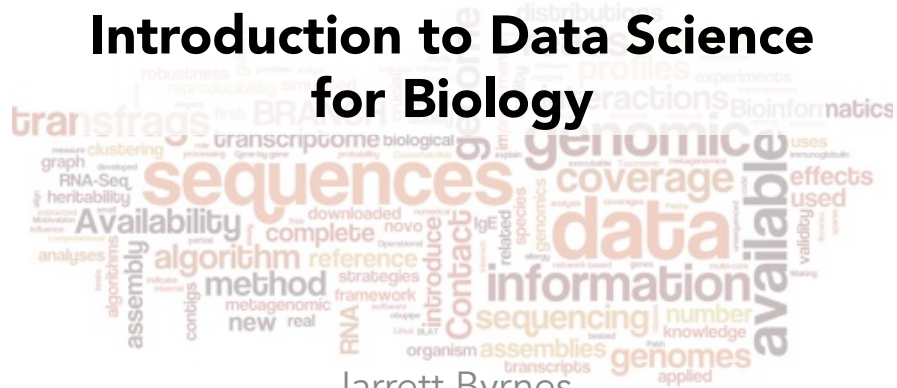


Introduction to Data Science for Biology




Jarrett Byrnes
UMass Boston
Spring 2018

We Are Awash in Data

FiveThirtyEight ESPN

Politics Sports Science & Health Economics Culture

First episode: Elections podcast



THE LATEST

- 7:00 PM You'd Have To Be Pretty Dumb To Fix A Tennis Match This Week
- 5:53 PM Elections Podcast: One Week To Iowa
- 4:43 PM What Do Anti-Abortion Demonstrators Want (Besides An End To Abortion)?
- 4:39 PM The Republican Party May Be Failing
- 3:34 PM Spurs-Warriors Is The Best Defense-Offense Clash In NBA History
- 12:56 PM Tom Brady Couldn't Take The Pressure

INTERACTIVES

Primary Forecasts

UPDATED 4 HOURS AGO

Chance of winning Iowa

	POLLS-ONLY FORECAST	POLLS-PLUS FORECAST
Clinton	67%	80%
Sanders	33%	20%
O'Malley	<1%	<1%

SEE FORECASTS FOR MORE PRIMARIES

NFL Predictions

UPDATED 1 DAY AGO

Chance of winning the Super Bowl

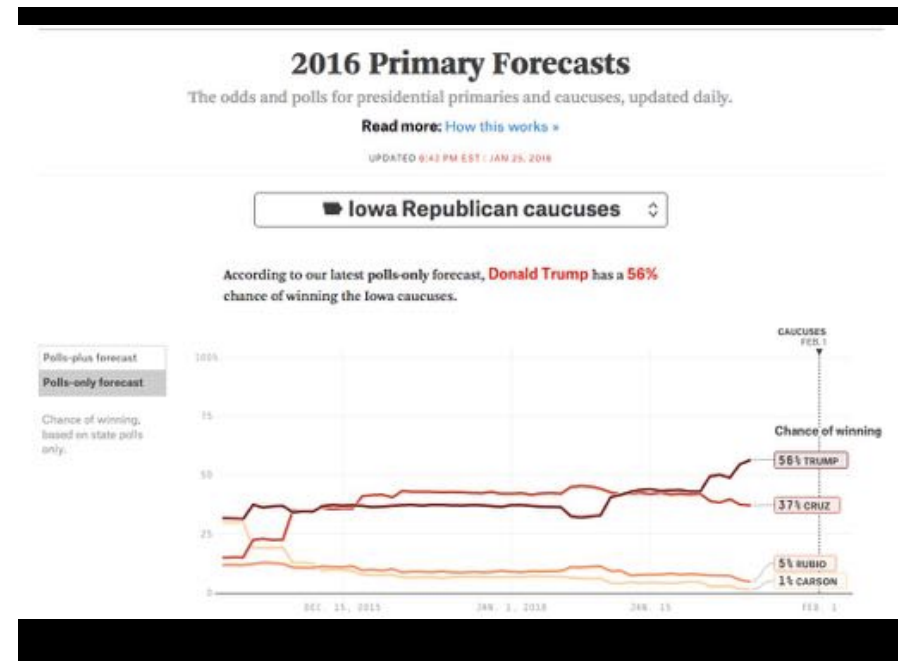
	WEEK 21
Carolina	59%
Denver	41%

2016 ELECTION

The Republican Party May Be Failing

By NATE SILVER

TENNIS





Brady's greatness hasn't expired, but Pats call it forever

PATRIOTS SCOREBOARD

Scoreboard

Sun, Jan 24

AFC Championship

Final	1	2	3	4	Tot
New England	6	3	3	6	18
Denver	7	10	0	3	20

Preview | Box | Gameview | Recap

Leader	New England	Denver
Passing	T. Brady 310	P. Manning 176
Rushing	T. Brady 13	C. Anderson 72
Receiving	R. Gronkowski 144	E. Sanders 62

