1	0	2	1 BY
19	2013	7	30	7/30/2013 BAKER	4	SIDE	0	0	0	0 BY
20	2013	7	24	7/24/2013 BAKER	4	HESA	0	0	0	2 BY
21	2013	7	24	7/24/2013 BAKER	4	PAGURUS	0	0	1	0 BY
22	2013	8	20	8/20/2013 BAKER	5	HOAM	1	1	2	3 ALL
23	2013	8	20	8/20/2013 BAKER	5	CAIR	0	1	1	0 ALL
24	2013	8	20	8/20/2013 BAKER	5	CABO	1	1	3	0 ALL
25	2013	8	20	8/20/2013 BAKER	5	ASFO	0	2	0	0 ALL
26	2013	8	20	8/20/2013 BAKER	5	CAMA	1	0	0	0 ALL
27	2013	8	20	8/20/2013 BAKER	6	HOAM	1	3	0	4 ALL
28	2013	8	20	8/20/2013 BAKER	6	CAIR	1	1	2	2 ALL
29	2013	8	20	8/20/2013 BAKER	6	SIDE	2	2	5	2 ALL
30	2013	8	20	8/20/2013 BAKER	6	CAMA	1	0	0	0 ALL
31	2013	8	20	8/20/2013 BAKER	6	CAIR	3	1	4	1 ALL



