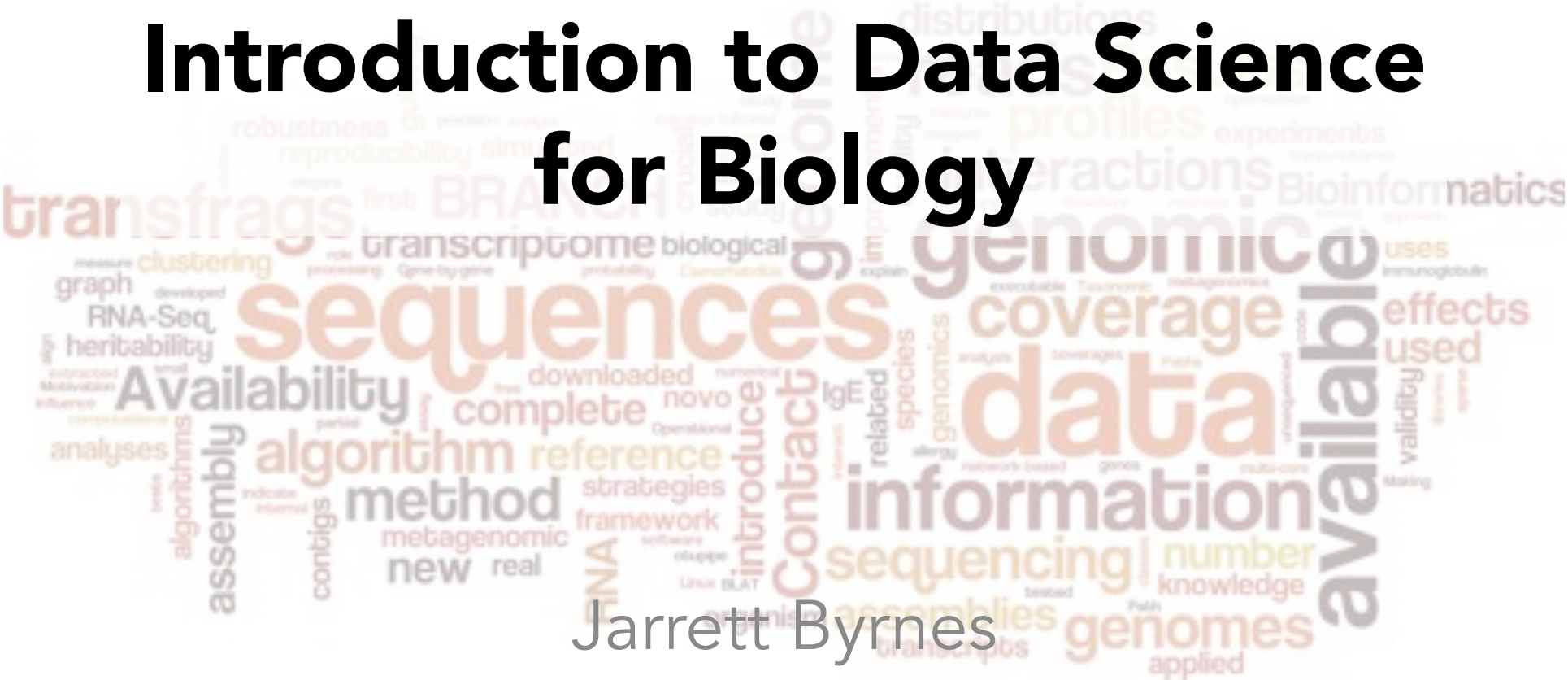
Introduction to Data Science for Biology



