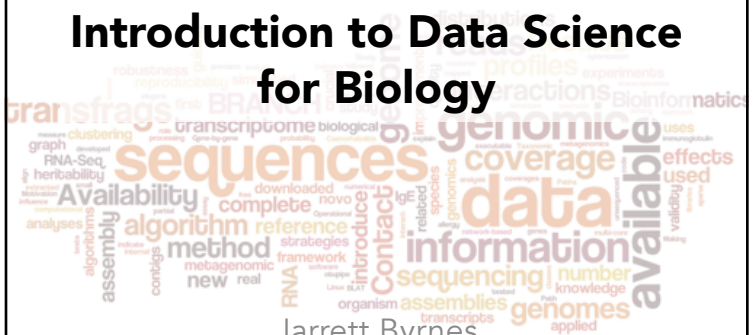
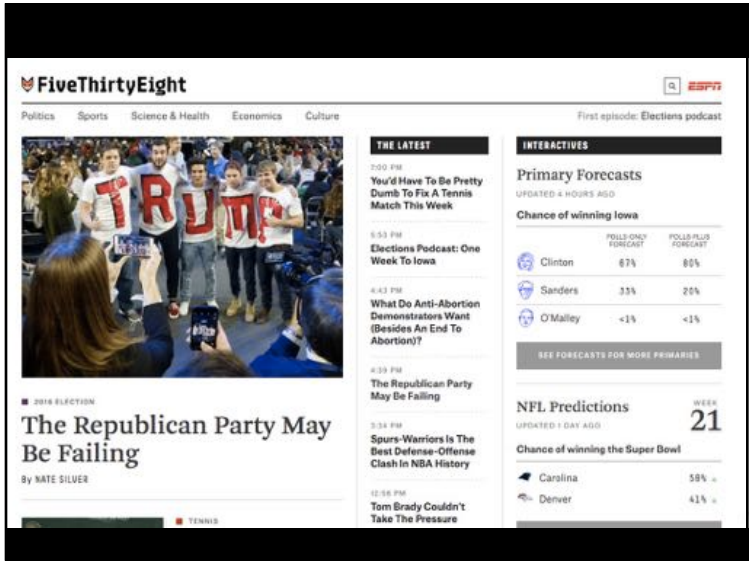
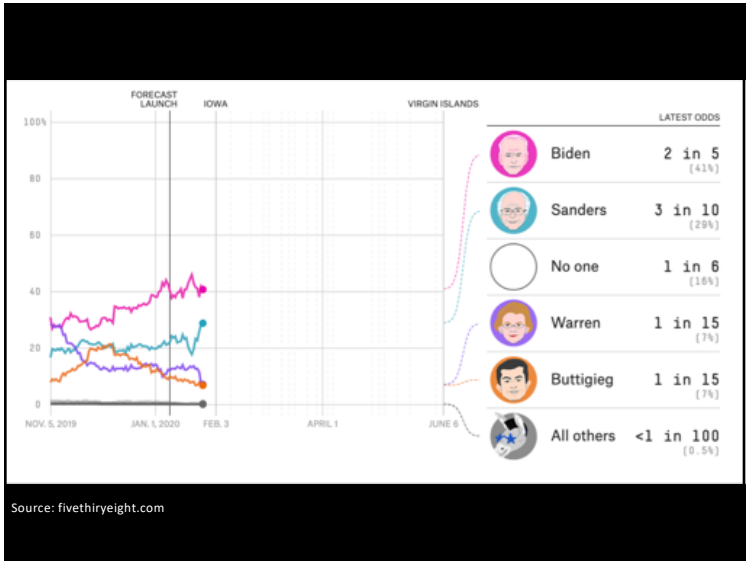
Introduction to Data Science for Biology



