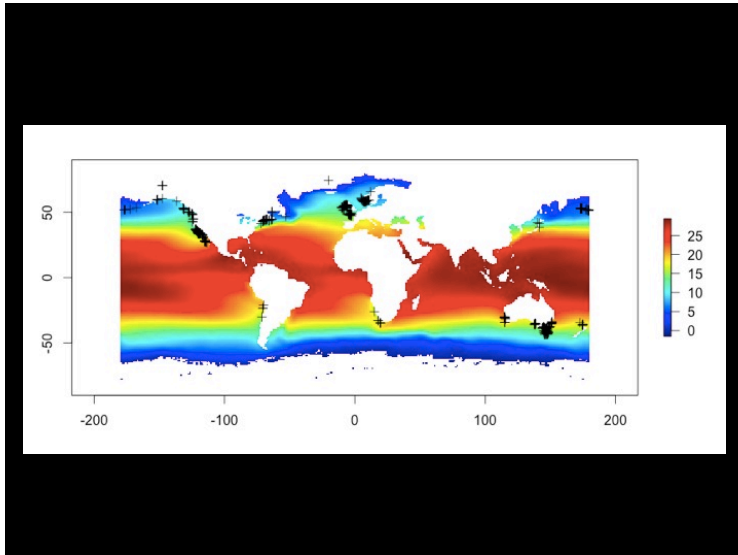






SalemSound\_Swath\_DATA\_2013.xlsx

YEAR	MONTH	DAY	DATE	SITE	TRANSECT	SP_CODE	0-20 IN	20-40 IN	40-20 OFF	20-0 OFF
2013	7	24	7/24/2013	BAKER	1	HOAM	1	1	2	1
2013	7	24	7/24/2013	BAKER	1	CAIR	1	1	1	0
2013	7	24	7/24/2013	BAKER	1	CABO	0	1	1	1
2013	7	24	7/24/2013	BAKER	1	ASFO	0	6	7	0
2013	7	24	7/24/2013	BAKER	2	HOAM	1	3	2	2
2013	7	24	7/24/2013	BAKER	2	CAIR	2	8	11	2
2013	7	24	7/24/2013	BAKER	2	CABO	0	1	2	0
2013	7	24	7/24/2013	BAKER	2	CAMA	1	9	5	2
2013	7	24	7/24/2013	BAKER	2	ASFO	0	4	6	0
2013	7	24	7/24/2013	BAKER	2	ASRU	0	0	1	0
2013	7	30	7/30/2013	BAKER	3	HOAM	13	3	6	2
2013	7	30	7/30/2013	BAKER	3	CAIR	3	3	0	3
2013	7	30	7/30/2013	BAKER	3	CABO	6	0	0	2
2013	7	30	7/30/2013	BAKER	3	HESA	1	0	1	0
2013	7	30	7/30/2013	BAKER	4	HOAM	3	1	4	2
2013	7	30	7/30/2013	BAKER	4	CAIR	1	1	1	1
2013	7	30	7/30/2013	BAKER	4	CABO	1	0	2	1
2013	7	30	7/30/2013	BAKER	4	SADE	8	0	0	3
2013	7	24	7/24/2013	BAKER	4	HESA	7	0	0	2
2013	7	24	7/24/2013	BAKER	4	PAGURUS	0	0	1	0
2013	8	20	8/20/2013	BAKER	5	HOAM	1	0	2	3
2013	8	20	8/20/2013	BAKER	5	CAIR	0	1	1	0
2013	8	20	8/20/2013	BAKER	5	CABO	1	1	3	0
2013	8	20	8/20/2013	BAKER	5	ASFO	0	0	0	0
2013	8	20	8/20/2013	BAKER	5	HESA	1	0	0	0
2013	8	20	8/20/2013	BAKER	6	HOAM	1	3	0	4
2013	8	20	8/20/2013	BAKER	6	CAIR	1	1	2	2
2013	8	20	8/20/2013	BAKER	6	CABO	2	2	5	2
2013	8	20	8/20/2013	BAKER	6	CAMA	1	0	0	0
2013	8	20	8/20/2013	BAKER	6	CAIR	3	1	4	1