SYNTENIC ASSEMBLIES FOR CG15386

```

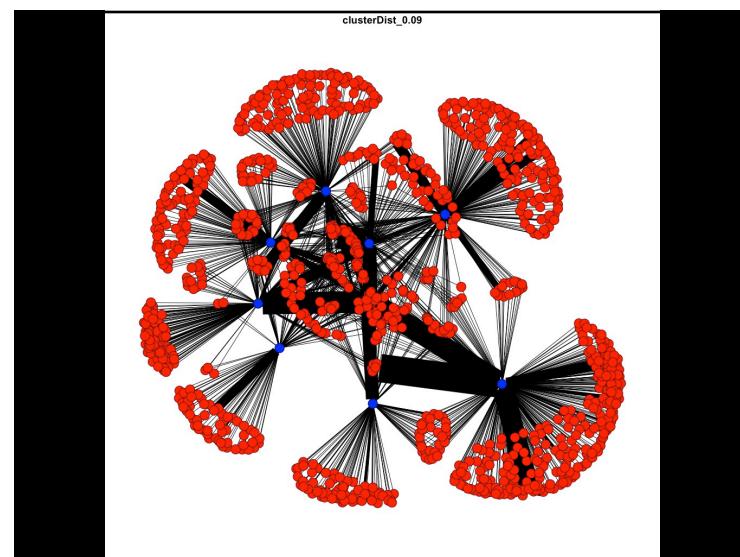
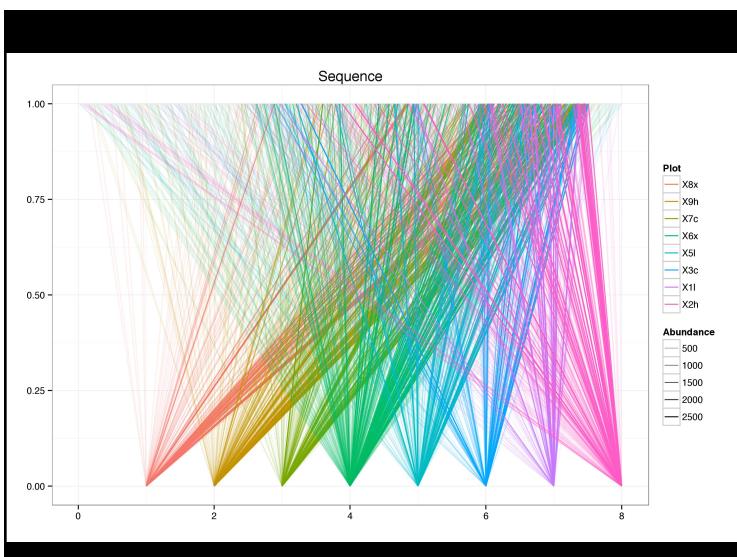
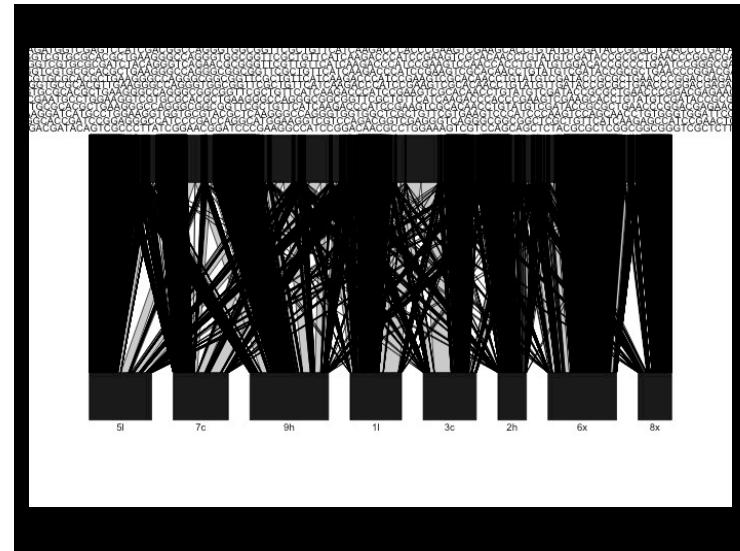
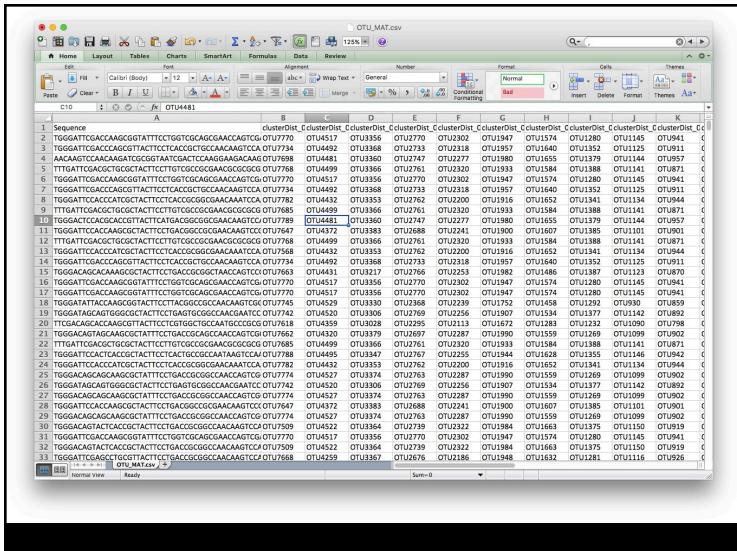
MD106 ATGCTTAGTAACTCCCTACTTTAGTCGCTTGTGGCTATGGCTTCGGAGGAATGGC
NEWC CCGTTTCAAGTACCAAACGTAGTCGCGATGAGCAGCGAAAGCTCTGTTTATGAGAAG
W501 CTACGGCCCTAAATGGTCGCTAACCGAGCCGAACCGTGCACAAAATAGGCCGATCTAAAGCCT
MD199 ATGCTTAGTAACTCCCTACTTTAGTCGCTTGTGGCTATGGCTTCGGAGGAATGGC
C1674 ATGCTTAGTAACTCCCTACTTTAGTCGCTTGTGGCTATGGCTTCGGAGGAATGGC
SIM4 ATGCTTAGTAACTCCCTACTTTAGTCGCTTGTGGCTATGGCTTCGGAGGAATGGC

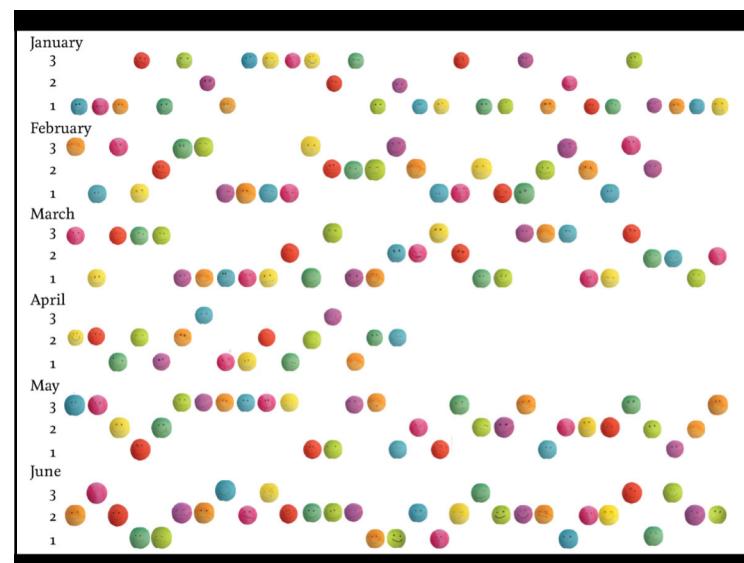
MD106 CTACGGCCCTAAATGGTCGCTAACCGAGCCGAACCGTGCACAAAATAGGCCGATCTAAAGCCT
NEWC CCGTTTCAAGTACCAAACGTAGTCGCGATGACCGAGCGAAAGCTCTGTTTATGAGAAG
W501 CTACGGCCCTAAATGGTCGCTAACCGAGCCGAACCGTGCACAAAATAGGCCGATCTAAAGCCT
MD199 CCGTTTCAAGTACCAAACGTAGTCGCGATGACCGAGCGAAAGCTCTGTTTATGAGAAG