Jarrett Byrnes

UMass Boston

Spring 2016

We Are Awash in Data

FiveThirtyEight

Politics Sports Science & Health Economics Culture

First episode: Elections podcast

2016 ELECTION

The Republican Party May Be Failing

By NATE SILVER

TENNIS

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

WEEK 21

UPDATED 1 DAY AGO

Chance of winning the Super Bowl

Carolina	59%
Denver	41%

2016 Primary Forecasts

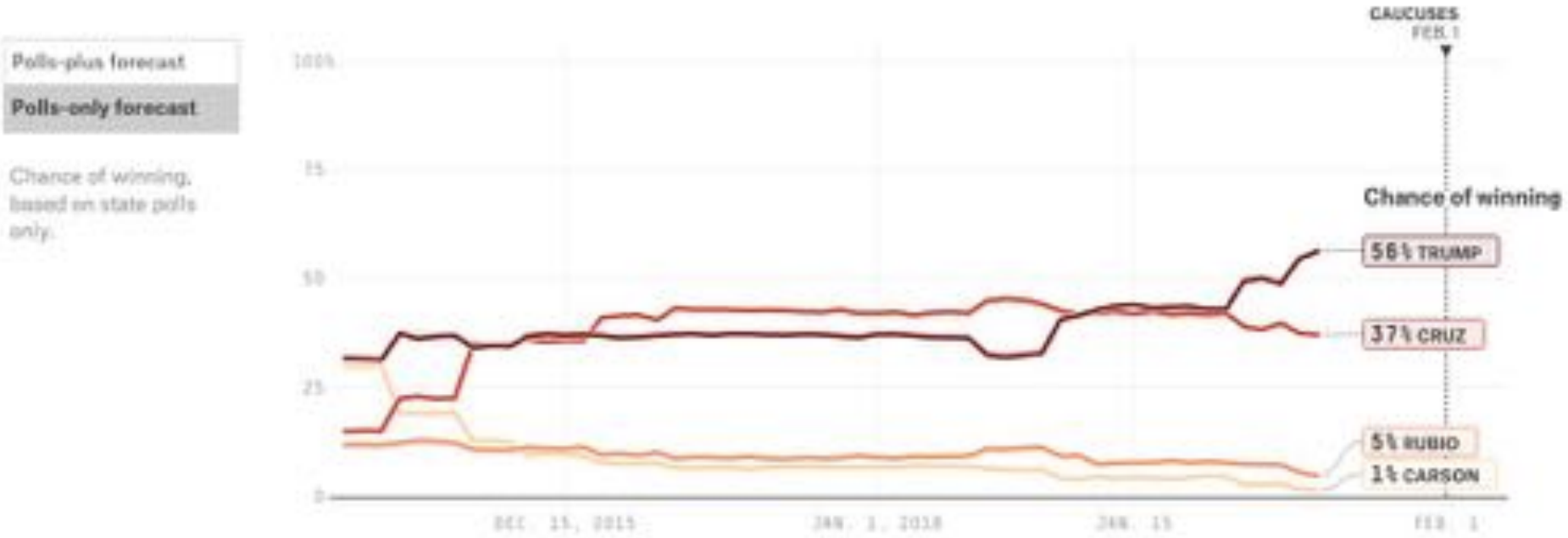
The odds and polls for presidential primaries and caucuses, updated daily.

Read more: How this works »

UPDATED 6:43 PM EST | JAN 25, 2016

Iowa Republican caucuses

According to our latest polls-only forecast, Donald Trump has a 56% chance of winning the Iowa caucuses.



HOME NEWS WEATHER SPORTS BUSINESS ENTERTAINM

Brady's greatness hasn't expired, but Pats ca 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)

Glassdoor Team | October 20, 2015

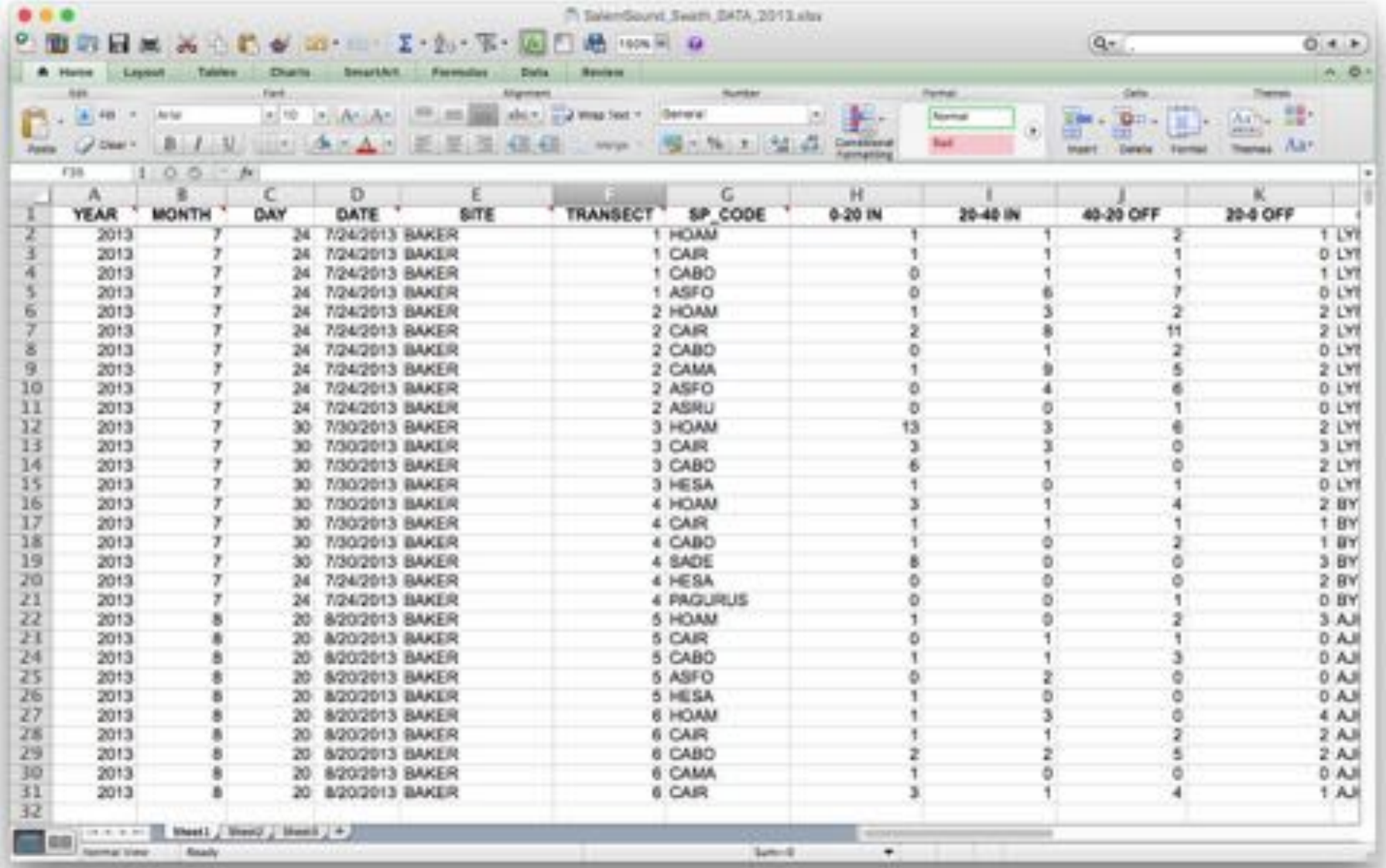
Maintaining a healthy work-life balance can be tough in today's work environment, but some jobs