SYNTENIC ASSEMBLES FOR CG15386

```

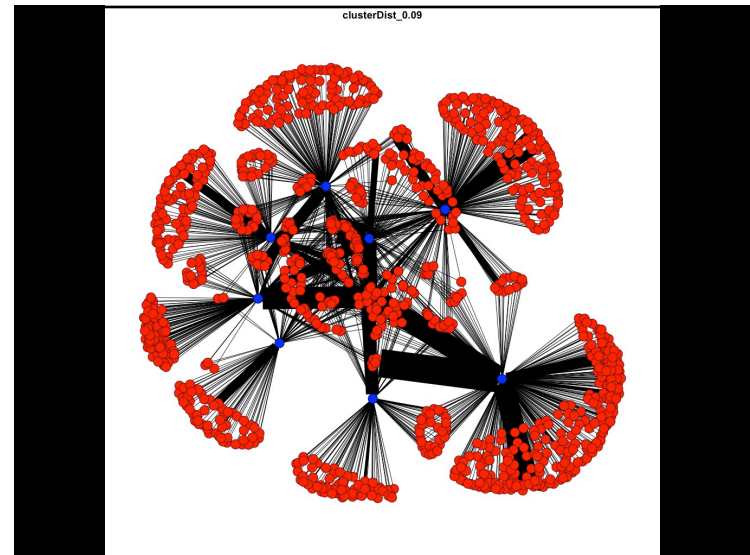
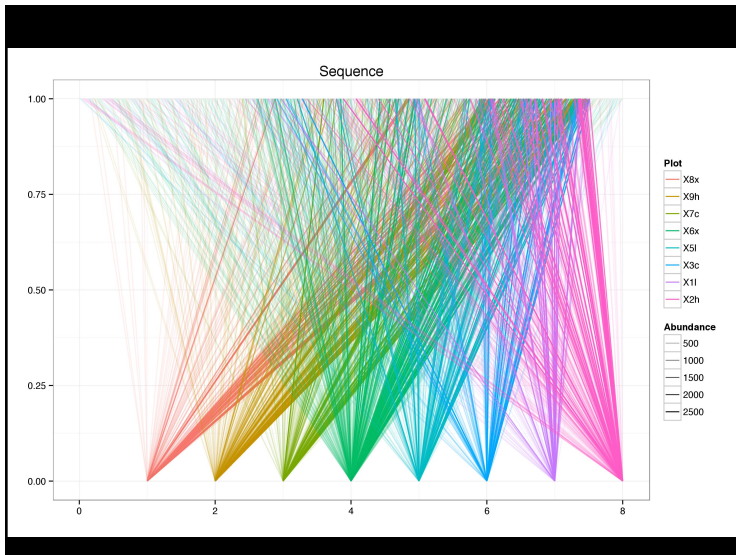
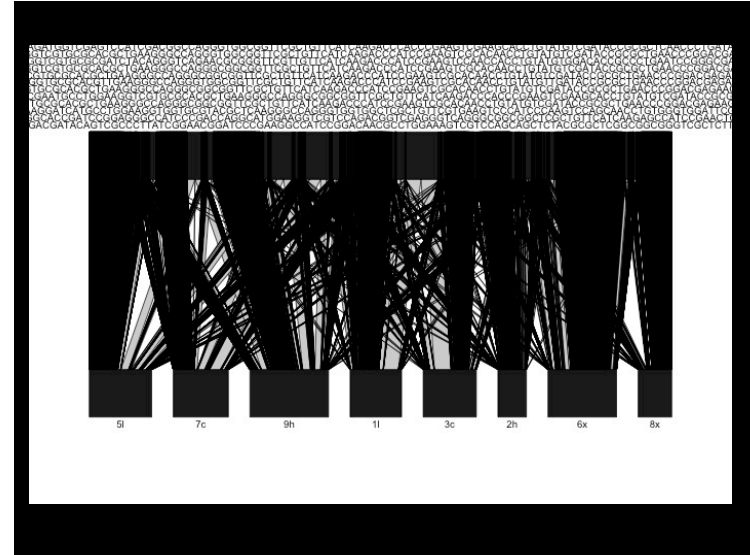
MD106 ATGCTTAGTAATCCCTACTTTAAGTCGGTTTTGTCGGTGATTGGCTTCGGAGGAATGGG
NEWC ATGCTTAGTAATCCCTACTTTAAGTCGGTTTTGTCGGTGATTGGCTTCGGAGGAATGGG
W501 ATGCTTAGTAATCCCTACTTTAAGTCGGTTTTGTCGGTGATTGGCTTCGGAGGAATGGG
MD199 ATGCTTAGTAATCCCTACTTTAAGTCGGTTTTGTCGGTGATTGGCTTCGGAGGAATGGG
C1674 ATGCTTAGTAATCCCTACTTTAAGTCGGTTTTGTCGGTGATTGGCTTCGGAGGAATGGG
SIM4 ATGCTTAGTAATCCCTACTTTAAGTCGGTTTTGTCGGTGATTGGCTTCGGAGGAATGGG

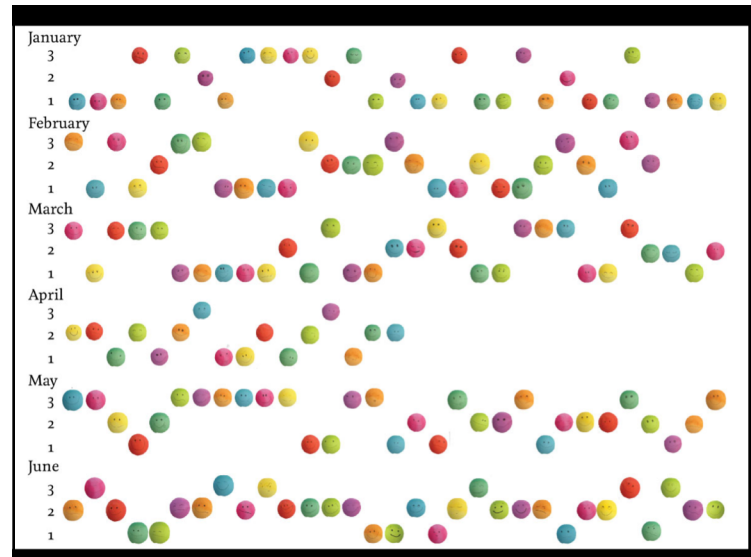
MD106 CTACGGCCTAATGGTGCTAACAGAGCCGAACGTCGACAAAATAGAGCGCATCAAAGCCT
NEWC CTACGGCCTAATGGTGCTAACAGAGCCGAACGTCGACAAAATAGAGCGCATCAAAGCCT
W501 CTACGGCCTAATGGTGCTAACAGAGCCGAACGTCGACAAAATAGAGCGCATCAAAGCCT
MD199 CTACGGCCTAATGGTGCTAACAGAGCCGAACGTCGACAAAATAGAGCGCATCAAAGCCT
C1674 CTACGGCCTAATGGTGCTAACAGAGCCGAACGTCGACAAAATAGAGCGCATCAAAGCCT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