C1674 CCGTTTCAAGTACCAAACGTAGTCGCGATGACCGAGCGAAAGCTCTGTTTATGAGAAG
SIM4 CCGTTTCAAGTACCAAACGTAGTCGCGATGACCGAGCGAAAGCTCTGTTTATGAGAAG

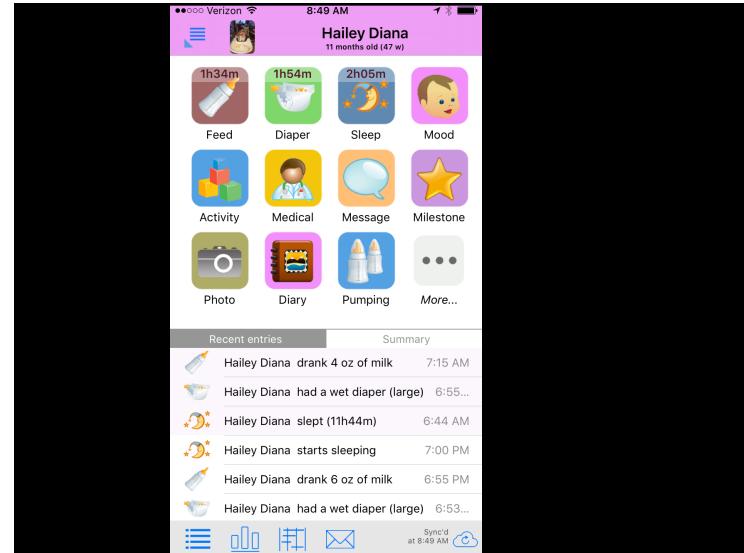
MD106 CCGTTTCAAGTACCAAACGTAGTCGCGATGACCGAGCGAAAGCTCTGTTTATGAGAAG
NEWC CCGTTTCAAGTACCAAACGTAGTCGCGATGACCGAGCGAAAGCTCTGTTTATGAGAAG
W501 CCGTTTCAAGTACCAAACGTAGTCGCGATGACCGAGCGAAAGCTCTGTTTATGAGAAG
MD199 CCGTTTCAAGTACCAAACGTAGTCGCGATGACCGAGCGAAAGCTCTGTTTATGAGAAG
C1674 CCGTTTCAAGTACCAAACGTAGTCGCGATGACCGAGCGAAAGCTCTGTTTATGAGAAG
SIM4 CCGTTTCAAGTACCAAACGTAGTCGCGATGACCGAGCGAAAGCTCTGTTTATGAGAAG

MD106 CTGCGAGGAGGGCTGACGACCAATGCCGGAATCTACAGTCGCGGGCGAGGAATAG
NEWC CTGCGAGGAGGGCTGACGACCAATCTACAGTCGCGGGCGAGGAATAG
W501 CTGCGAGGAGGGCTGACGACCAATCTACAGTCGCGGGCGAGGAATAG
MD199 CTGCGAGGAGGGCTGACGACCAATCTACAGTCGCGGGCGAGGAATAG
C1674 CTGCGAGGAGGGCTGACGACCAATCTACAGTCGCGGGCGAGGAATAG
SIM4 CTGCGAGGAGGGCTGACGACCAATCTACAGTCGCGGGCGAGGAATAG