Leaders

Pass Yds:	T. Brady	4770
Pass TDs:	T. Brady	36
Rush Yds:	L. Blount	703
Rush TDs:	L. Blount	6
Rec Yds:	R. Gronkowski	1176
Rec TDs:	R. Gronkowski	11
Int:	L. Ryan	4
Sacks:	C. Jones	12.5

AFC East Division

Team	W	L	T	Pct
xy-New England	12	4	0	.750
New York	10	6	0	.625
Buffalo	8	8	0	.500
Miami	6	10	0	.375

Full Standings



25 Best Jobs For Work-Life Balance (2015)

Classroom Team | October 20, 2015

Maintaining a healthy work-life balance can be tough in today's work environment, but some jobs

1. Data Scientist

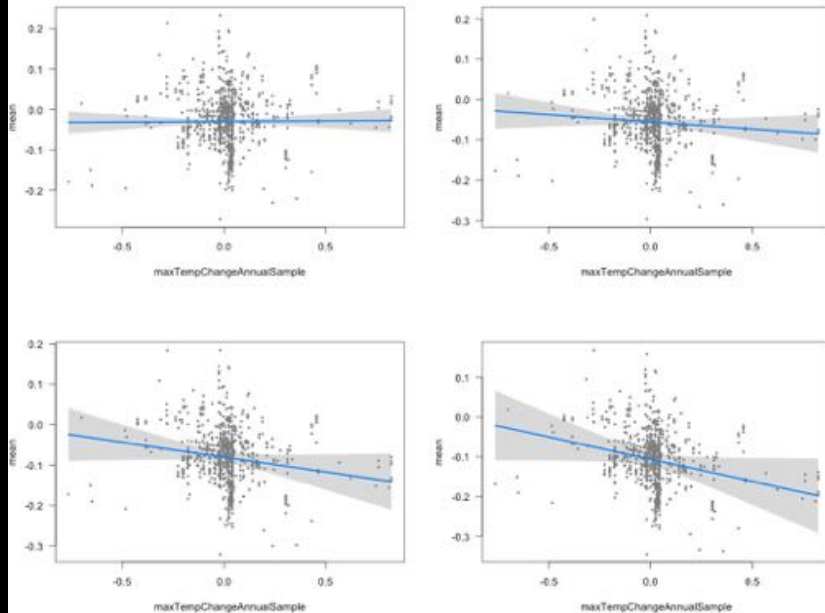
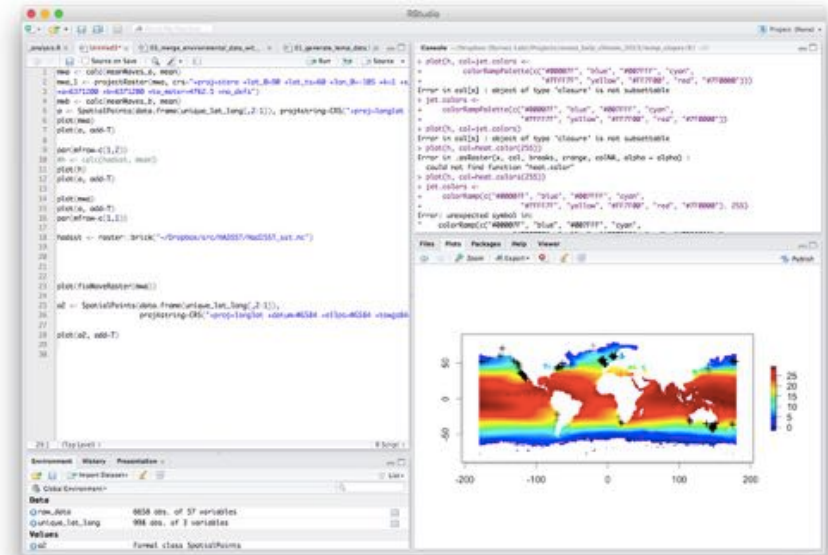
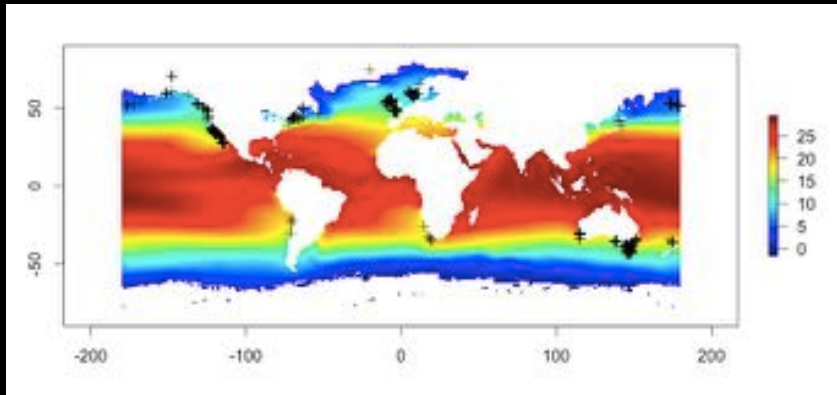
- Work-Life Balance Rating: 4.2
- Salary: \$114,808
- Number of Job Openings: 1,315