1. Data Scientist

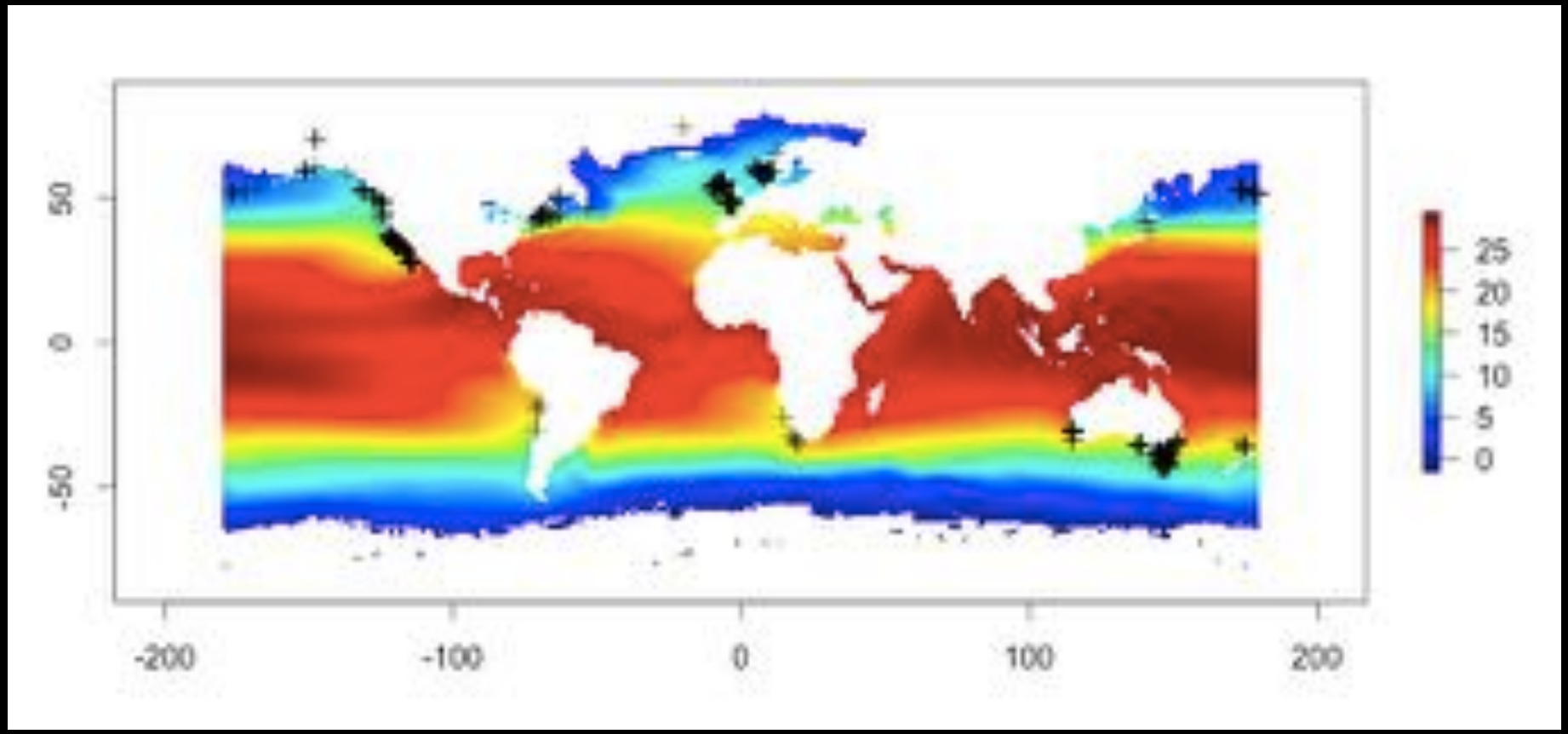
- Work-Life Balance Rating: 4.2
- Salary: $114,808
- Number of Job Openings: 1,315