```







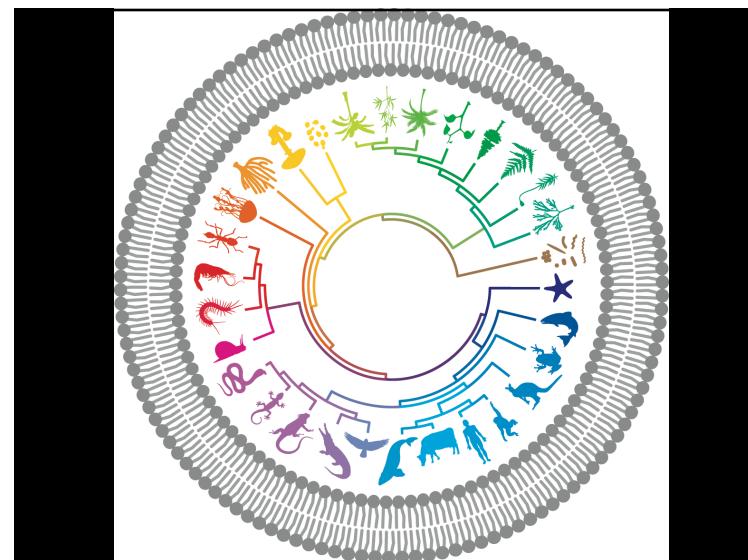
## Data Takes Many Forms

- Athletic performance
- Timeseries of polls
- Sequence Data
- Measurements of physical properties
- Maps (often with many layers) with information
- Timings of events
- Images
- Network descriptions
- Plain text

# What do you want to learn from data?

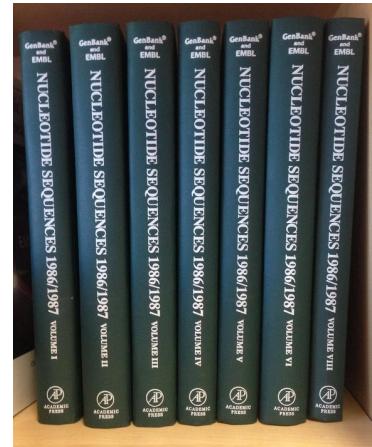
- Go to <https://datasetsearch.research.google.com/>
  - Find something cool
  - Write a few sentences or sketch a picture of what you want to learn from it
  - Tell your neighbor about it
  - You will introduce each other's "projects"

## Data is at the Center of Biology

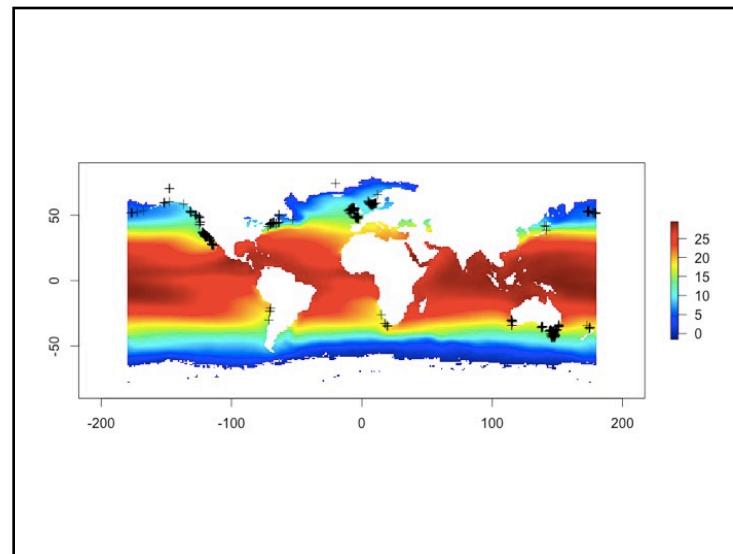
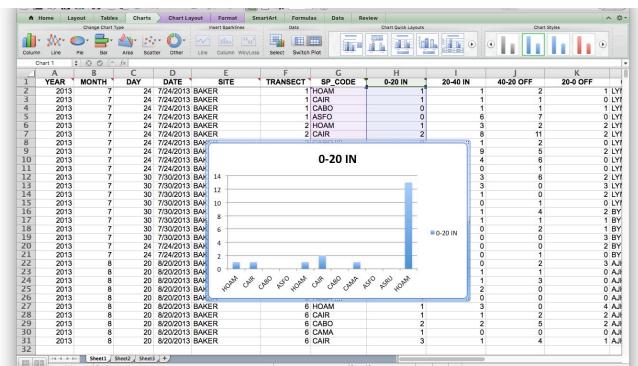