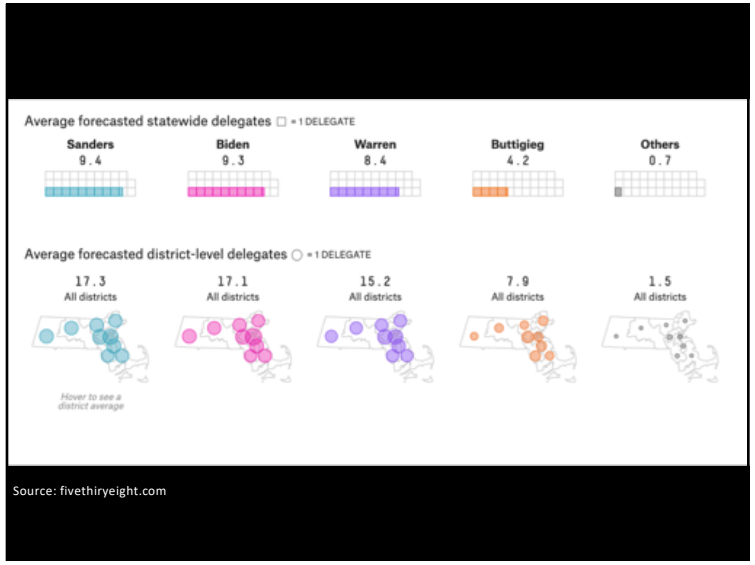
Jarrett Byrnes
UMass Boston
Spring 2020

We Are Awash in Data

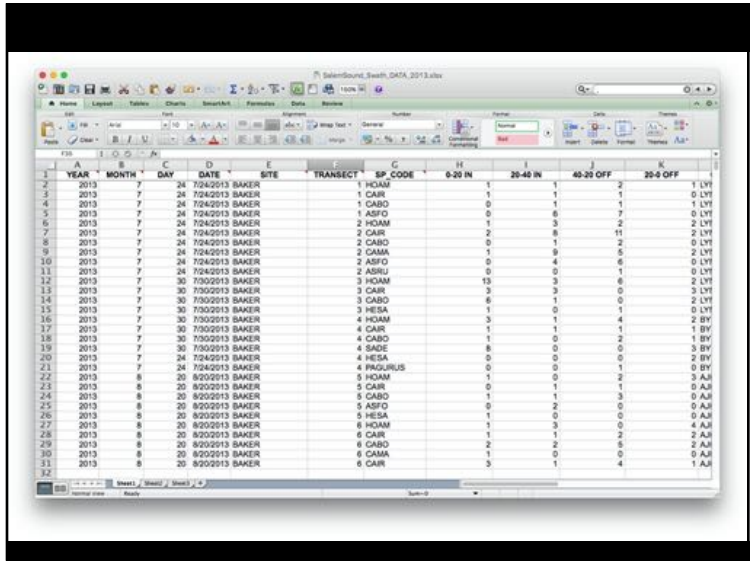


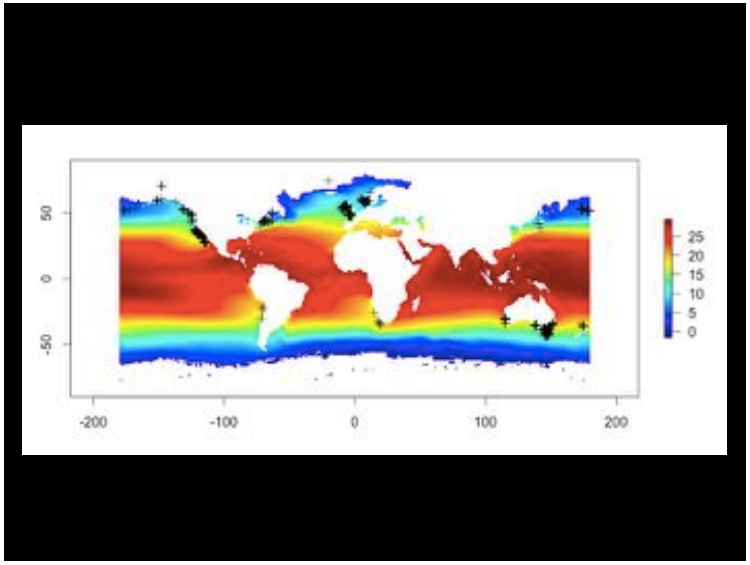


Source: fivethiryeight.com



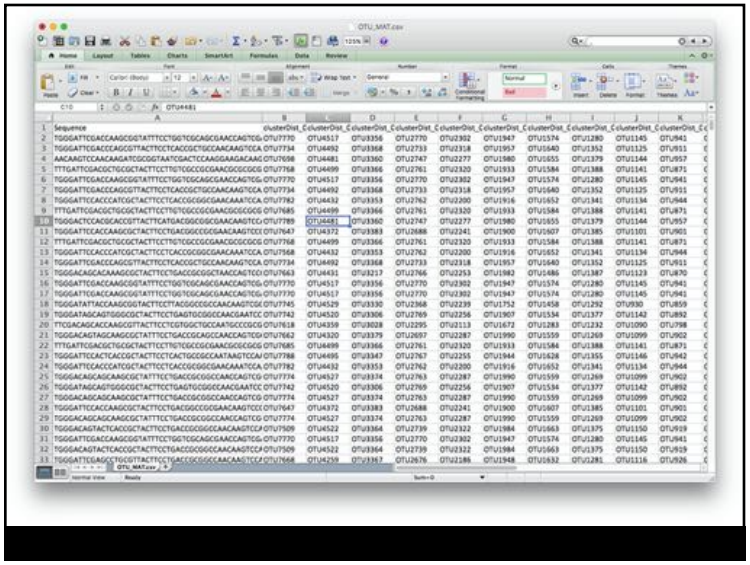
Source: fivethiryeight.com

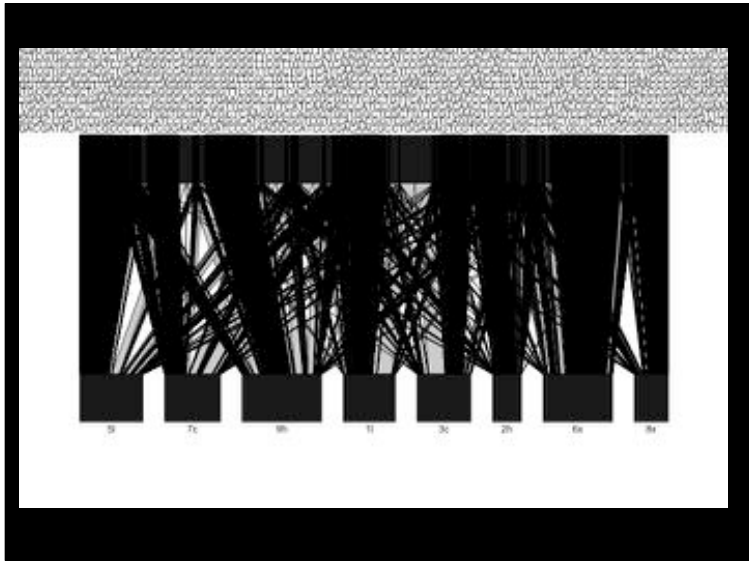


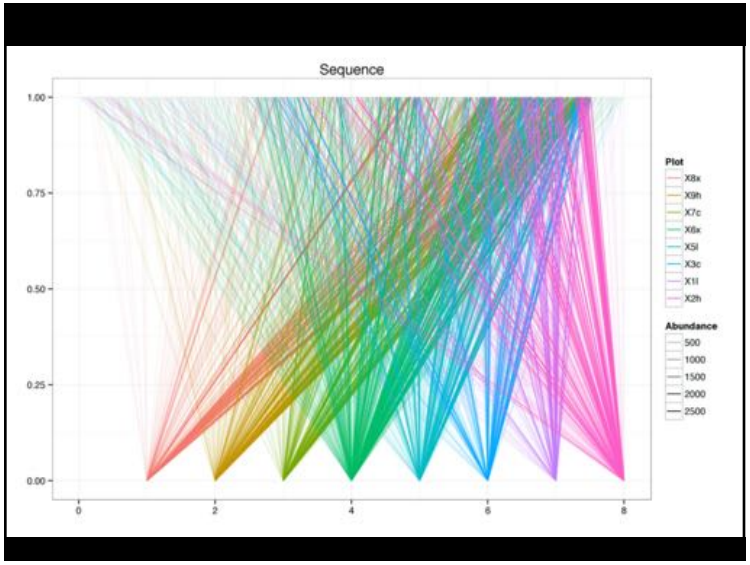


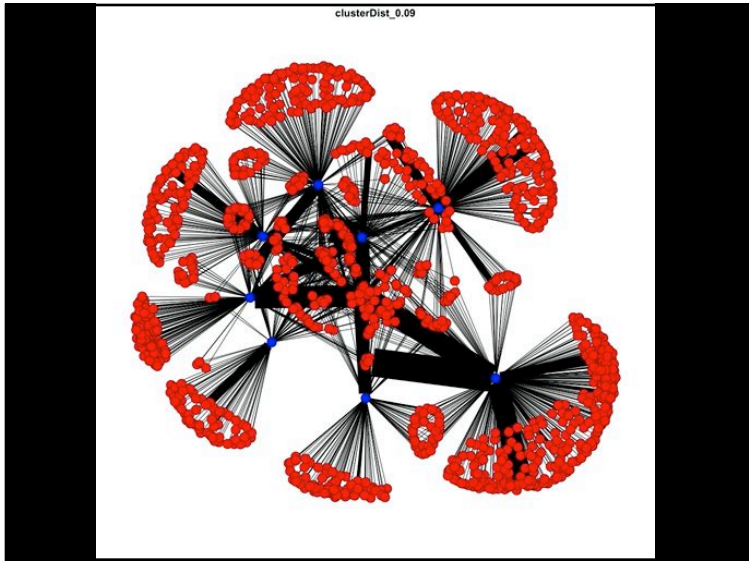
SYNTENIC ASSEMBLIES FOR CG15386

```
MD106   ATGCTTAGTAATCCCTACTTTAAGTCCGTTTTGTGGCTGATTGGCTTCGGAGGAATGGG
NEWC    ATGCTTAGTAATCCCTACTTTAAATCCGTTTTGTGGCTGATTGGCTTCGGAGGAATGGG
W501    ATGCTTAGTAATCCCTACTTTAAGTCCGTTTTGTGGCTGATTGGCTTCGGAGGAATGGG
MD199   ATGCTTAGTAATCCCTACTTTAAGTCCGTTTTGTGGCTGATTGGCTTCGGAGGAATGGG
C1674   ATGCTTAGTAATCCCTACTTTAAGTCCGTTTTGTGGCTGATTGGCTTCGGAGGAATGGG
SIM4    ATGCTTAGTAATCCCTACTTTAAGTCCGTTTTGTGGCTGATTGGCTTCGGAGGAATGGG

MD106   CTACGGCCTAATGGTGCTAACAGAGCCGAACGTCGACAAAATAGAGCGCATCAAAGCCT
NEWC    CTACGGCCTAATGGTGCTAACCGAGCCGAACGTCGACAAAATAGAGCGCATCAAAGCCT
W501    CTACGGCCTAATGGTGCTAACCGAGCCGAACGTCGACAAAATAGAGCGCATCAAAGCCT