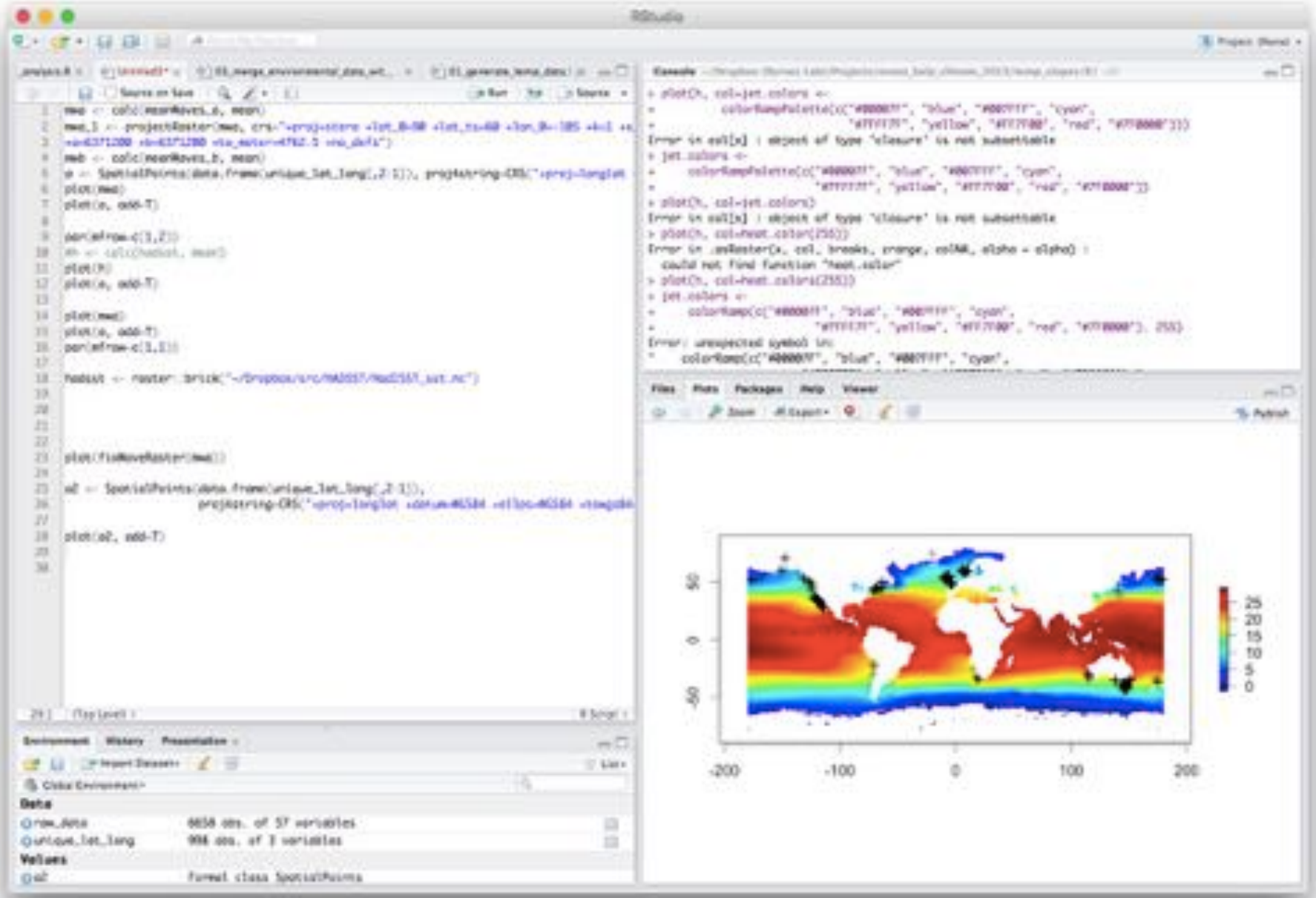
1. Data Scientist

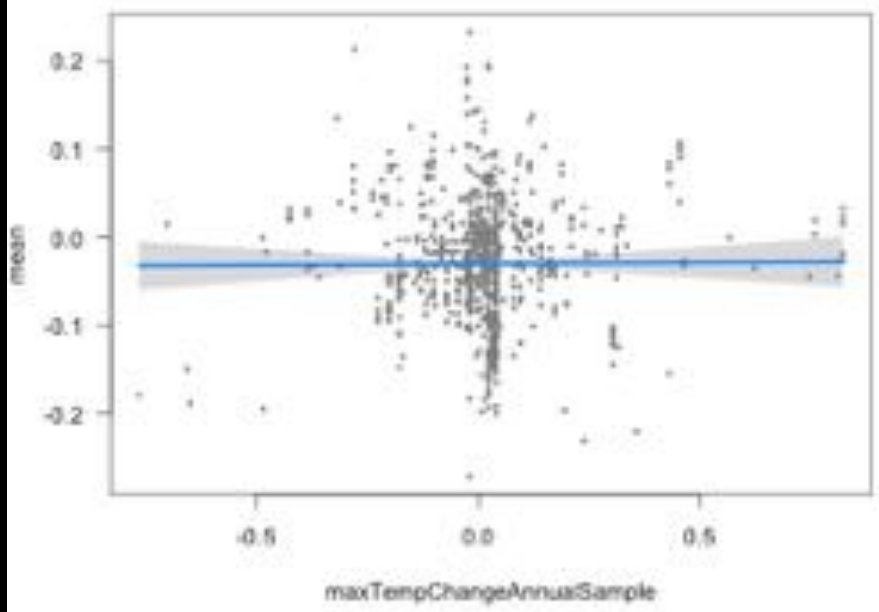
- Work-Life Balance Rating: 4.2
- Salary: $114,808
- Number of Job Openings: 1,315