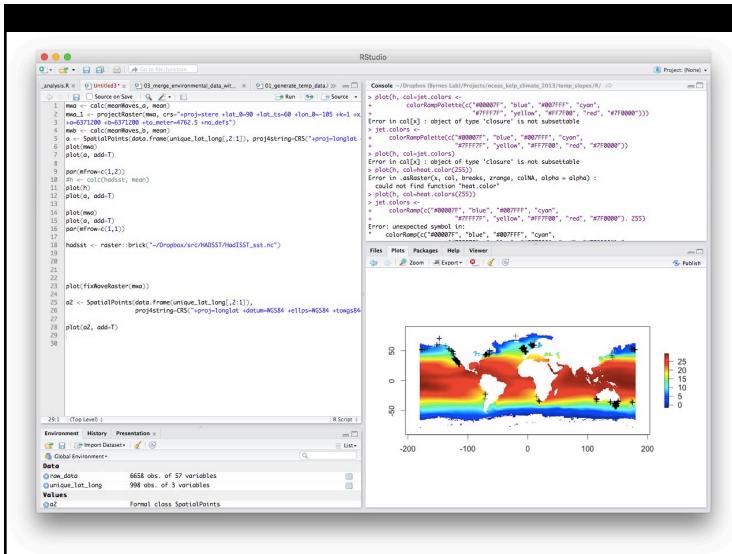


## Classical Tools Not Up to the Task



## Classical Tools Not Up to the Task

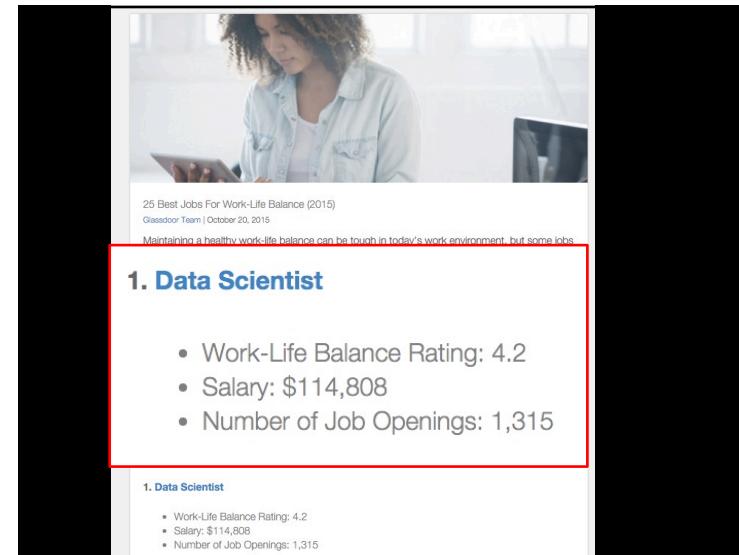


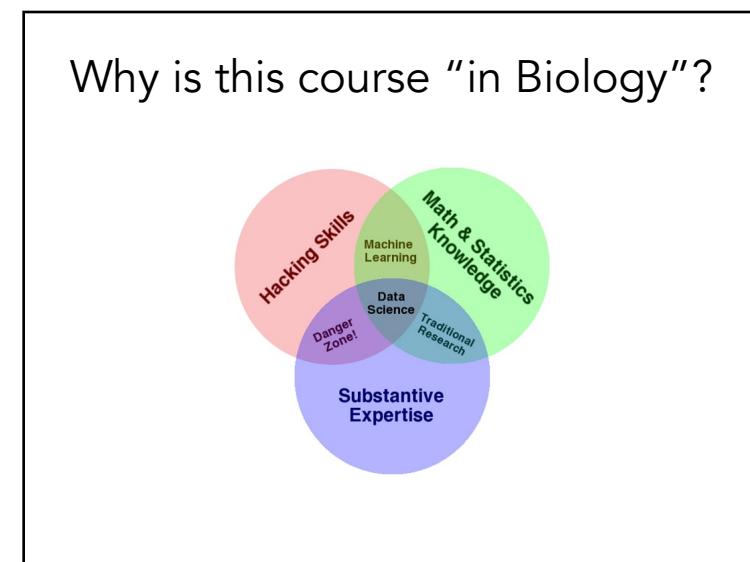
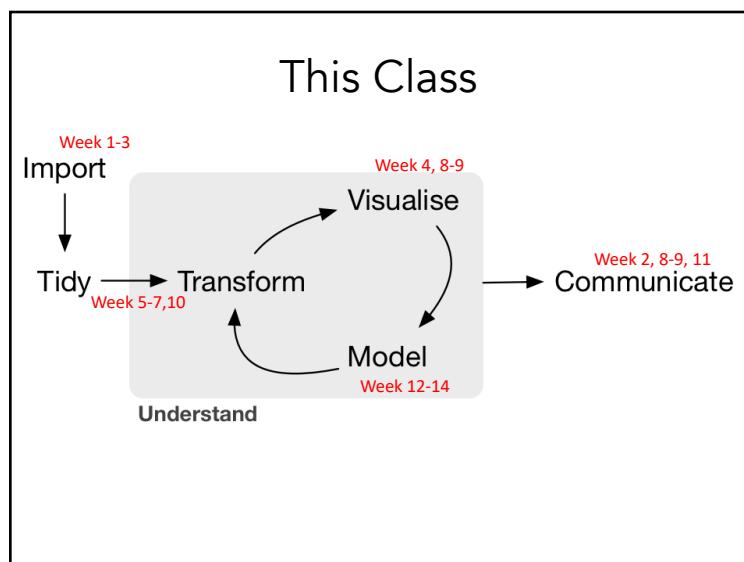
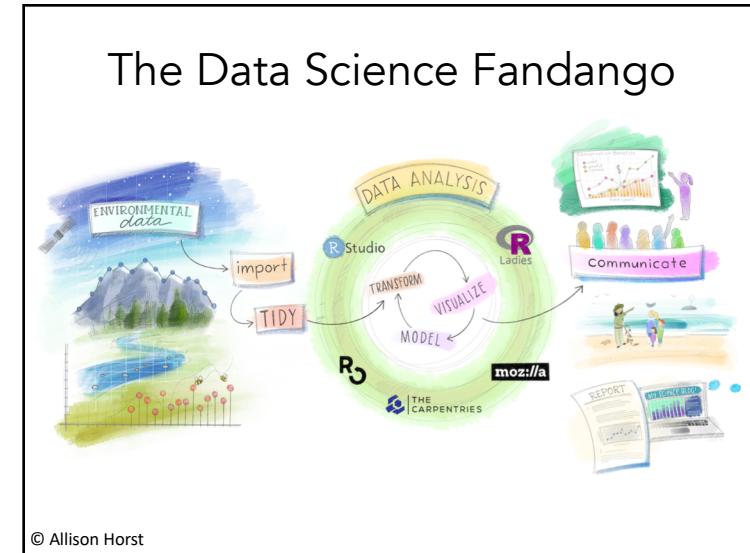
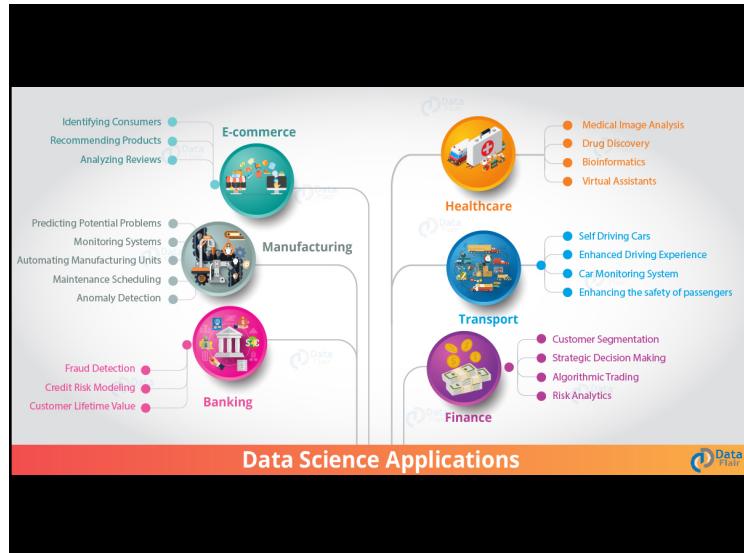


# So, programming...

- Write a few sentences about your experience with programming or, if you haven't before, how programming makes you feel.
  - Share with the four people in a breakout room
  - Report back about common themes and impressions

# What is Data Science?





## Introduction to Data Science for Biology

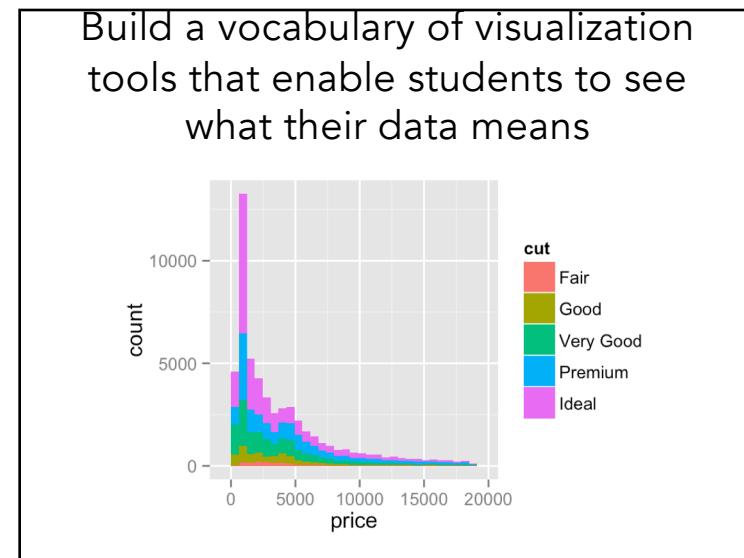
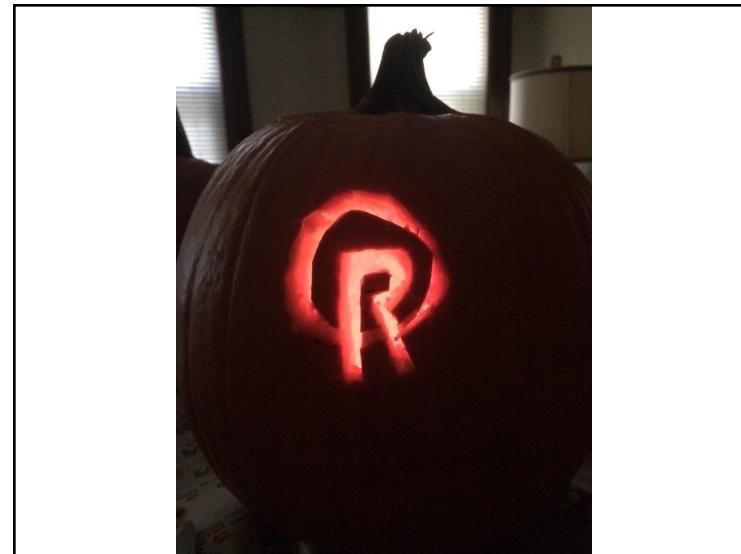
## Our Semester