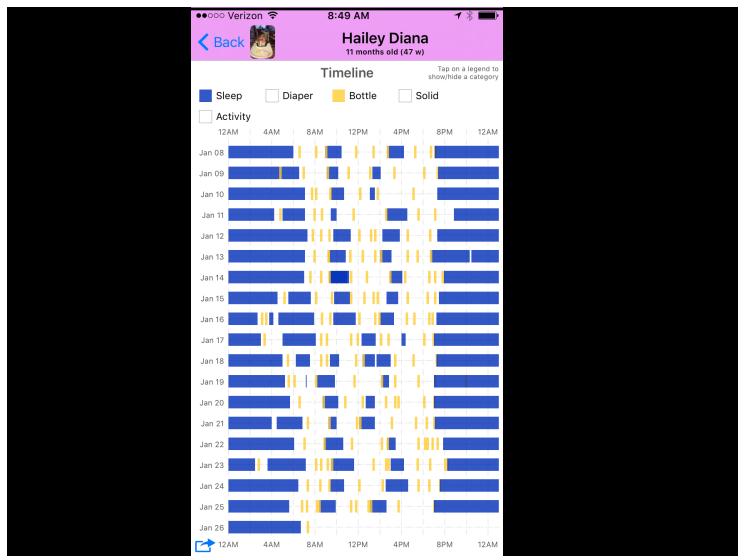
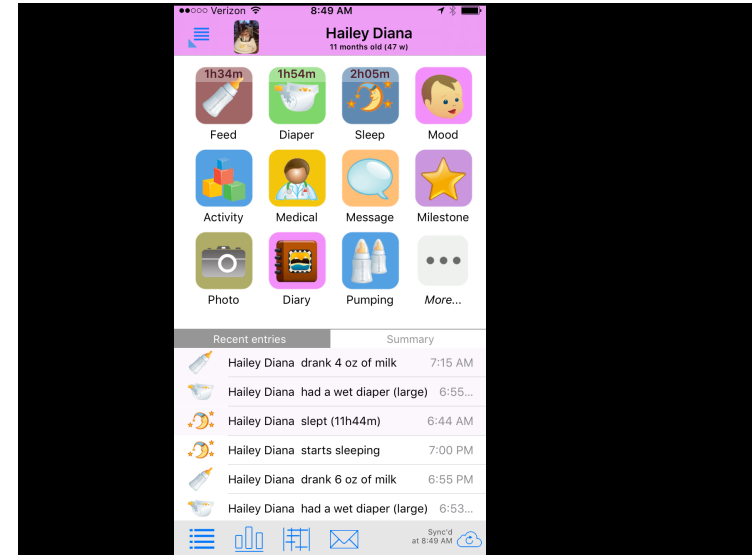
MD106 CCGTTTTCAAGTACCAAACTGAGTGGGATGAGCAGCGAAAGGCTCTGTTTATGAAGAAG
NEWC CCGTTTTCAAGTACCAAACTGAGTGGGATGAGCAGCGAAAGGCTCTGTTTATGAAGAAG
W501 CCGTTTTCAAGTACCAAACTGAGTGGGATGAGCAGCGAAAGGCTCTGTTTATGAAGAAG
MD199 CCGTTTTCAAGTACCAAACTGAGTGGGATGAGCAGCGAAAGGCTCTGTTTATGAAGAAG
C1674 CCGTTTTCAAGTACCAAACTGAGTGGGATGAGCAGCGAAAGGCTCTGTTTATGAAGAAG
SIM4 CCGTTTTCAAGTACCAAACTGAGTGGGATGAGCAGCGAAAGGCTCTGTTTATGAAGAAG

MD106 CTGCAGGAGGCTCCACCACCAAGTGCCCAATCTACAGGTCAGCGCCGAGAAATAG
NEWC CTGCAGGAGGCTCCACCACCAAGTGCCCAATCTACAGGTCATCGGCCGAGAAATAG
W501 CTGCAGGAGGCTCCACCACCAAGTGCCCAATCTACAGGTCATCGGCCGAGAAATAG
MD199 CTGCAGGAGGCTCCACCACCAAGTGCCCAATCTACAGGTCATCGGCCGAGAAATAG
C1674 CTGCAGGAGGCTCCACCACCAAGTGCCCAATCTACAGGTCATCGGCCGAGAAATAG
SIM4 CTGCAGGAGGCTCCACCACCAAGTGCCCAATCTACAGGTCATCGGCCGAGAAATAG
    
```