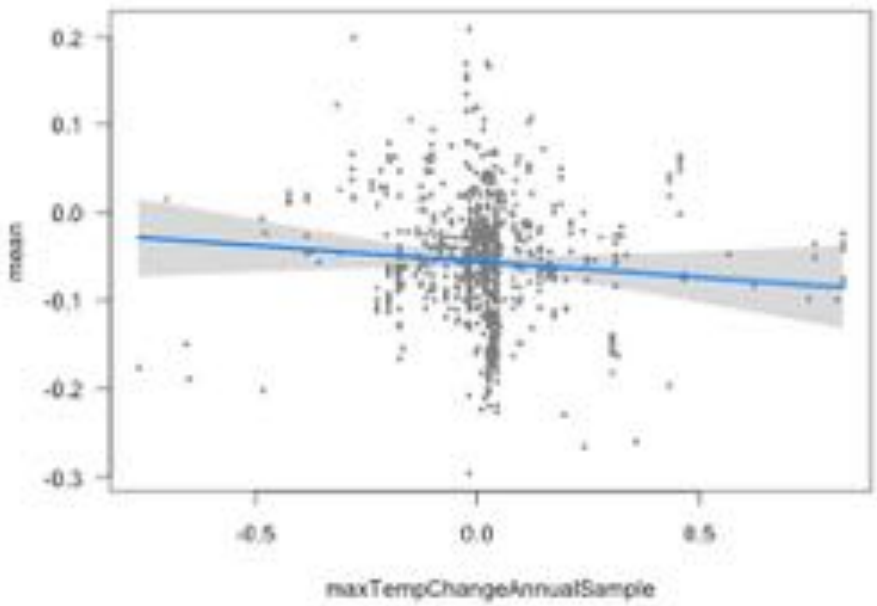


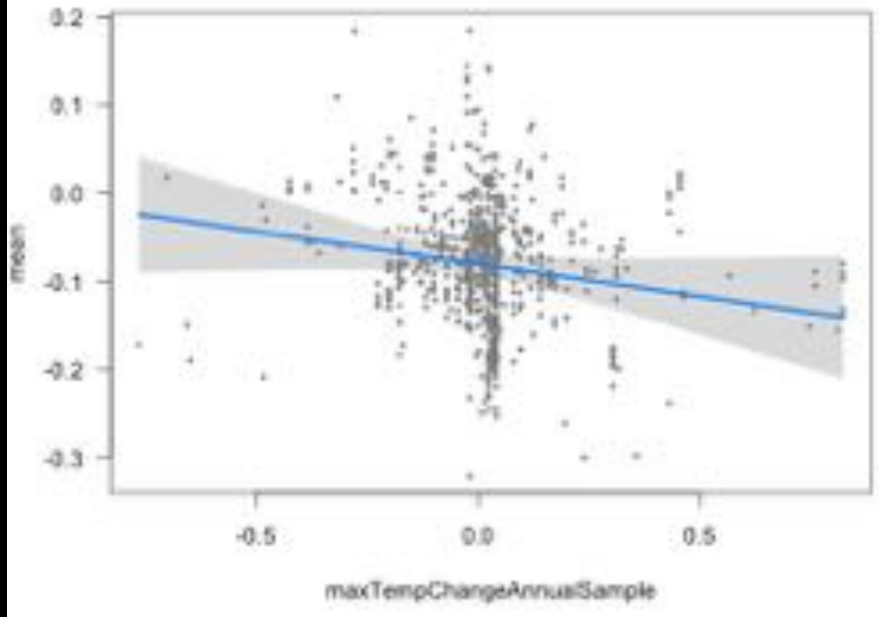
	A	B	C	D	E	F	G	H	I	J	K
1	YEAR	MONTH	DAY	DATE	SITE	TRANSECT	SP_CODE	0-20 IN	20-40 IN	40-20 OFF	20-0 OFF
2	2013	7	24	7/24/2013	BAKER	1	HOAM	1	1	2	1
3	2013	7	24	7/24/2013	BAKER	1	CAIR	1	1	1	0
4	2013	7	24	7/24/2013	BAKER	1	CABO	0	1	1	1
5	2013	7	24	7/24/2013	BAKER	1	ASFO	0	6	7	0
6	2013	7	24	7/24/2013	BAKER	2	HOAM	1	3	2	2
7	2013	7	24	7/24/2013	BAKER	2	CAIR	2	8	11	2
8	2013	7	24	7/24/2013	BAKER	2	CABO	0	1	2	0
9	2013	7	24	7/24/2013	BAKER	2	CAMA	1	9	5	2
10	2013	7	24	7/24/2013	BAKER	2	ASFO	0	4	6	0
11	2013	7	24	7/24/2013	BAKER	2	ASRU	0	0	1	0
12	2013	7	30	7/30/2013	BAKER	3	HOAM	13	3	6	2
13	2013	7	30	7/30/2013	BAKER	3	CAIR	3	3	0	3
14	2013	7	30	7/30/2013	BAKER	3	CABO	6	1	0	2
15	2013	7	30	7/30/2013	BAKER	3	HESA	1	0	1	0
16	2013	7	30	7/30/2013	BAKER	4	HOAM	3	1	4	2
17	2013	7	30	7/30/2013	BAKER	4	CAIR	1	1	1	1
18	2013	7	30	7/30/2013	BAKER	4	CABO	1	0	2	1
19	2013	7	30	7/30/2013	BAKER	4	SADE	8	0	0	3
20	2013	7	24	7/24/2013	BAKER	4	HESA	0	0	0	2
21	2013	7	24	7/24/2013	BAKER	4	PAGURUS	0	0	1	0
22	2013	8	20	8/20/2013	BAKER	5	HOAM	1	0	2	3
23	2013	8	20	8/20/2013	BAKER	5	CAIR	0	1	1	0
24	2013	8	20	8/20/2013	BAKER	5	CABO	1	1	3	0
25	2013	8	20	8/20/2013	BAKER	5	ASFO	0	2	0	0
26	2013	8	20	8/20/2013	BAKER	5	HESA	1	0	0	0
27	2013	8	20	8/20/2013	BAKER	6	HOAM	1	3	0	4
28	2013	8	20	8/20/2013	BAKER	6	CAIR	1	1	2	2
29	2013	8	20	8/20/2013	BAKER	6	CABO	2	2	5	2
30	2013	8	20	8/20/2013	BAKER	6	CAMA	1	0	0	0
31	2013	8	20	8/20/2013	BAKER	6	CAIR	3	1	4	1
32											