1. Data Scientist

- Work-Life Balance Rating: 4.2
- Salary: \$114,808
- Number of Job Openings: 1,315



YEAR	MONTH	DAY	DATE	SITE	TRANSECT	SP CODE	H	0-20 IN	20-40 IN	40-60 OFF	20-4 OFF
2013	7	24	7/24/2013	BAKER	1	HOAM	1	1	1	2	1 LYI
2013	7	24	7/24/2013	BAKER	1	CABR	1	1	1	1	0 LYI
2013	7	24	7/24/2013	BAKER	1	CABO	0	0	1	1	1 LYI
2013	7	24	7/24/2013	BAKER	1	ASFO	0	0	6	7	0 LYI
2013	7	24	7/24/2013	BAKER	2	HOAM	1	3	2	2	2 LYI
2013	7	24	7/24/2013	BAKER	2	CABR	2	8	11	11	2 LYI
2013	7	24	7/24/2013	BAKER	2	CABO	0	1	2	0	0 LYI
2013	7	24	7/24/2013	BAKER	2	CAMA	1	9	5	2	2 LYI
2013	7	24	7/24/2013	BAKER	2	ASFO	0	4	6	0	0 LYI
2013	7	24	7/24/2013	BAKER	3	HOAM	13	3	6	2	2 LYI
2013	7	30	7/30/2013	BAKER	3	CABR	3	3	0	3	3 LYI
2013	7	30	7/30/2013	BAKER	3	CABO	6	1	0	2	2 LYI
2013	7	30	7/30/2013	BAKER	3	HESA	1	0	1	0	0 LYI
2013	7	30	7/30/2013	BAKER	4	HOAM	3	1	4	2	2 BY
2013	7	30	7/30/2013	BAKER	4	CABR	1	1	1	1	1 BY
2013	7	30	7/30/2013	BAKER	4	CABO	1	0	2	1	1 BY
2013	7	30	7/30/2013	BAKER	4	BADE	0	0	0	0	3 BY
2013	7	24	7/24/2013	BAKER	4	HESA	0	0	0	0	2 BY
2013	7	24	7/24/2013	BAKER	4	PRGURUS	0	0	0	1	0 BY
2013	8	20	8/20/2013	BAKER	5	HOAM	1	0	2	1	3 A/J
2013	8	20	8/20/2013	BAKER	5	CABR	0	1	1	0	0 A/J
2013	8	20	8/20/2013	BAKER	5	CABO	1	1	3	0	0 A/J
2013	8	20	8/20/2013	BAKER	5	ASFO	0	2	0	0	0 A/J
2013	8	20	8/20/2013	BAKER	5	HESA	1	0	0	0	0 A/J
2013	8	20	8/20/2013	BAKER	6	HOAM	1	3	0	4	4 A/J
2013	8	20	8/20/2013	BAKER	6	CABR	1	1	2	2	2 A/J
2013	8	20	8/20/2013	BAKER	6	CABO	2	1	2	2	2 A/J
2013	8	20	8/20/2013	BAKER	6	CAMA	1	0	0	0	0 A/J
2013	8	20	8/20/2013	BAKER	6	CABR	3	1	4	1	1 A/J