The image shows a spreadsheet titled 'OTU\_MAT.csv' with columns for 'Sequence' and multiple 'clusterDist' columns. The data consists of 33 rows of DNA sequences and their corresponding distances to various clusters. The sequences are listed in column A, and the distances are listed in columns B through K. The spreadsheet interface includes standard menu options like File, Edit, and View, and a status bar at the bottom.



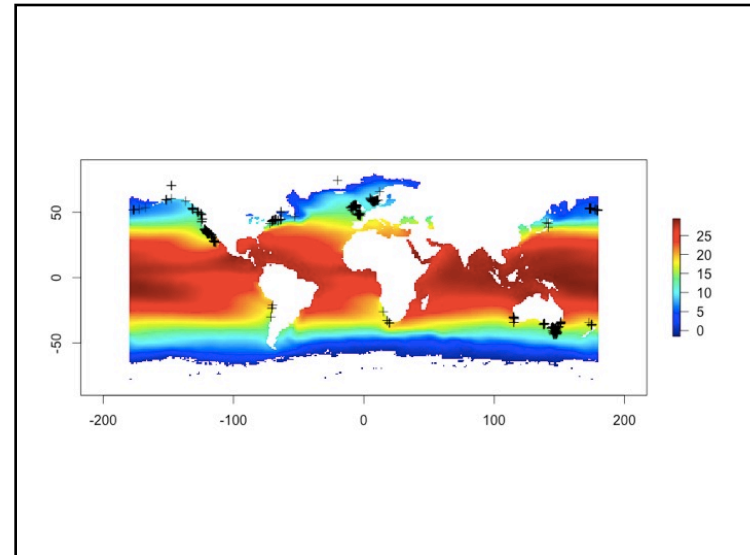
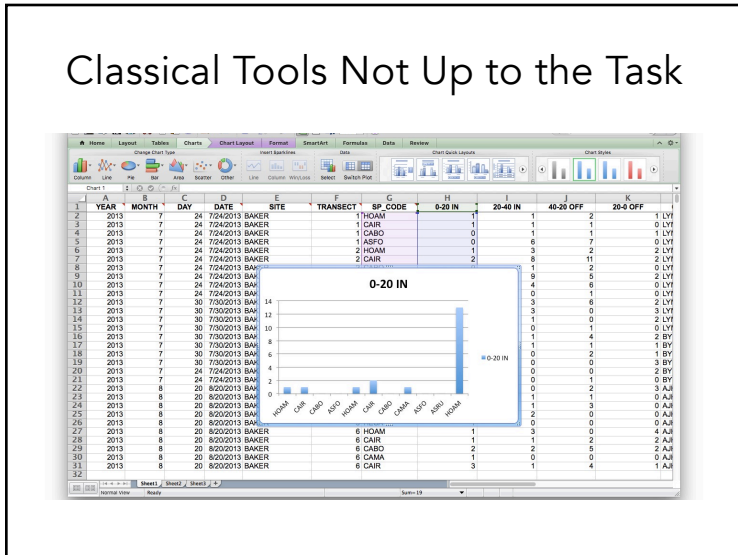
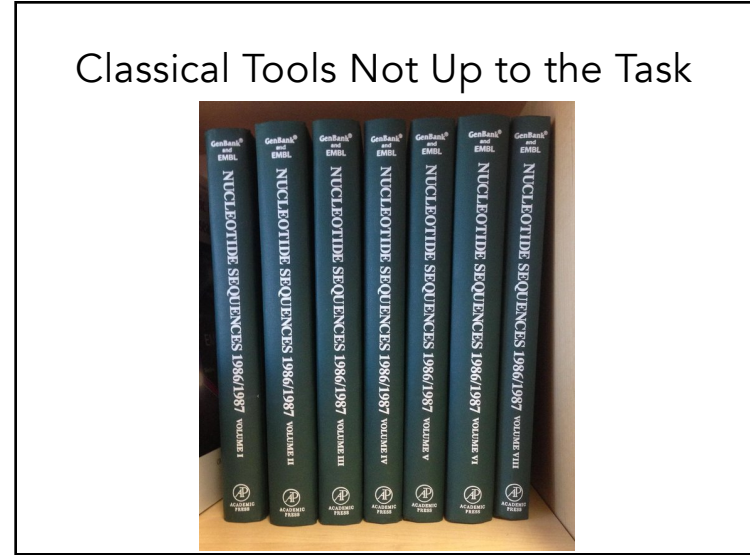