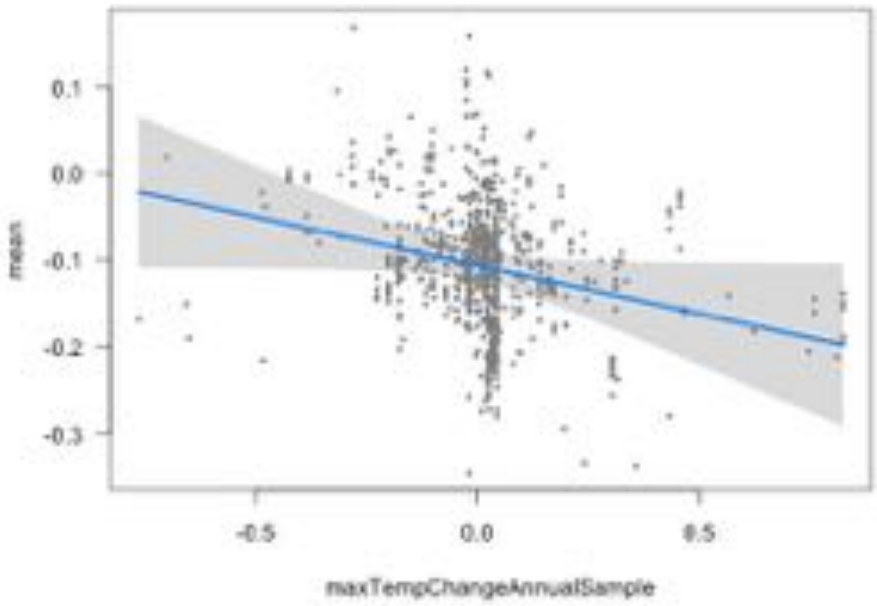

50
0
-50
-200
-100
0
100
200
25
20
15
10
5
0

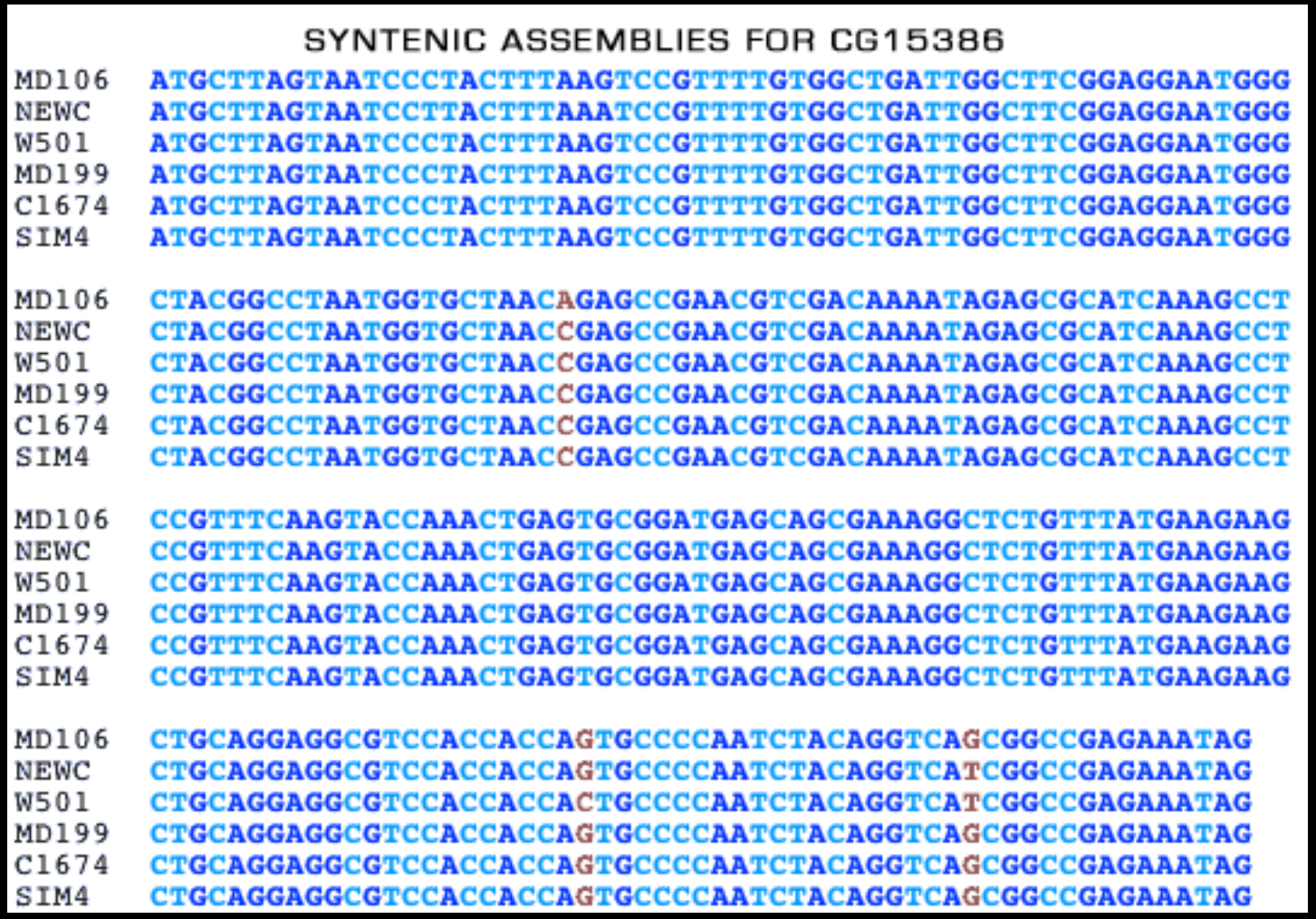

RStudio
Console
Files
Plots
Packages
Help
Viewer
Environment
History
Presentation
Data
raw_data
unique_lat_long
Values
-200
-100
0
100
200
50
0
-50
25
20
15
10
5
0