SYNTENIC ASSEMBLIES FOR CG15386

```

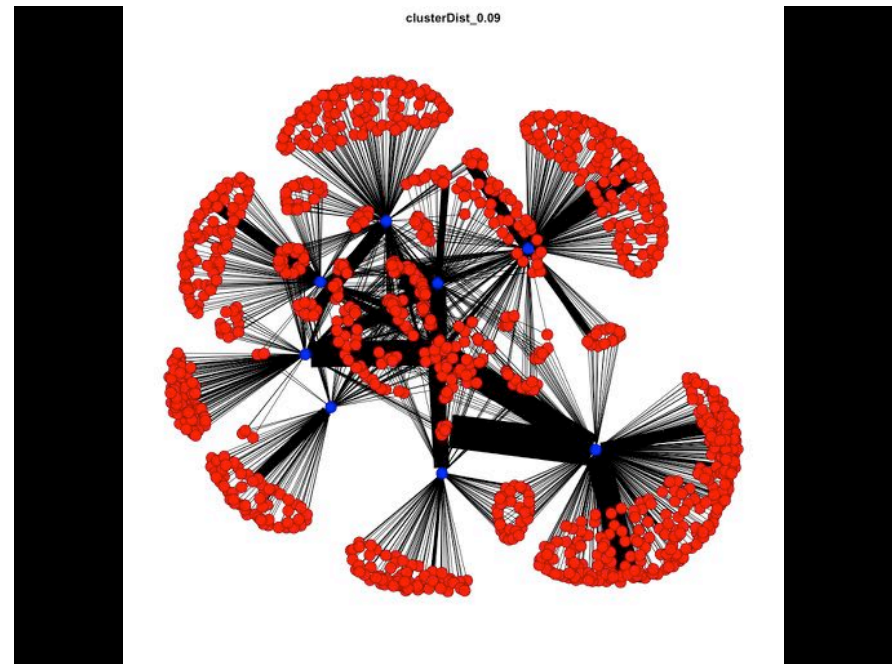
MD106 ATGCTTAGTAATCCCTACTTTAAAGTCCGTTTGTGGCTGATTGGCTTCGGAGGAATGGG
NEWC ATGCTTAGTAATCCCTACTTTAAAGTCCGTTTGTGGCTGATTGGCTTCGGAGGAATGGG
W501 ATGCTTAGTAATCCCTACTTTAAAGTCCGTTTGTGGCTGATTGGCTTCGGAGGAATGGG
MD199 ATGCTTAGTAATCCCTACTTTAAAGTCCGTTTGTGGCTGATTGGCTTCGGAGGAATGGG
C1674 ATGCTTAGTAATCCCTACTTTAAAGTCCGTTTGTGGCTGATTGGCTTCGGAGGAATGGG
SIM4 ATGCTTAGTAATCCCTACTTTAAAGTCCGTTTGTGGCTGATTGGCTTCGGAGGAATGGG

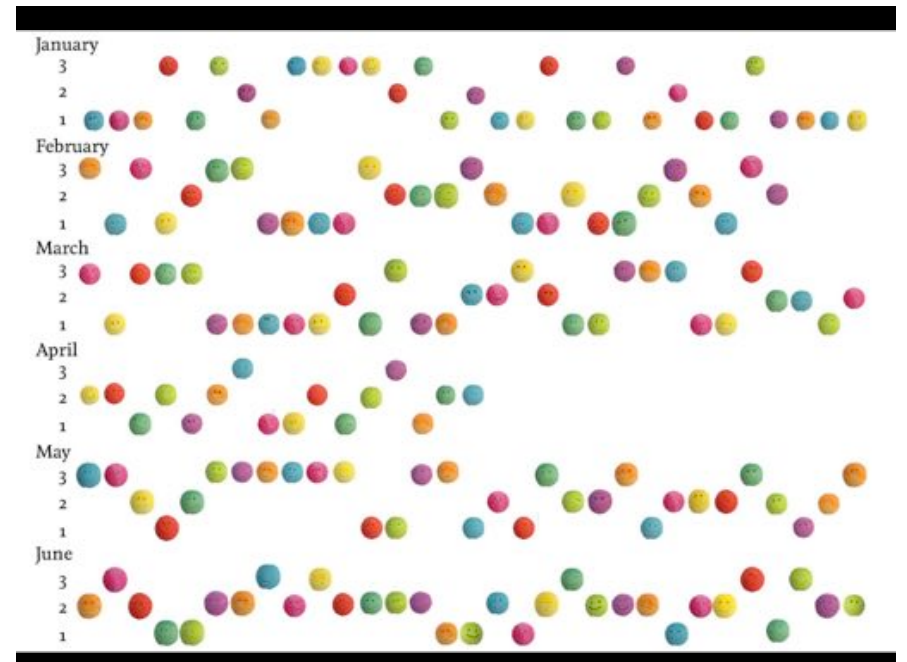
MD106 CTACGGCCCTAATGGTGCTAACGAGCGGAACGTCGACAAAATAGAGCGCATCAAAGCCT
NEWC CTACGGCCCTAATGGTGCTAACGAGCGGAACGTCGACAAAATAGAGCGCATCAAAGCCT
W501 CTACGGCCCTAATGGTGCTAACGAGCGGAACGTCGACAAAATAGAGCGCATCAAAGCCT
MD199 CTACGGCCCTAATGGTGCTAACGAGCGGAACGTCGACAAAATAGAGCGCATCAAAGCCT
C1674 CTACGGCCCTAATGGTGCTAACGAGCGGAACGTCGACAAAATAGAGCGCATCAAAGCCT
SIM4 CTACGGCCCTAATGGTGCTAACGAGCGGAACGTCGACAAAATAGAGCGCATCAAAGCCT