MD199   CTACGGCCTAATGGTGCTAACCGAGCCGAACGTCGACAAAATAGAGCGCATCAAAGCCT
C1674   CTACGGCCTAATGGTGCTAACCGAGCCGAACGTCGACAAAATAGAGCGCATCAAAGCCT
SIM4    CTACGGCCTAATGGTGCTAACCGAGCCGAACGTCGACAAAATAGAGCGCATCAAAGCCT

MD106   CCGTTTCAAGTACCAAACTGAGTGCGGATGAGCAGCGAAAGGCTCTGTTTATGAAGAAG
NEWC    CCGTTTCAAGTACCAAACTGAGTGCGGATGAGCAGCGAAAGGCTCTGTTTATGAAGAAG
W501    CCGTTTCAAGTACCAAACTGAGTGCGGATGAGCAGCGAAAGGCTCTGTTTATGAAGAAG
MD199   CCGTTTCAAGTACCAAACTGAGTGCGGATGAGCAGCGAAAGGCTCTGTTTATGAAGAAG
C1674   CCGTTTCAAGTACCAAACTGAGTGCGGATGAGCAGCGAAAGGCTCTGTTTATGAAGAAG
SIM4    CCGTTTCAAGTACCAAACTGAGTGCGGATGAGCAGCGAAAGGCTCTGTTTATGAAGAAG

MD106   CTGCAGGAGGCGTCCACCACCAGTGCCCCAATCTACAGGTCAGCGGCCGAGAAATAG
NEWC    CTGCAGGAGGCGTCCACCACCAGTGCCCCAATCTACAGGTCATCGGCCGAGAAATAG
W501    CTGCAGGAGGCGTCCACCACCACTGCCCCAATCTACAGGTCATCGGCCGAGAAATAG
MD199   CTGCAGGAGGCGTCCACCACCAGTGCCCCAATCTACAGGTCAGCGGCCGAGAAATAG
C1674   CTGCAGGAGGCGTCCACCACCAGTGCCCCAATCTACAGGTCAGCGGCCGAGAAATAG
SIM4    CTGCAGGAGGCGTCCACCACCAGTGCCCCAATCTACAGGTCAGCGGCCGAGAAATAG
```