## Data Takes Many Forms

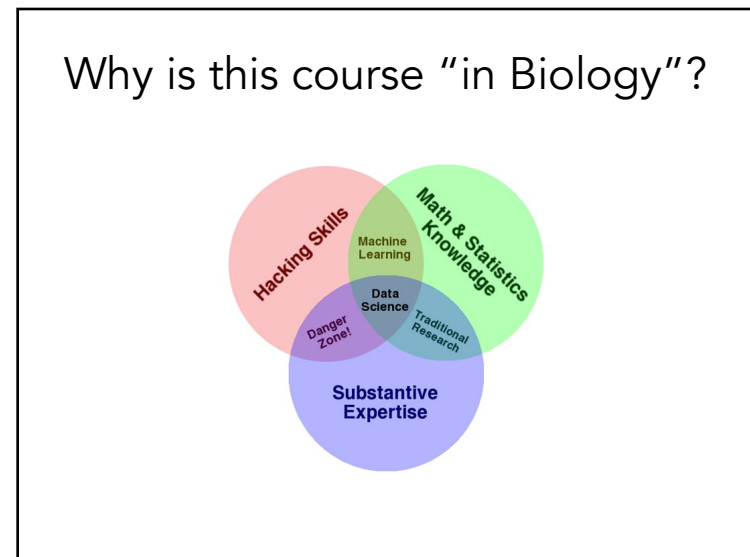
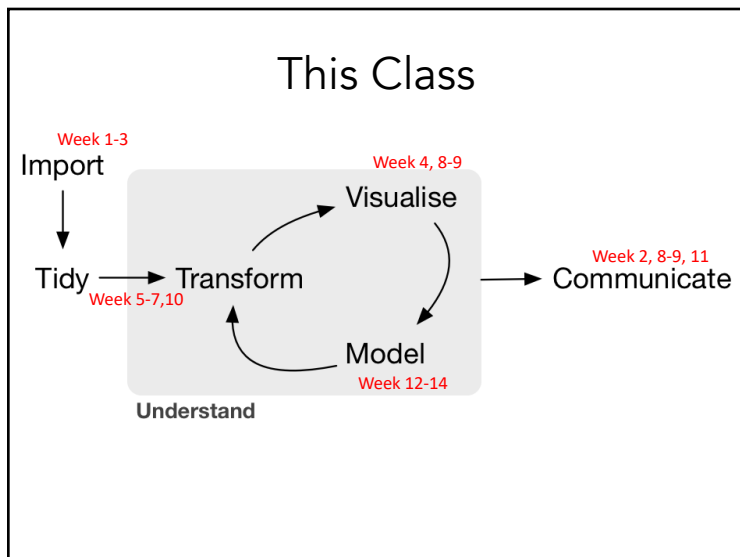
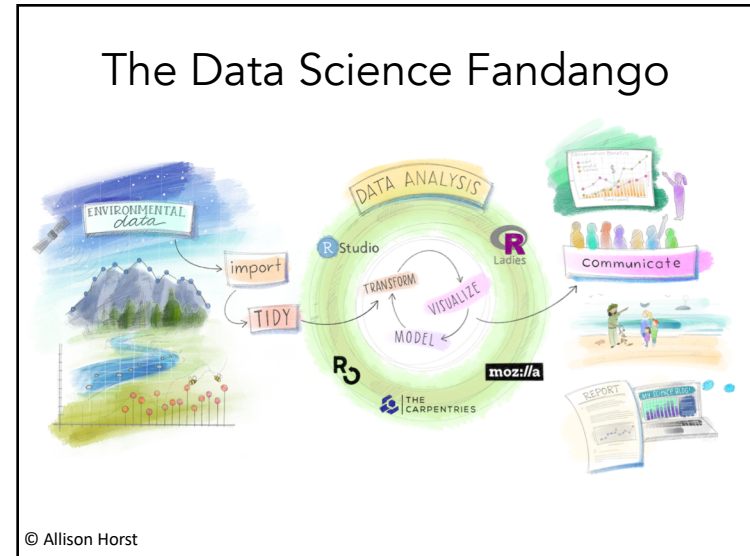
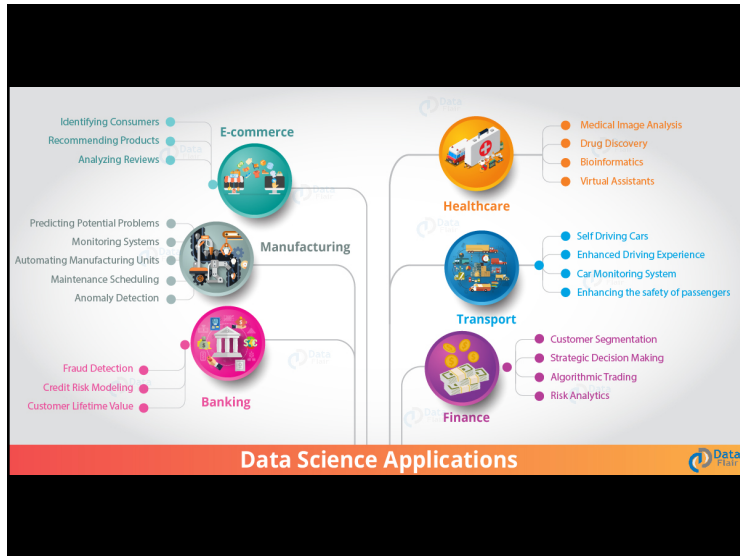
- Athletic performance
- Timeseries of polls
- Sequence Data
- Measurements of physical properties
- Maps (often with many layers) with information
- Timings of events
- Images
- Network descriptions
- Plain text