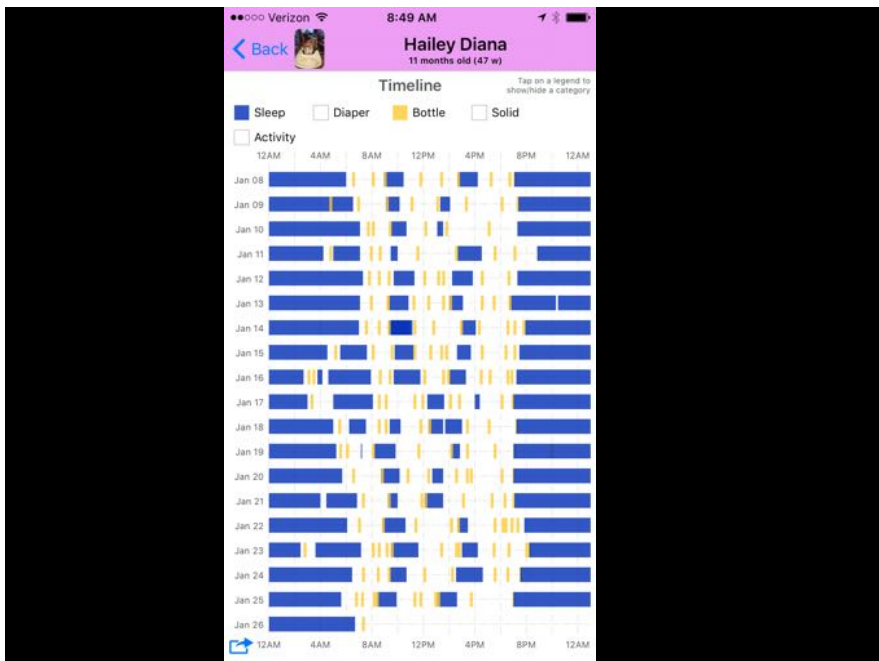
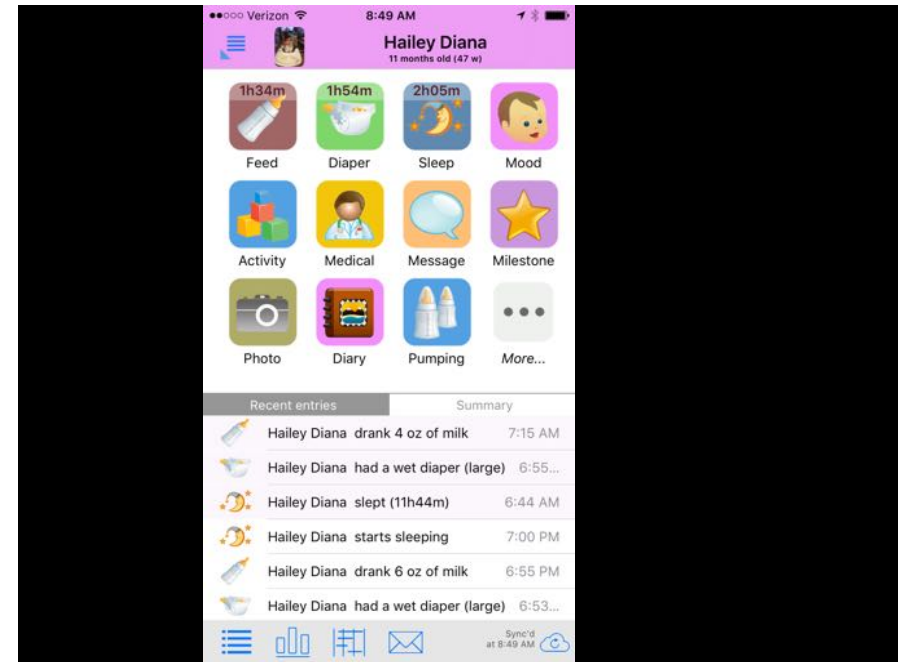
MD106 CCGTTTCAAGTACCAAACCTGAGTGGCGGATGAGCAGCGAAAGGCTCTGTTTATGAAGAAG
NEWC CCGTTTCAAGTACCAAACCTGAGTGGCGGATGAGCAGCGAAAGGCTCTGTTTATGAAGAAG
W501 CCGTTTCAAGTACCAAACCTGAGTGGCGGATGAGCAGCGAAAGGCTCTGTTTATGAAGAAG
MD199 CCGTTTCAAGTACCAAACCTGAGTGGCGGATGAGCAGCGAAAGGCTCTGTTTATGAAGAAG
C1674 CCGTTTCAAGTACCAAACCTGAGTGGCGGATGAGCAGCGAAAGGCTCTGTTTATGAAGAAG
SIM4 CCGTTTCAAGTACCAAACCTGAGTGGCGGATGAGCAGCGAAAGGCTCTGTTTATGAAGAAG

MD106 CTGCAGGAGGCGTCCACCACCACTGCCCCAATCTACAGGTCACTCGGCCGAGAAATAG
NEWC CTGCAGGAGGCGTCCACCACCACTGCCCCAATCTACAGGTCACTCGGCCGAGAAATAG
W501 CTGCAGGAGGCGTCCACCACCACTGCCCCAATCTACAGGTCACTCGGCCGAGAAATAG
MD199 CTGCAGGAGGCGTCCACCACCACTGCCCCAATCTACAGGTCACTCGGCCGAGAAATAG
C1674 CTGCAGGAGGCGTCCACCACCACTGCCCCAATCTACAGGTCACTCGGCCGAGAAATAG
SIM4 CTGCAGGAGGCGTCCACCACCACTGCCCCAATCTACAGGTCACTCGGCCGAGAAATAG