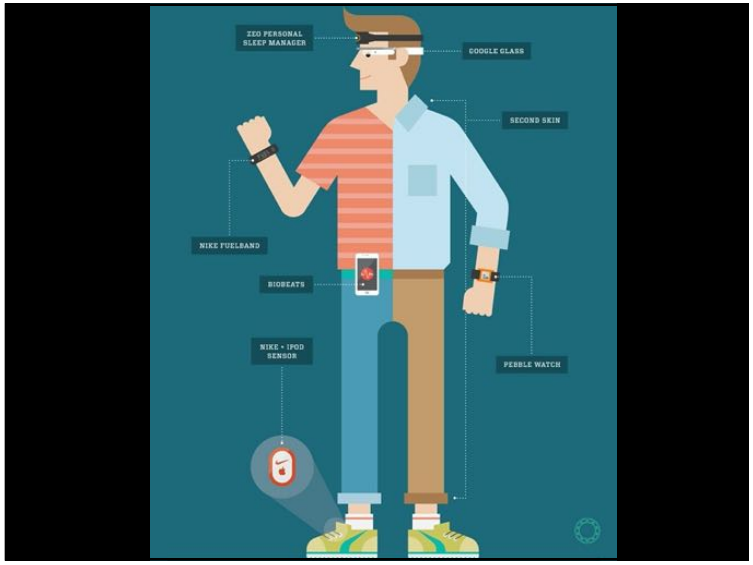




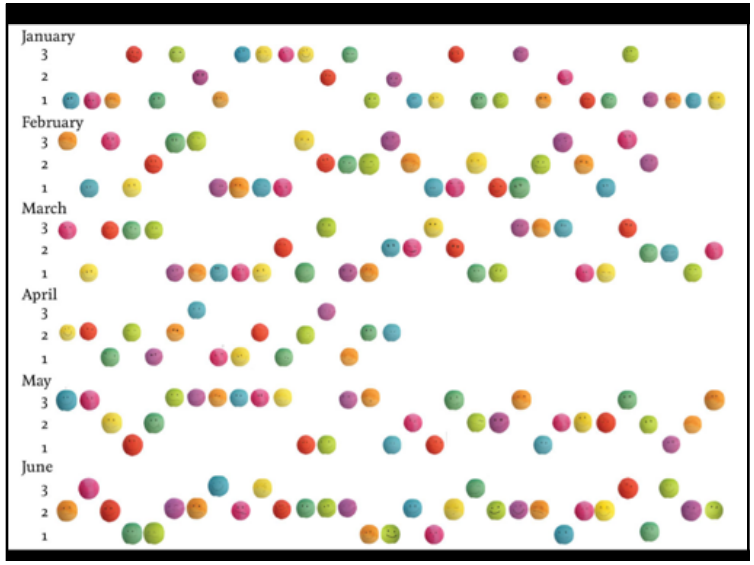


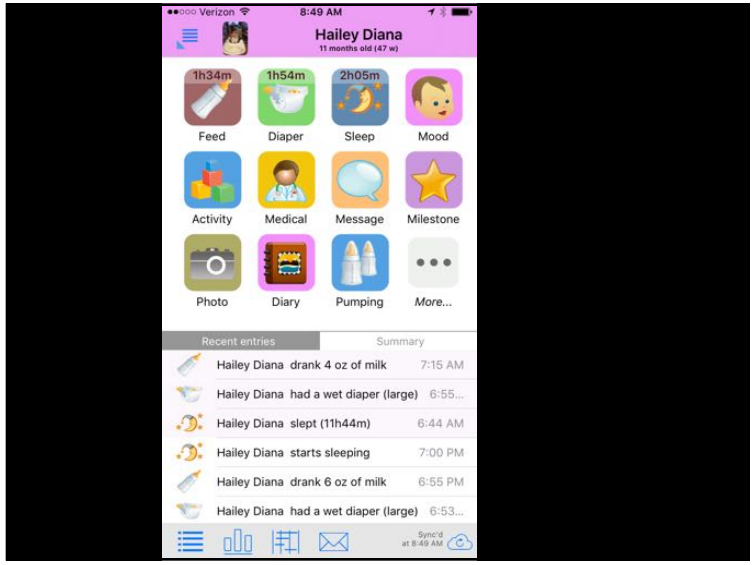

clusterDist_0.09


8:44


ZEO PERSONAL SLEEP MANAGER
GOOGLE GLASS
SECOND SKIN
NIKE FUELBAND
BIOBEATS
NIKE + IPOD SENSOR
PEBBLE WATCH


QS
Quantified Self
self knowledge through numbers


January
3
2
1
February
3
2
1
March
3
2
1
April
3
2
1
May
3
2
1
June
3
2
1


Verizon
8:49 AM
Hailey Diana
11 months old (47 w)
1h34m
1h54m
2h05m
Feed
Diaper
Sleep
Mood
Activity
Medical
Message
Milestone
Photo
Diary
Pumping
More...
Recent entries
Summary
Hailey Diana drank 4 oz of milk 7:15 AM
Hailey Diana had a wet diaper (large) 6:55...
Hailey Diana slept (11h44m) 6:44 AM
Hailey Diana starts sleeping 7:00 PM
Hailey Diana drank 6 oz of milk 6:55 PM
Hailey Diana had a wet diaper (large) 6:53...
Sync'd at 8:49 AM


Verizon
8:49 AM
Back
Hailey Diana
11 months old (47 w)
Timeline
Sleep
Diaper
Bottle
Solid
Activity
12AM
4AM
8AM
12PM
4PM
8PM
12AM