## Introduction to Data Science for Biology

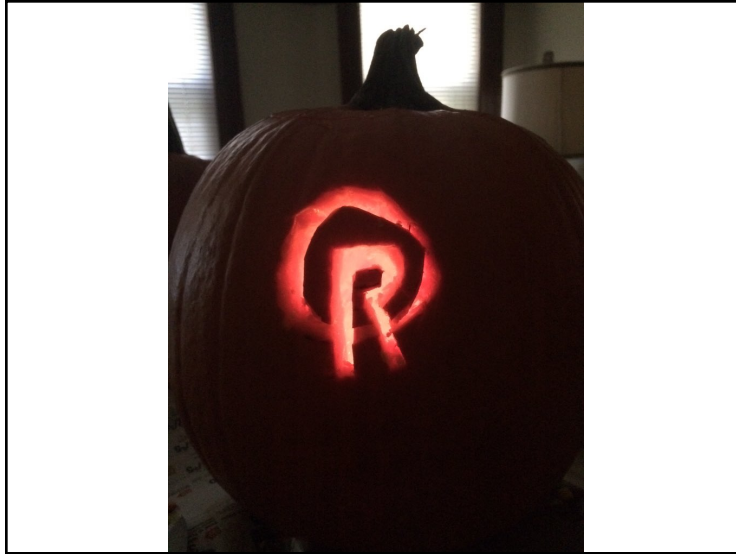
## Our Semester

Learn how to create efficient  
understandable datasets for  
biological research

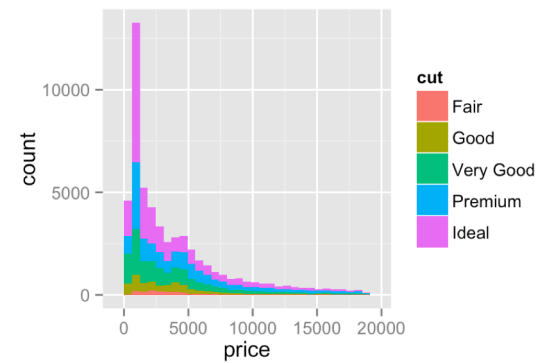
YEAR	MONTH	DAY	DATE	SITE	TRANSECT	SP_CODE	0-20 IN	20-40 IN	40-20 OFF
2013	7	24	7/24/2013	BAKER	1HOAM		1	1	2
2013	7	24	7/24/2013	BAKER	1CAIR		1	1	1
2013	7	24	7/24/2013	BAKER	1CABO		0	1	1
2013	7	24	7/24/2013	BAKER	1ASFO		0	6	7
2013	7	24	7/24/2013	BAKER	2HOAM		1	3	2
2013	7	24	7/24/2013	BAKER	2CAIR		2	8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

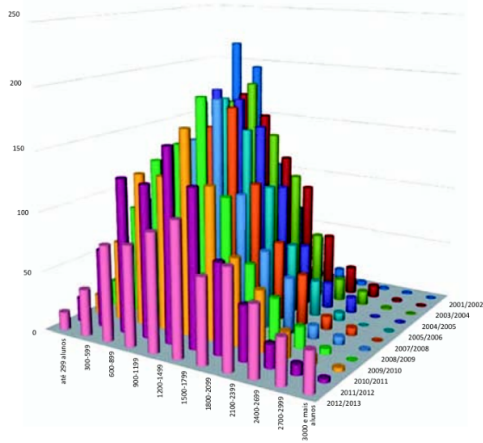




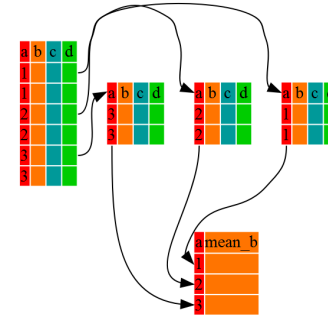
Build a vocabulary of visualization tools that enable students to see what their data means