```





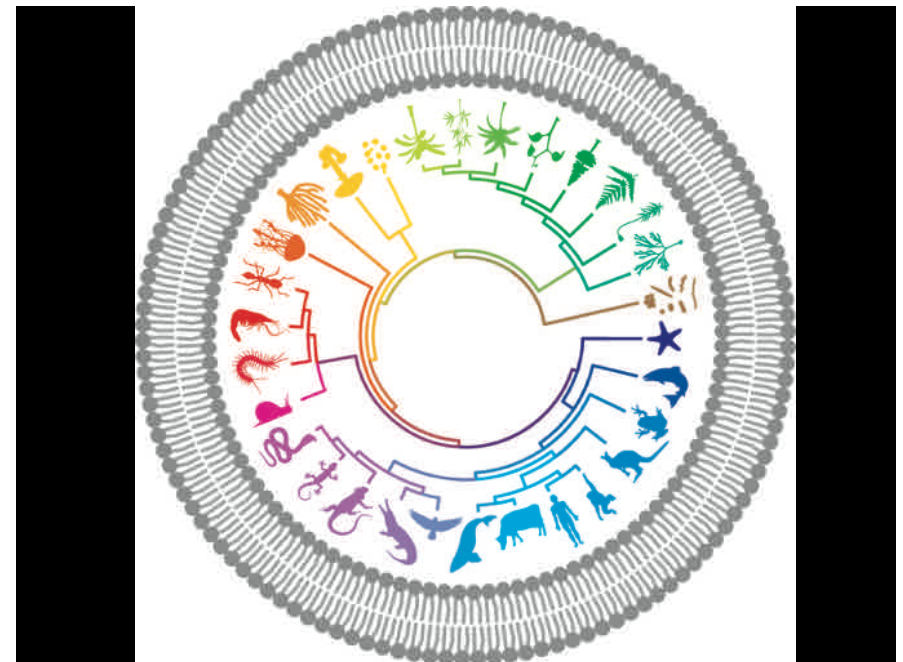


We Are Awash in Data

Data Takes Many Forms

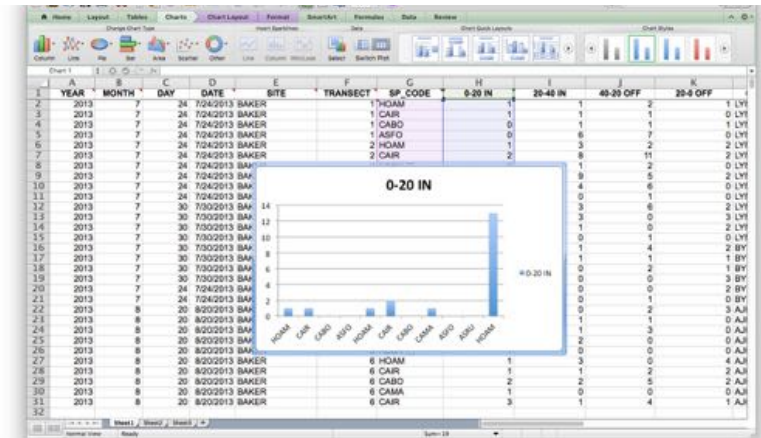
- Athletic performance
- Timeseries of polls
- Sequence Data
- Measurements of physical properties
- Maps (often with many layers) with information
- Timings of events
- Images
- Network descriptions
- Plain text

Data is at the Center of Biology





Classical Tools Not Up to the Task

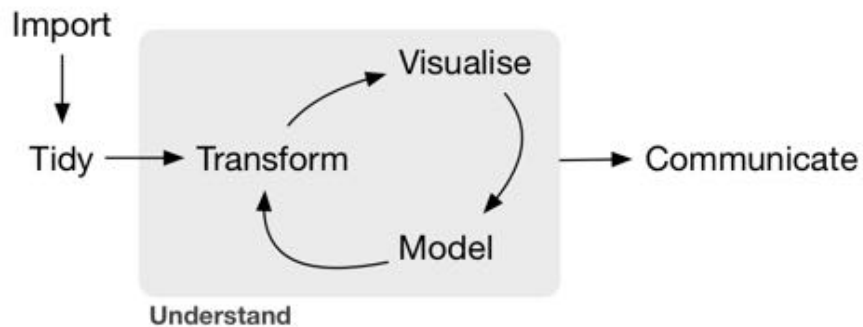


Classical Tools Not Up to the Task



Introduction to Data Science for Biology

Who Are You?



Our Semester

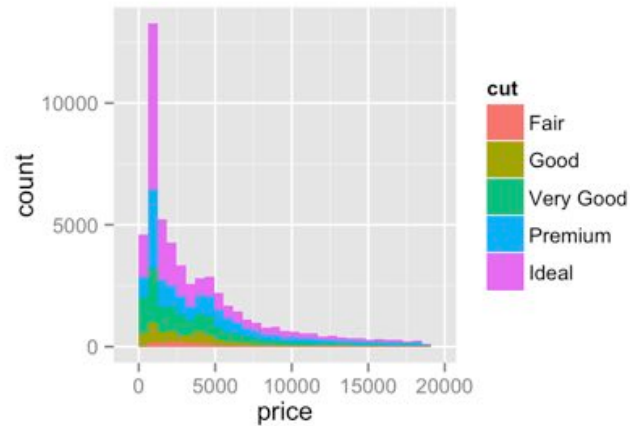