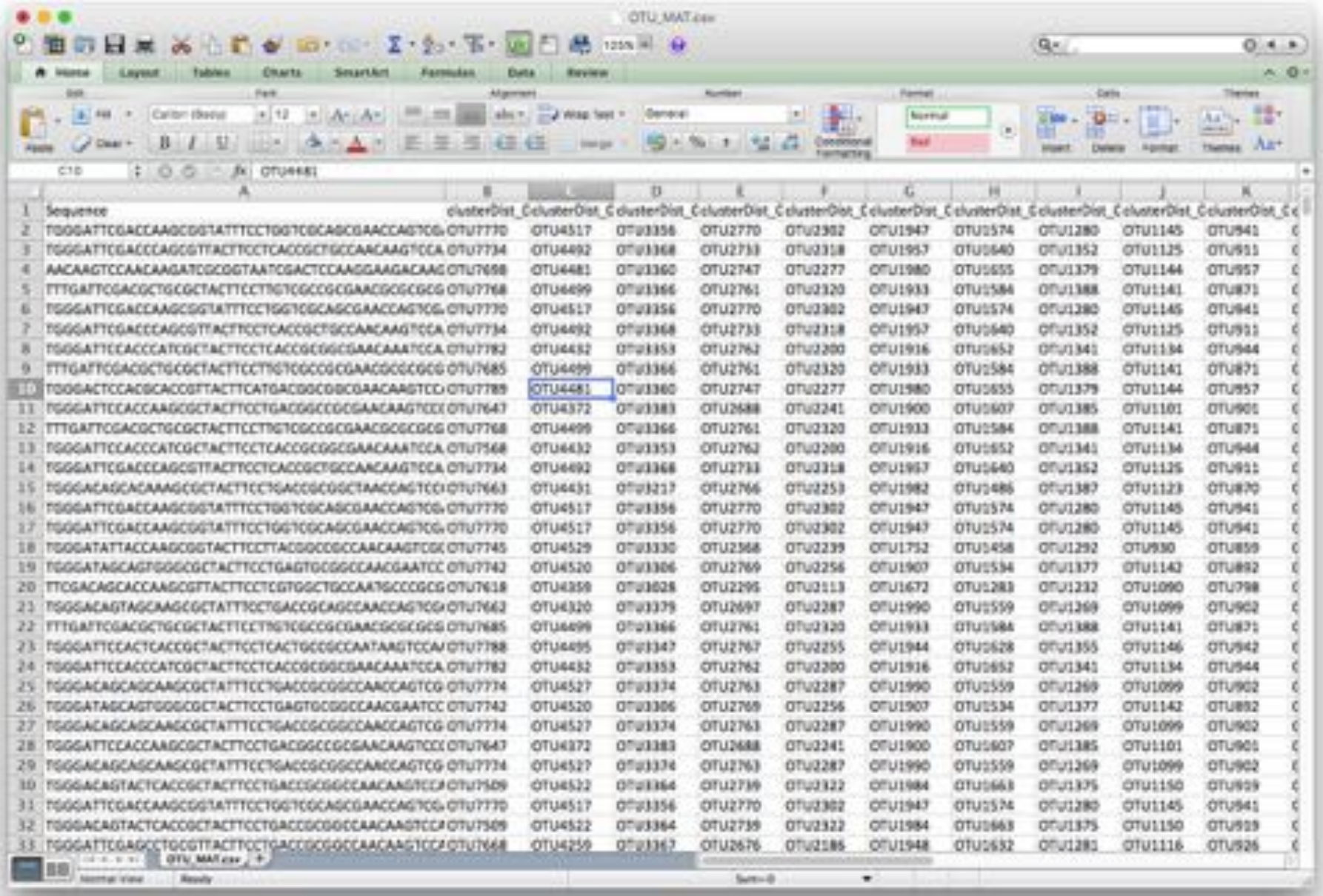

mean
0.2
0.1
0.0
-0.1
-0.2
-0.5
0.0
0.5
maxTempChangeAnnualSample