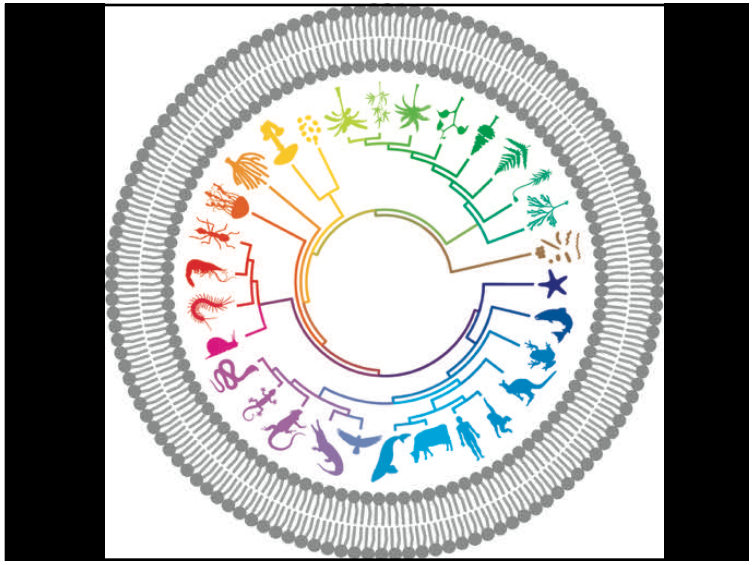
Data Takes Many Forms

- 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

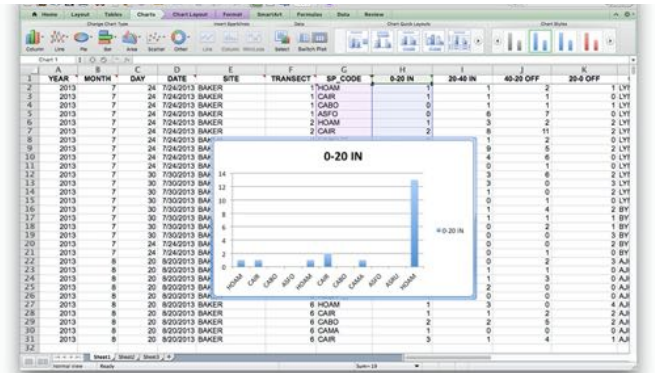
What do you want to learn from data?

- 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"