Learn how to create efficient understandable datasets for biological research

YEAR	MONTH	DAY	DATE	SITE	TRANSECT	SP_CODE	0-20 IN	20-40 IN	40-20 OFF
2013	7	24	247/24/2013	BAKER	1	HOAM	1	1	2
2013	7	24	247/24/2013	BAKER	1	CAIR	1	1	1
2013	7	24	247/24/2013	BAKER	1	CABO	0	1	1
2013	7	24	247/24/2013	BAKER	1	ASFO	0	6	7
2013	7	24	247/24/2013	BAKER	2	HOAM	1	3	2
2013	7	24	247/24/2013	BAKER	2	CAIR	2	8	11
2013	7	24	247/24/2013	BAKER	2	CABO	0	1	2
2013	7	24	247/24/2013	BAKER	2	CAMA	1	9	5
2013	7	24	247/24/2013	BAKER	2	ASFO	0	4	6
2013	7	24	247/24/2013	BAKER	2	ASRU	0	0	1

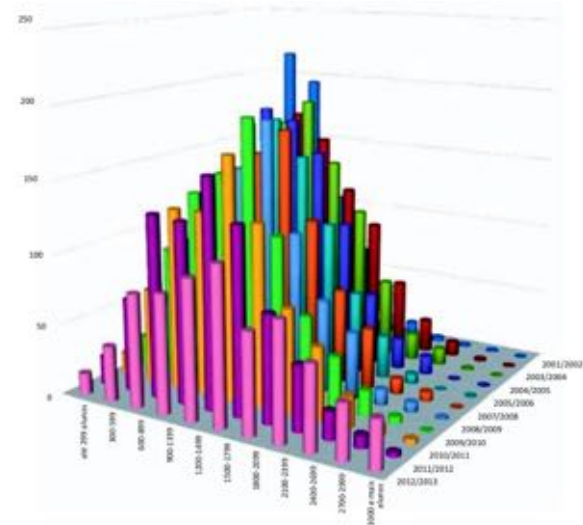
Learn common programming
language(s) associated with data
science