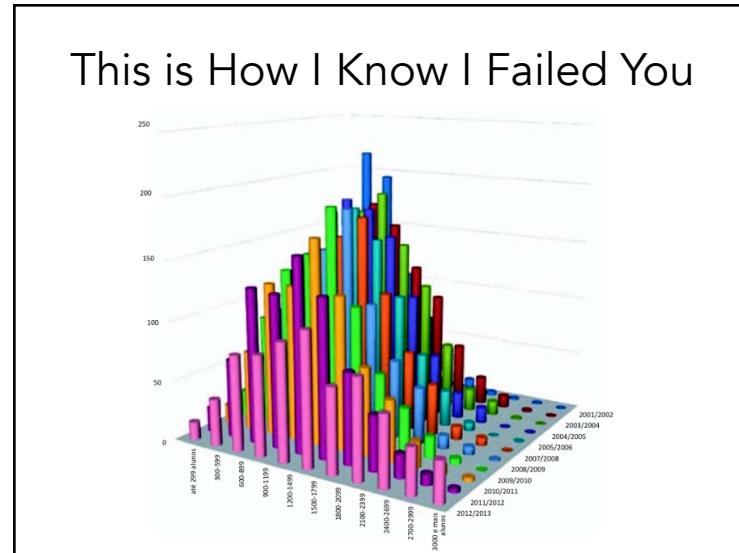
Learn how to create efficient  
understandable datasets for  
biological research

YEAR	MONTH	DAY	DATE	SITE	TRANSECT	SP_CODE	0-20 IN	20-40 IN	40-20 OFF
2013	7	24	3	BAKER	1HOAM		1	1	2
2013	7	24	3	BAKER	1CAIR		1	1	1
2013	7	24	3	BAKER	1CABO		0	1	1
2013	7	24	3	BAKER	1ASFO		0	6	7
2013	7	24	3	BAKER	2HOAM		1	3	2
2013	7	24	3	BAKER	2CAIR		2	8	11
2013	7	24	3	BAKER	2CABO		0	1	2
2013	7	24	3	BAKER	2CAMA		1	9	5
2013	7	24	3	BAKER	2ASFO		0	4	6
2013	7	24	3	BAKER	2ASRU		0	0	1

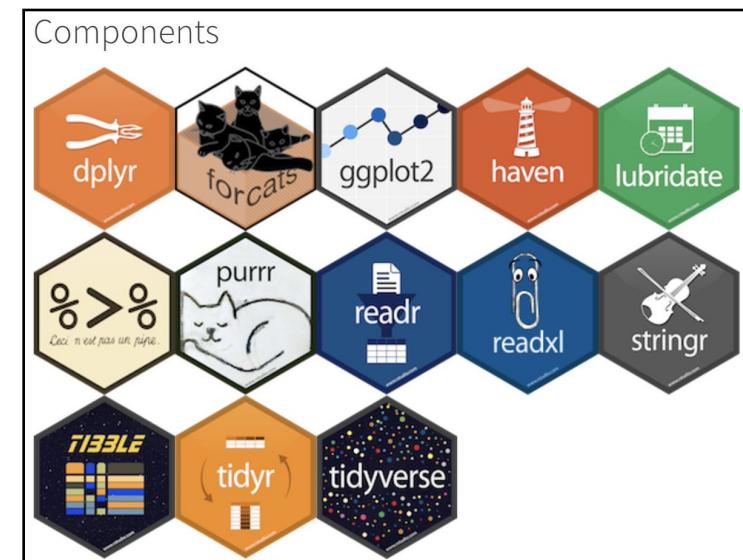
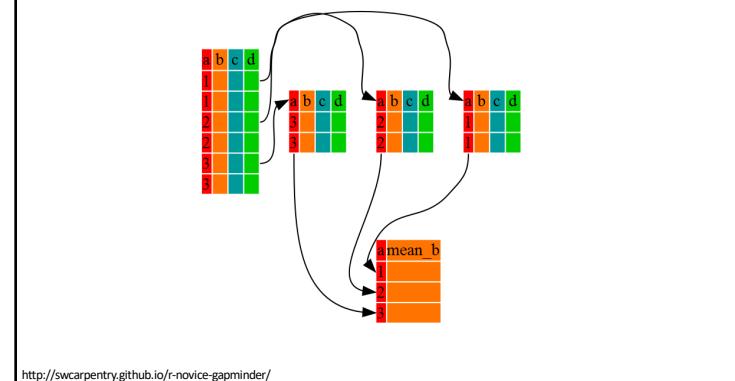
Learn common programming  
language(s) associated with data  
science