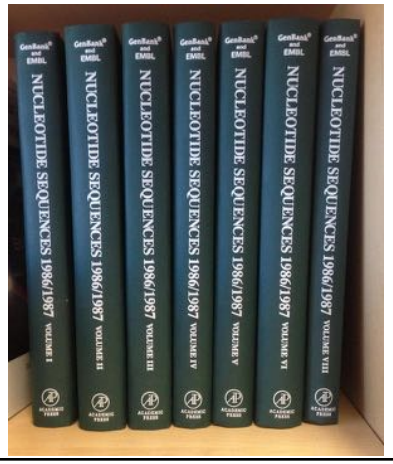
Data is at the Center of Biology

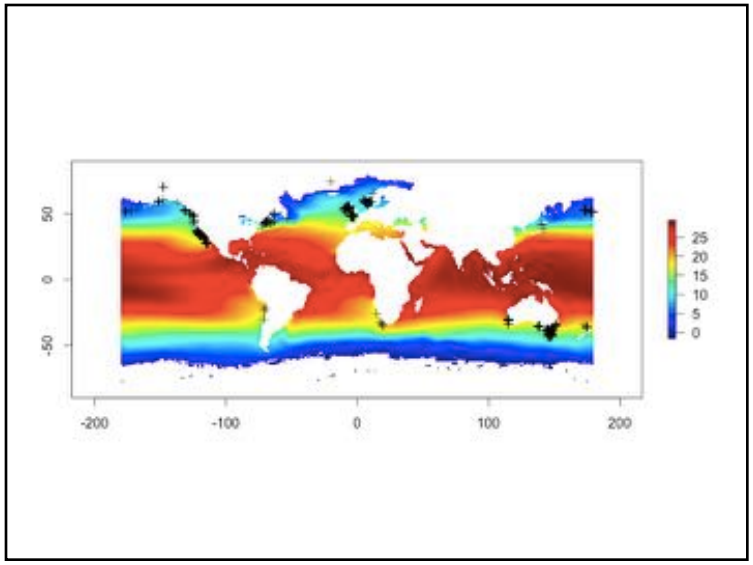


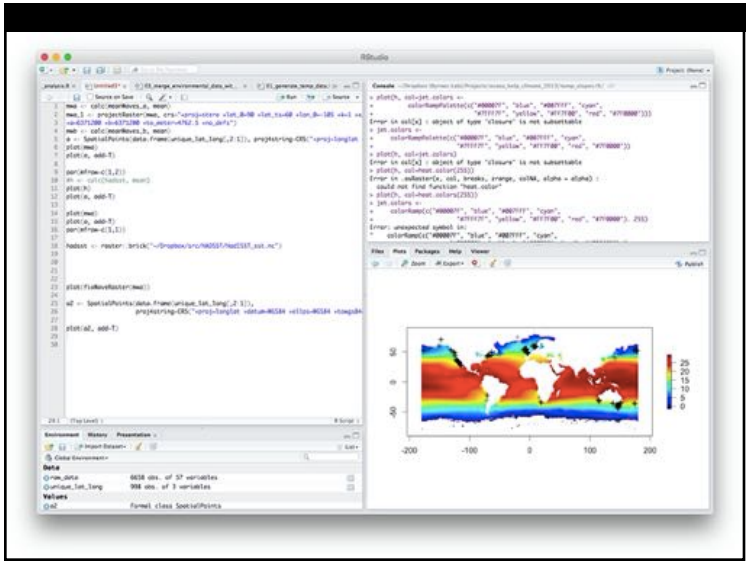
Classical Tools Not Up to the Task



Classical Tools Not Up to the Task



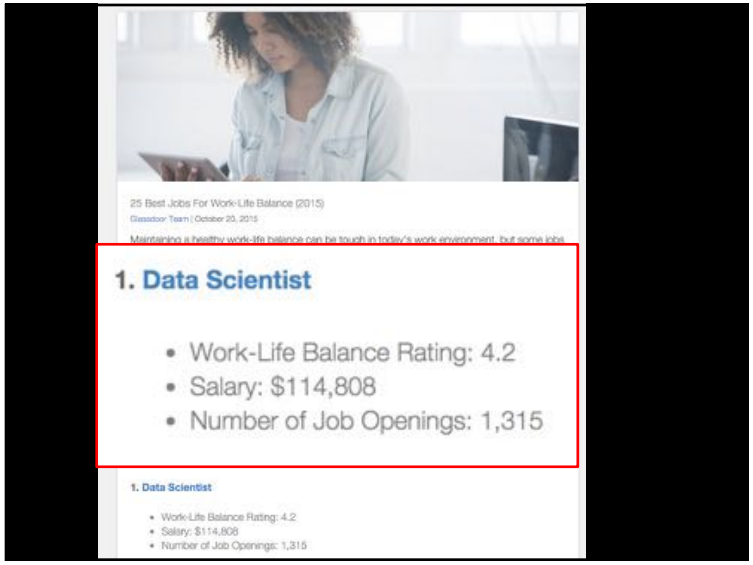


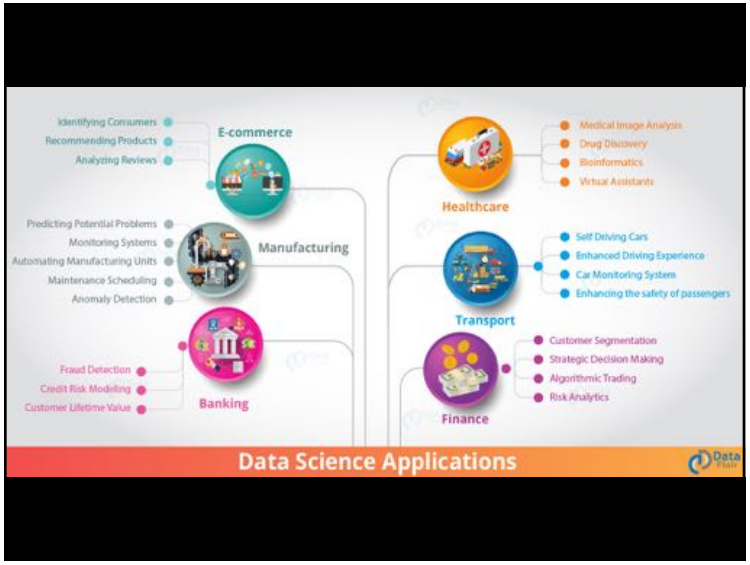


So, programming…

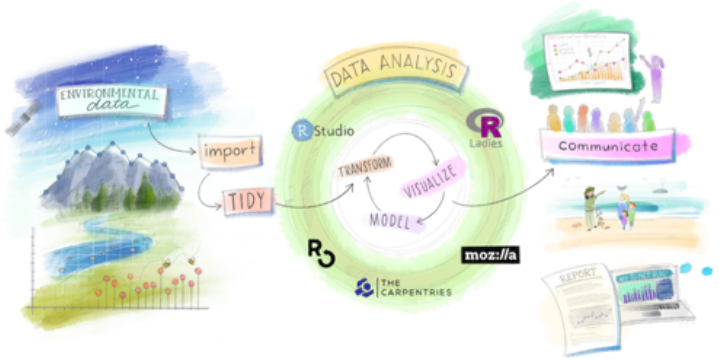
- Write a few sentences about your experience with programming or, if you haven't before, how programming makes you feel.
- Share with the four people around you
- Report back about common themes and impressions

What is Data Science?

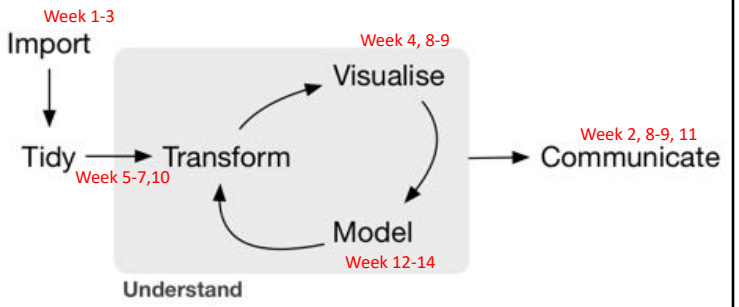




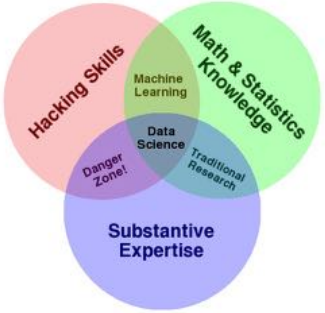
The Data Science Fandango



This Class



Why is this course "in Biology"?



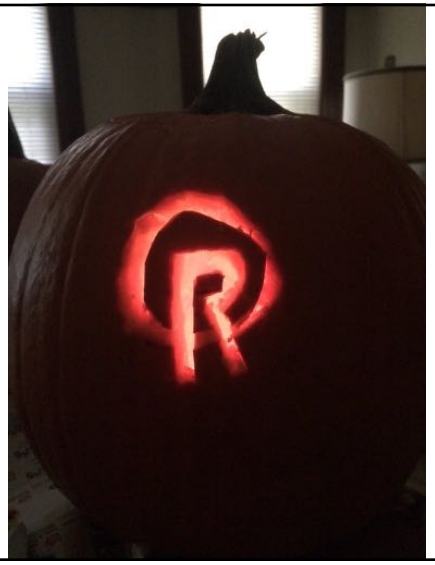
Introduction to Data Science for Biology

Our Semester

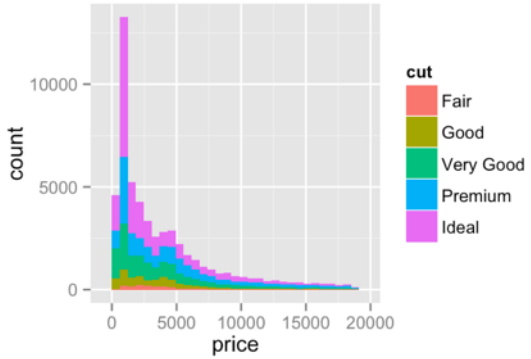
Learn how to create efficient understandable datasets for biological research