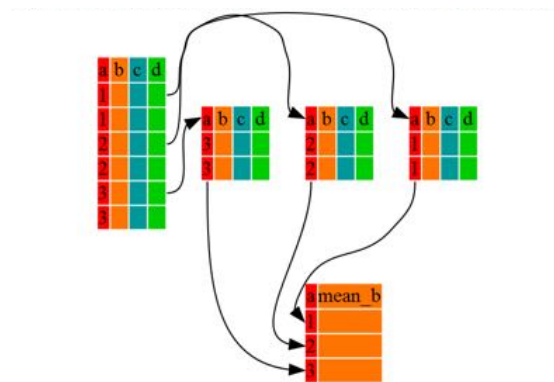
Build a vocabulary of visualization tools that enable students to see what their data means



This is How I Know I Failed You

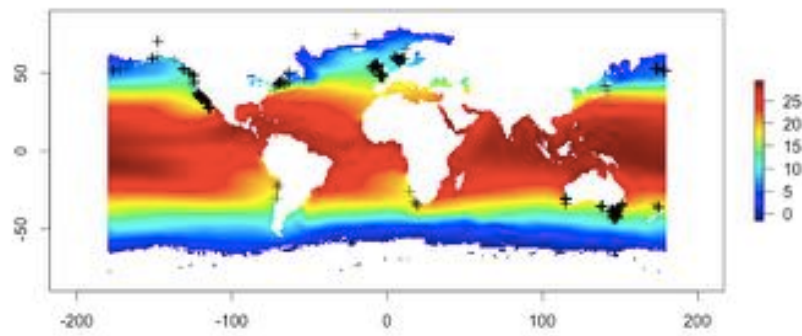


Develop an understanding of how to manipulate data for the purposes of seeing useful patterns





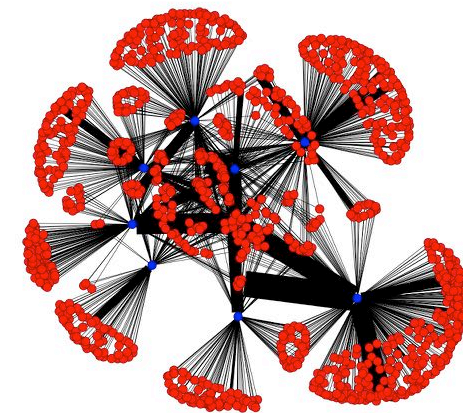
Understand how to unify data from disparate sources to build a larger picture of biological phenomena



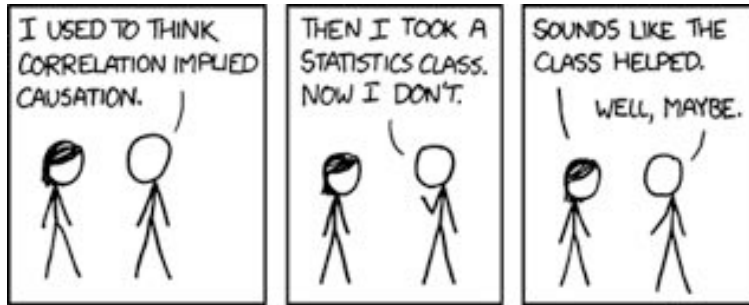
Components



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



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

10 11 9 1
4 8 5 7
6 2 3

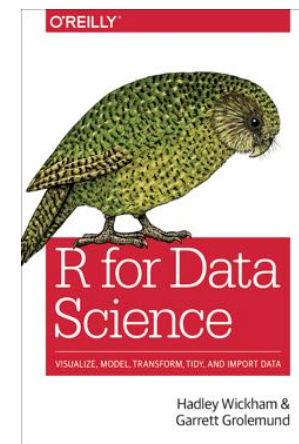
DNA Optical Density

Protein Concentration

Instructor: Jarrett Byrnes, PhD.
Email: jarrett.byrnes@umb.edu
TA: Isaac Rosenthal
Email: 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/>



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

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

Lab

- Coding!
- TA: Isaac Rosenthal
- Guided examples and then challenge problems

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 two weeks
 - What data will you be using?
 - What question do you want to answer?

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

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)