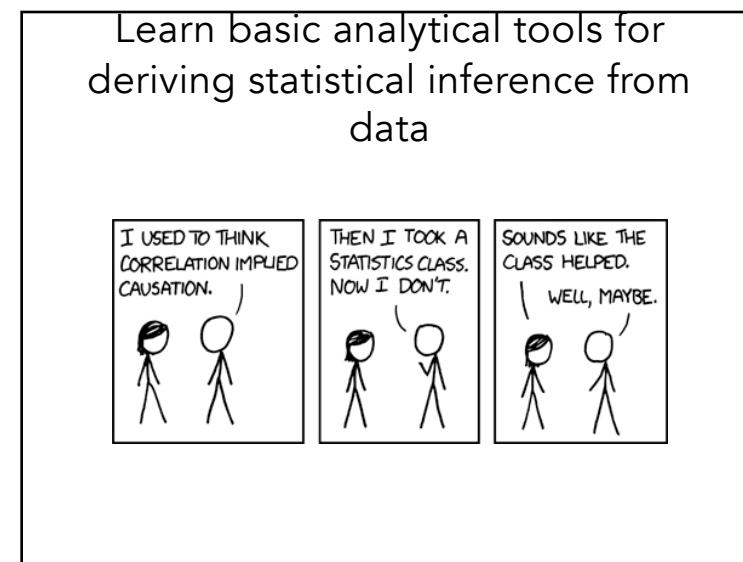
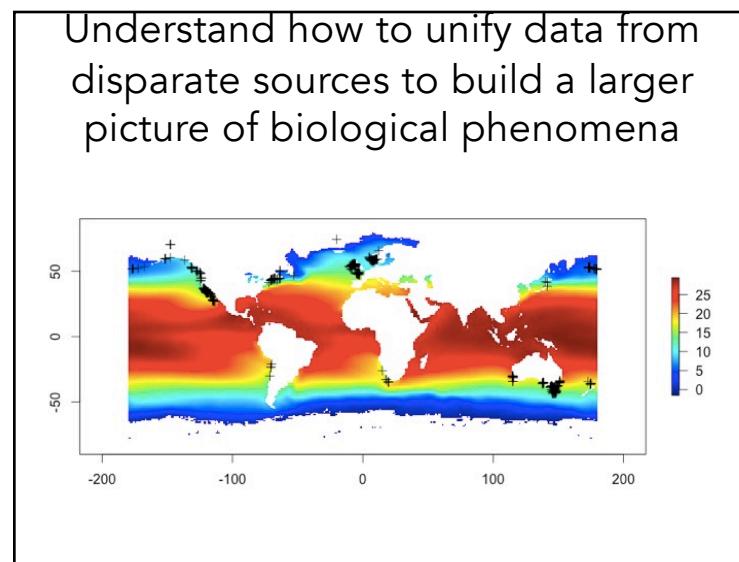
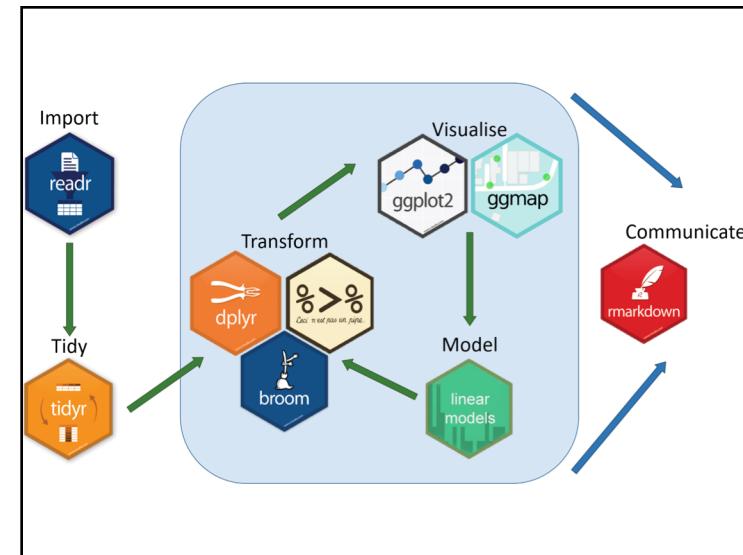
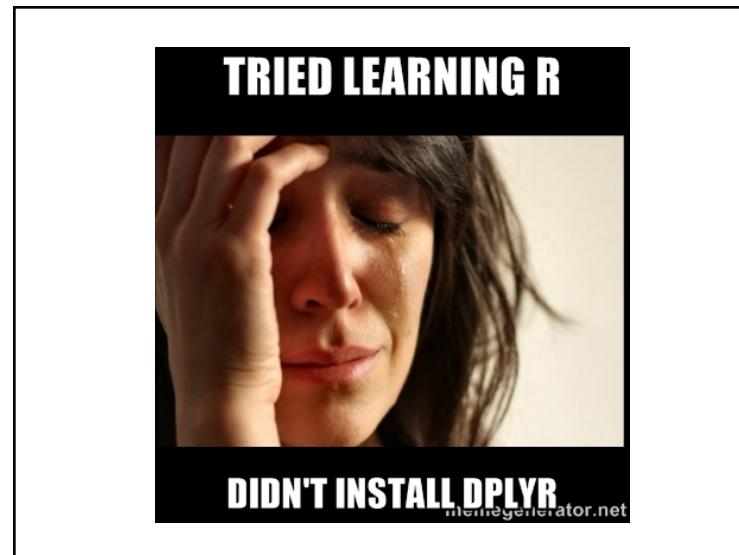






Develop an understanding of how to manipulate data for the purposes of seeing useful patterns





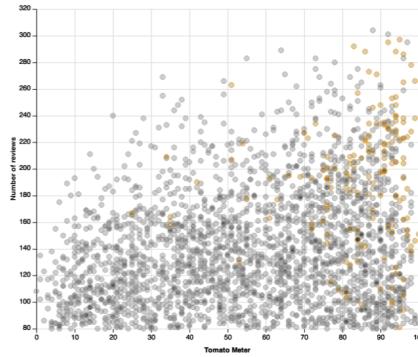
## Determine Strategies for Communicating Results of Data Explorations for Use by Others

**Movie explorer**

**Filter**

- Minimum number of reviews on Rotten Tomatoes: 10 to 300
- Year released: 1,900 to 2,014
- Minimum number of Oscar wins (all categories): 0 to 4
- Dollars at Box Office (millions): 0 to 800
- Genre (a movie can have multiple genres): All

X-axis variable: Tomato Meter

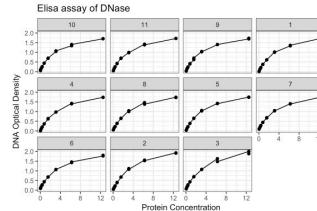


## This Class

## Course Web Page

[Home](#) [Overview](#) [Schedule](#) [R Errors](#) [Final Project](#) [Syllabus](#) [Resources](#) [Local Meetups](#)

### Biol 355/356: Intro to Data Science for Biology



Instructor: Jarrett Byrnes, PhD.