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

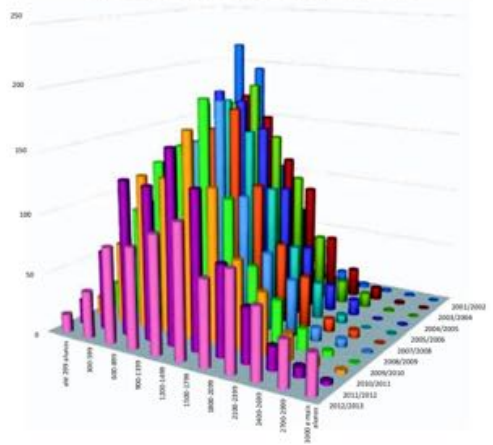
Learn common programming language(s) associated with data science



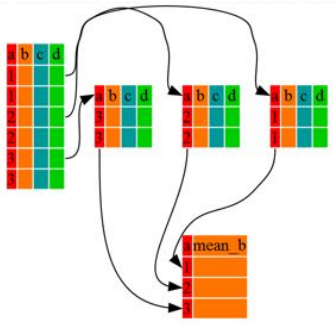
Build a vocabulary of visualization tools that enable students to see what their data means



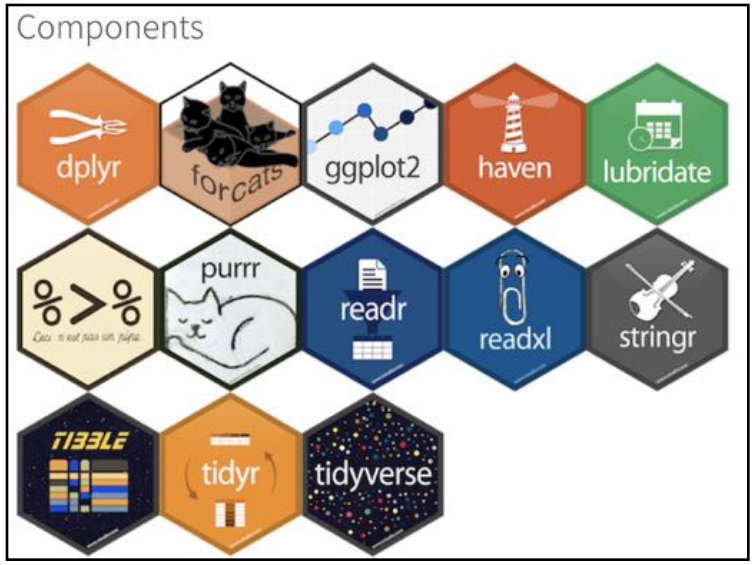
This is How I Know I Failed You

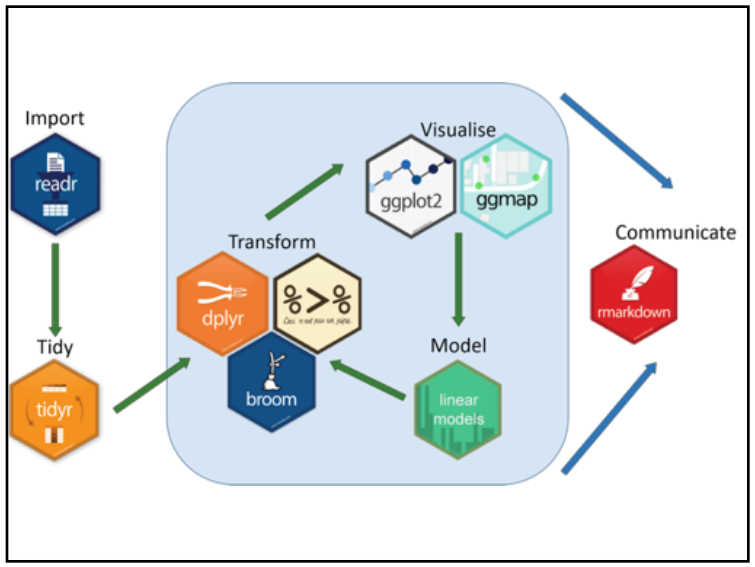


Develop an understanding of how to manipulate data for the purposes of seeing useful patterns

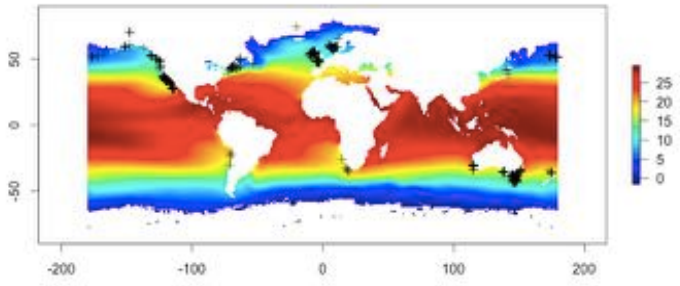


http://swcarpentry.github.io/r-novice-gapminder/

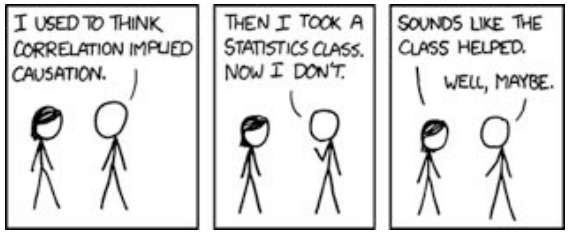




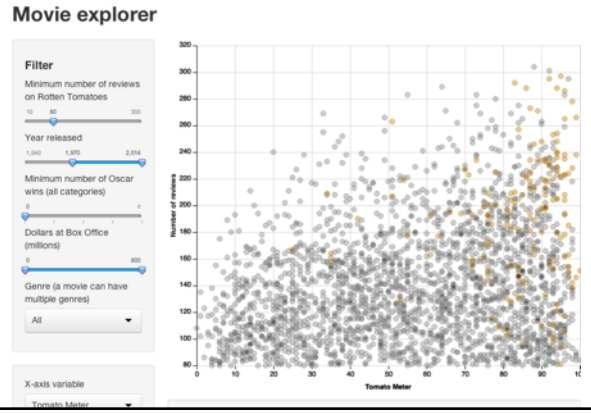
Understand how to unify data from disparate sources to build a larger picture of biological phenomena