# This is How I Know I Failed You



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

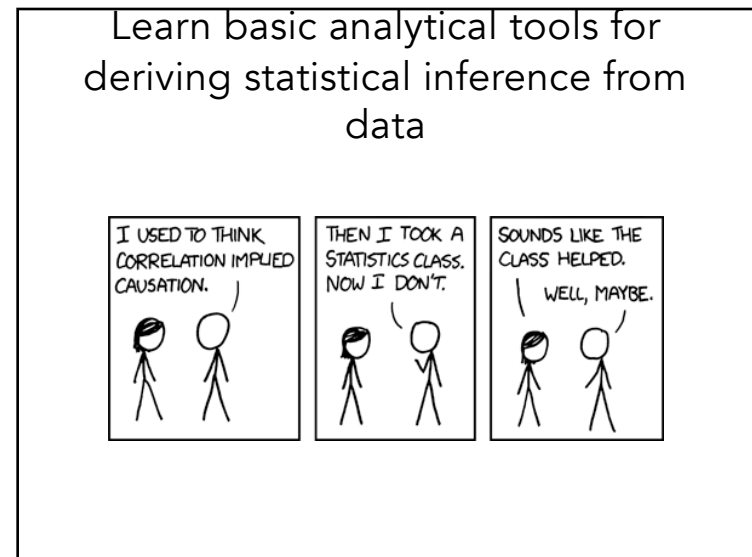
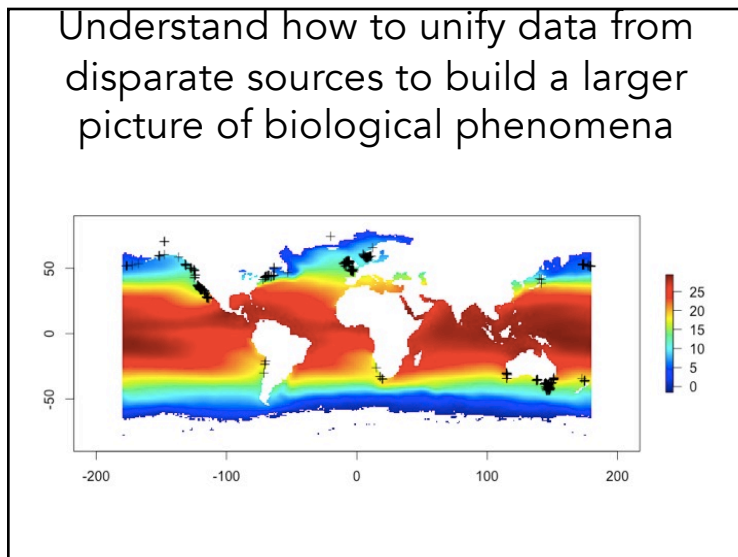
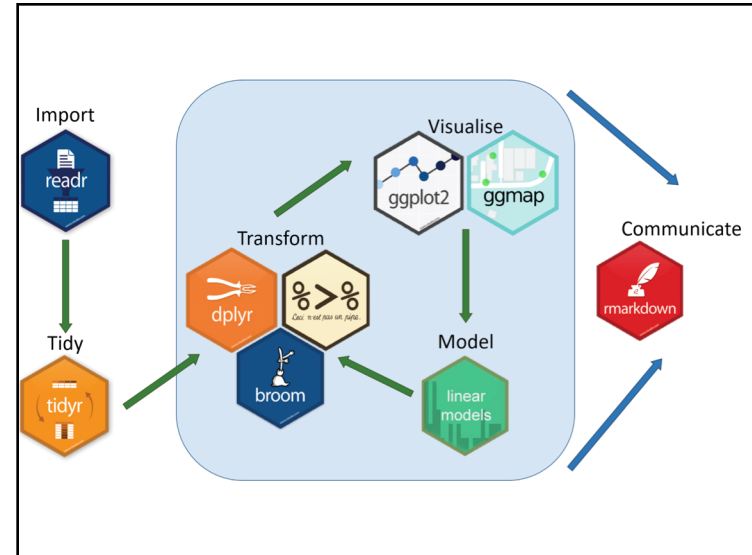
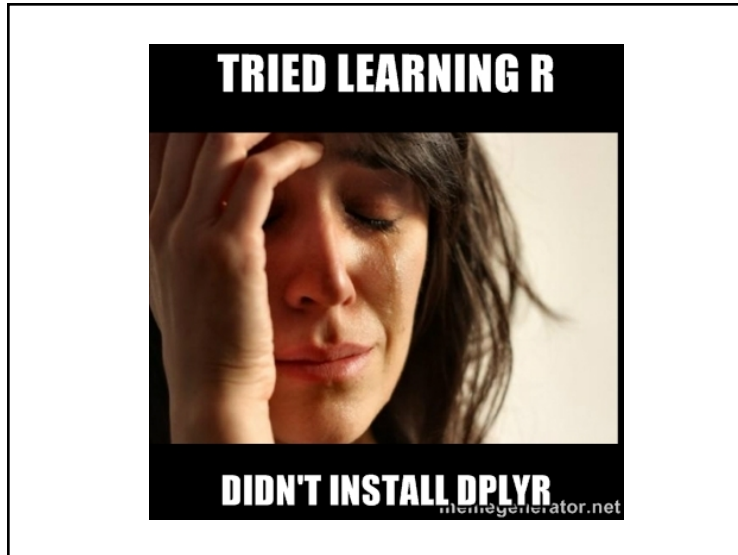