Email: [jarrett.byrnes@umb.edu](mailto:jarrett.byrnes@umb.edu)

TA: Isaac Rosenthal

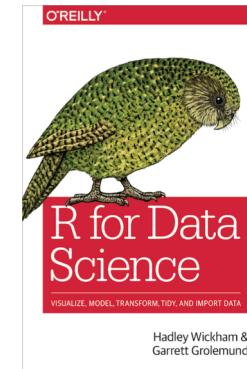
Email: [isaacrosenthal001@umb.edu](mailto:isaacrosenthal001@umb.edu)

Weekly Schedule: Tuesday & Thursday 9:30-12:00, Lab Wednesday 12:30-3:30

Office Hours: Prof. Byrnes will hold office hours Thursday from 1:30-3 in ISC 3130

<http://biol355.github.io>

## "Text"book & Weekly Readings



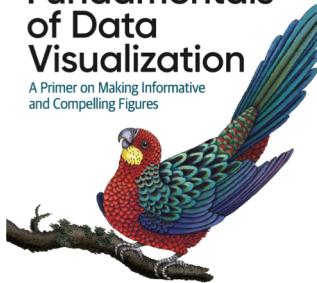
<http://r4ds.had.co.nz/>

## Additional Resources

O'REILLY®

**Fundamentals of Data Visualization**

A Primer on Making Informative and Compelling Figures



Claus O. Wilke

<http://r4ds.had.co.nz/>

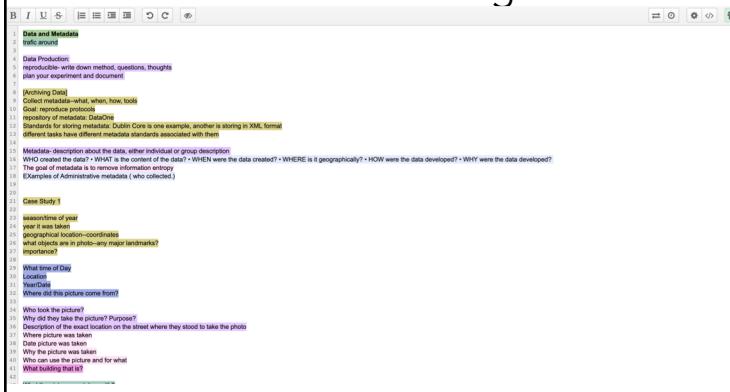


**DATA CARPENTRY**

BUILDING COMMUNITIES TEACHING UNIVERSAL DATA LITERACY

<http://www.datacarpentry.org/>

## Etherpad: How we will Communicate During Class



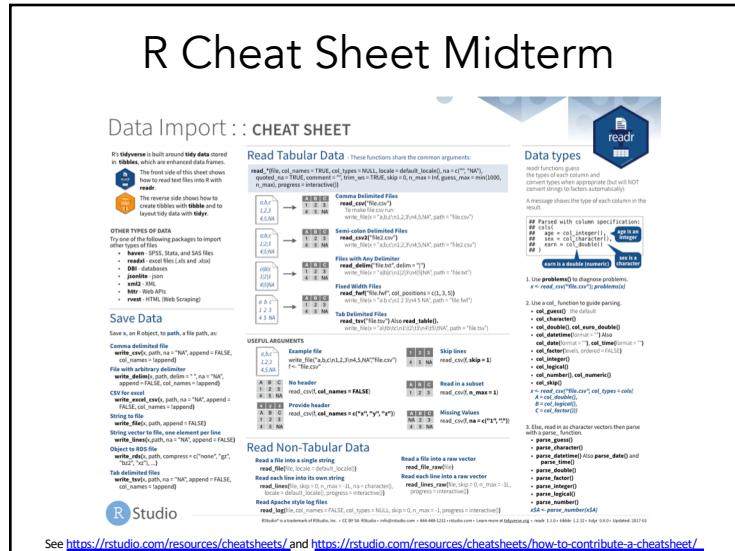
## Lab

- Coding!
- TA: Michael Roy
- Guided examples and then challenge problems

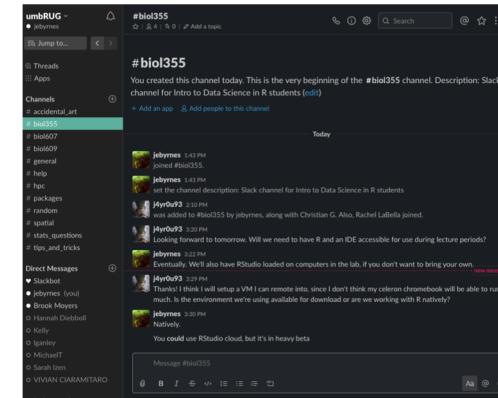
# Assignments

- Weekly problem sets
    - Variable in scope!
    - May involve elements of your final project
  - Can be started in lab
    - Will highlight concepts from that week

# R Cheat Sheet Midterm



## Slack for out of class interactions



- Weird R errors?
  - Questions?
  - Something nerdy and funny

## Final Project

- Interactive presentation of data
    - Can be from your own work
    - See examples at  
[https://biol355.github.io/projects\\_2020.html](https://biol355.github.io/projects_2020.html)
  - Data mashups encouraged!
    - Bring together multiple public sources of data
  - Proposals due in three weeks
    - What data will you be using?
    - What question do you want to answer?

Next Time: Data Collection,  
Entry, and How to Make Your  
Data Usable

(and have future you avoid  
wanting to kill now you)

**(And listen to the Not So Standard Deviations Podcast)**

Friday: Lab – what does a data  
collection process look like?