Learn basic analytical tools for deriving statistical inference from data

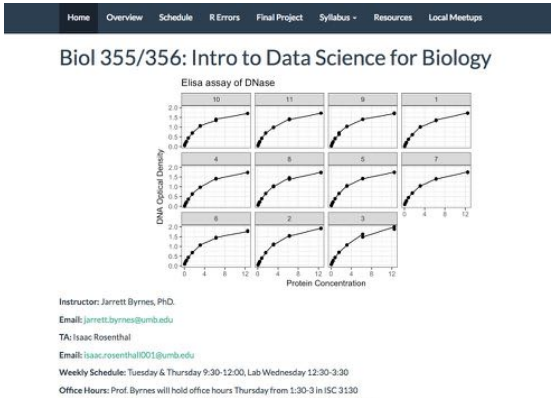


Determine Strategies for Communicating Results of Data Explorations for Use by Others



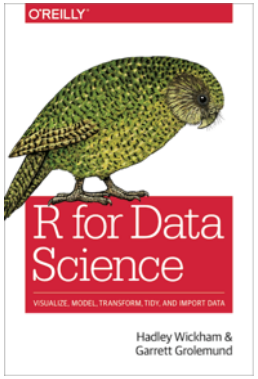
This Class

Course Web Page



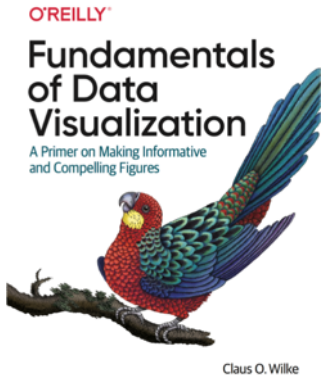
http://biol355.github.io

"Text"book & Weekly Readings



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

Additional Resources



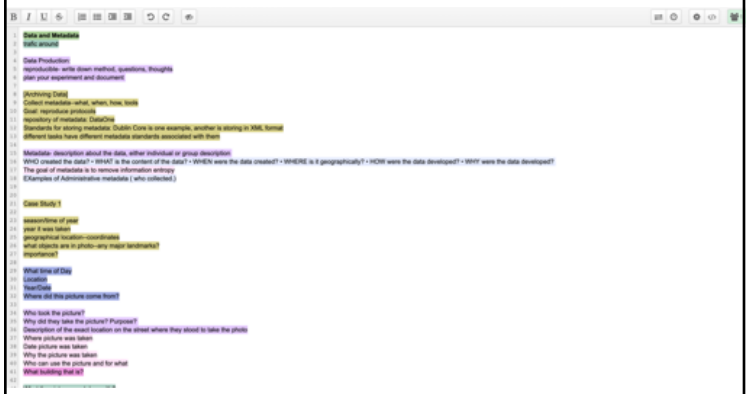
http://r4ds.had.co.nz/

DATA CARPENTRY

BUILDING COMMUNITIES TEACHING UNIVERSAL DATA LITERACY

http://www.datacarpentry.org/

Collaborative Note Taking



Lab

- Coding!
- TA: Rachel La Bella
- Guided examples and then challenge problems

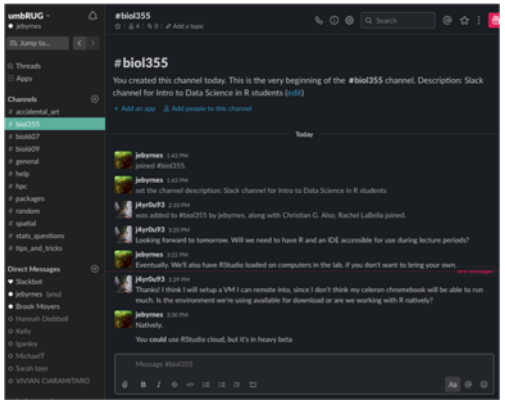
Keeping Track of Interacting

- Green means good!
- Red means HALP!
- Blue is for feedback….

Assignments

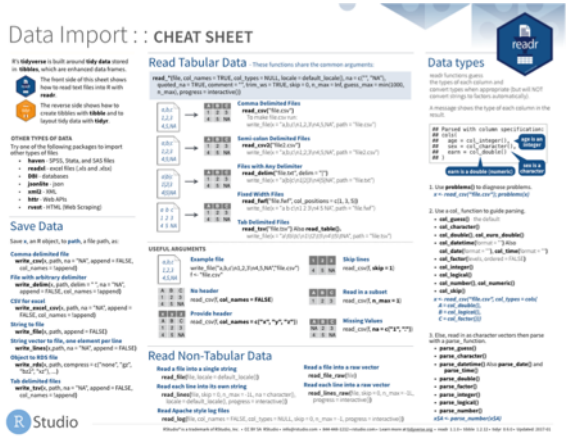
- Weekly problem sets
  - Variable in scope!
  - May involve elements of your final project
- Can be started in lab
  - Will highlight concepts from that week

Slack for out of class interactions



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

R Cheat Sheet Midterm



See https://rstudio.com/resources/cheatsheets/ and https://rstudio.com/resources/cheatsheets/how-to-contribute-a-cheatsheet/

Final Project

- Analysis of a data set of your choosing
  - From your own work
  - Found data
- 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?