<http://swcarpentry.github.io/r-novice-gapminder/>

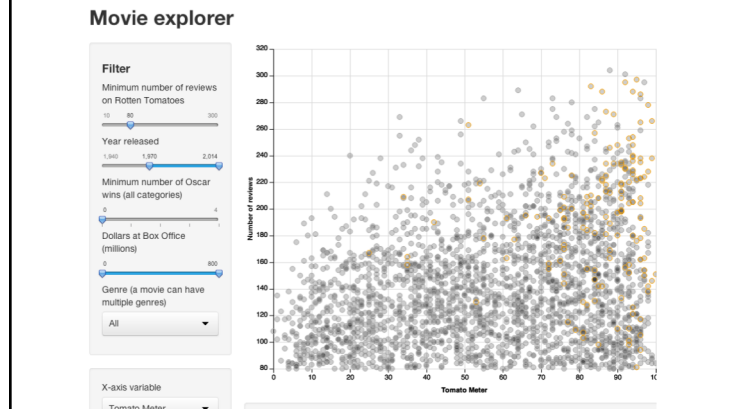


## Components





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



## This Class

## Course Web Page

Home Overview Schedule R Errors Final Project Syllabus Resources Local Meetups

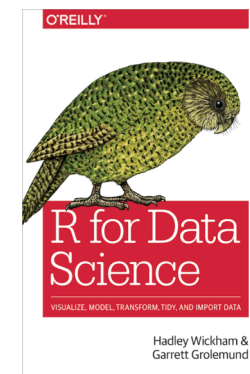
**Biol 355/356: Intro to Data Science for Biology**

Elisa assay of DNase

Instructor: Jarrett Byrnes, PhD.  
 Email: [jarrett.byrnes@umb.edu](mailto:jarrett.byrnes@umb.edu)  
 TA: Isaac Rosenthal  
 Email: [isaac.rosenthal001@umb.edu](mailto:isaac.rosenthal001@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



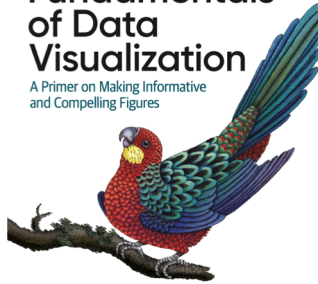
<http://r4ds.had.co.nz/>

## Additional Resources

**O'REILLY**


### Fundamentals of Data Visualization

A Primer on Making Informative and Compelling Figures



Claus O. Wilke

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

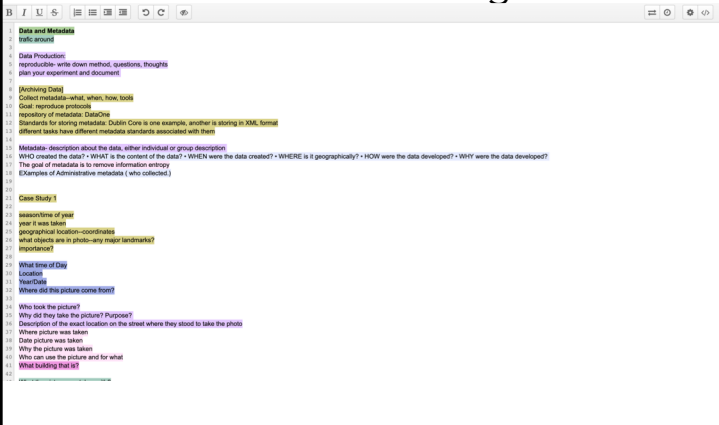


## DATA CARPENTRY

BUILDING COMMUNITIES TEACHING UNIVERSAL DATA LITERACY

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

## Etherpad: How we will Communicate During Class



**Data and Metadata**

**Basic protocol**

**Data Production**  
reproducible: write down method, questions, thoughts  
plus your experiment and document

**Archiving Data**  
Collect metadata—what, when, how, tools  
Goal: reproducible protocol  
repository of metadata: DataONE  
Standards for storing metadata: Dublin Core is one example, another is storing in XML format  
different tasks have different metadata standards associated with them

**Metadata: description about the data, either individual or group description**  
WHO created the data? - WHO? is the content of the data? - WHO? were the data created? - WHERE is it geographically? - HOW were the data developed? - WHY were the data developed?  
The goal of metadata is to remove information entropy  
(Examples of Administrative metadata (who collected))

**Case Study 1**

season/time of year  
near it was taken  
geographical location—coordinates  
what objects are in photo—any major landmarks?  
timestamp

**What time of Day**  
Location  
Year/Date  
Where did this picture come from?

**Who took the picture?**  
Who did they take the picture? Purpose?  
Description of the exact location on the street where they stood to take the photo  
Where picture was taken  
Date picture was taken  
Why the picture was taken  
Who can use the picture and for what  
What building that is?

## Lab

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





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?