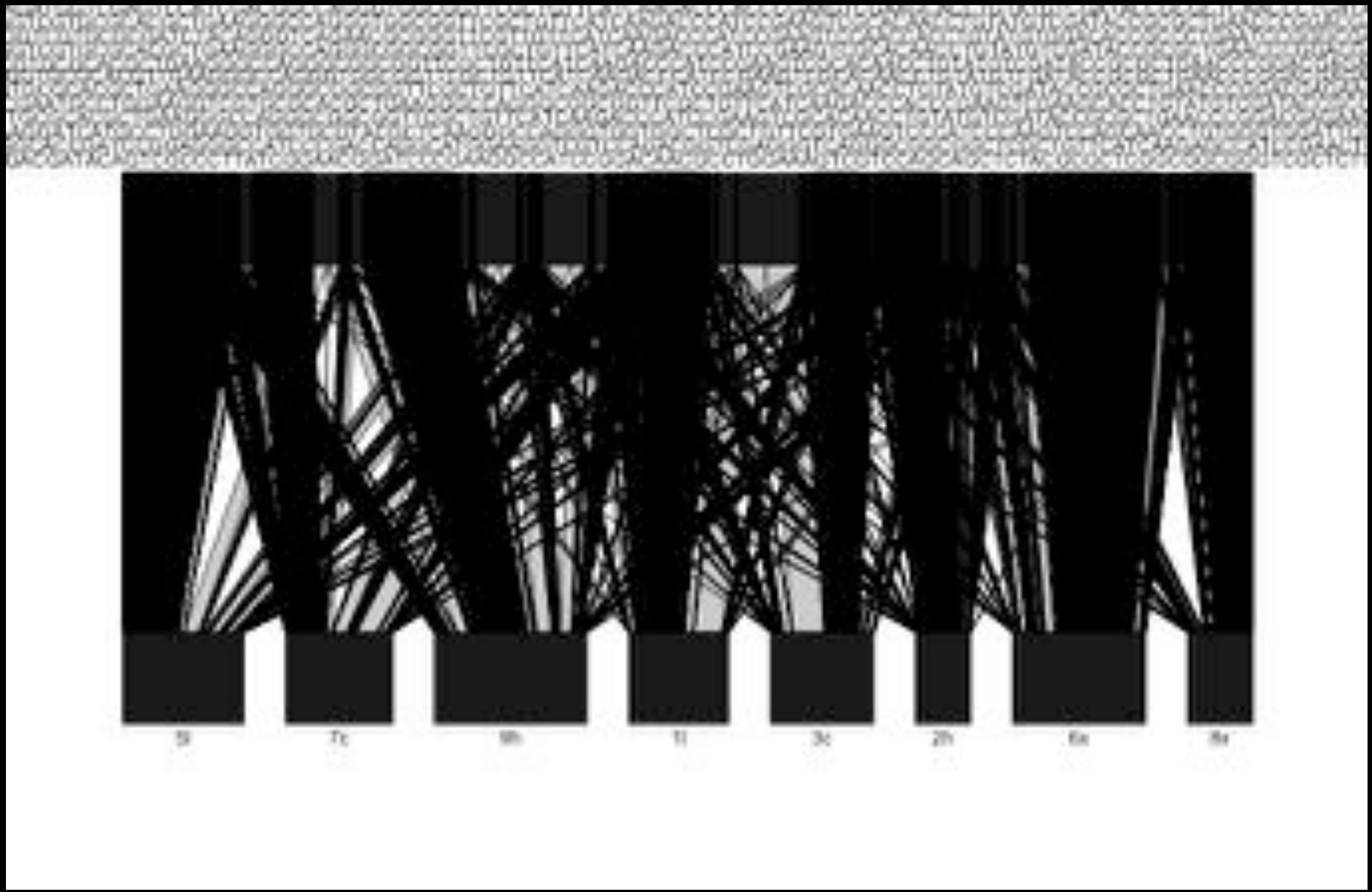

mean
0.2
0.1
0.0
-0.1
-0.2
-0.3
-0.5
0.0
0.5
maxTempChangeAnnualSample

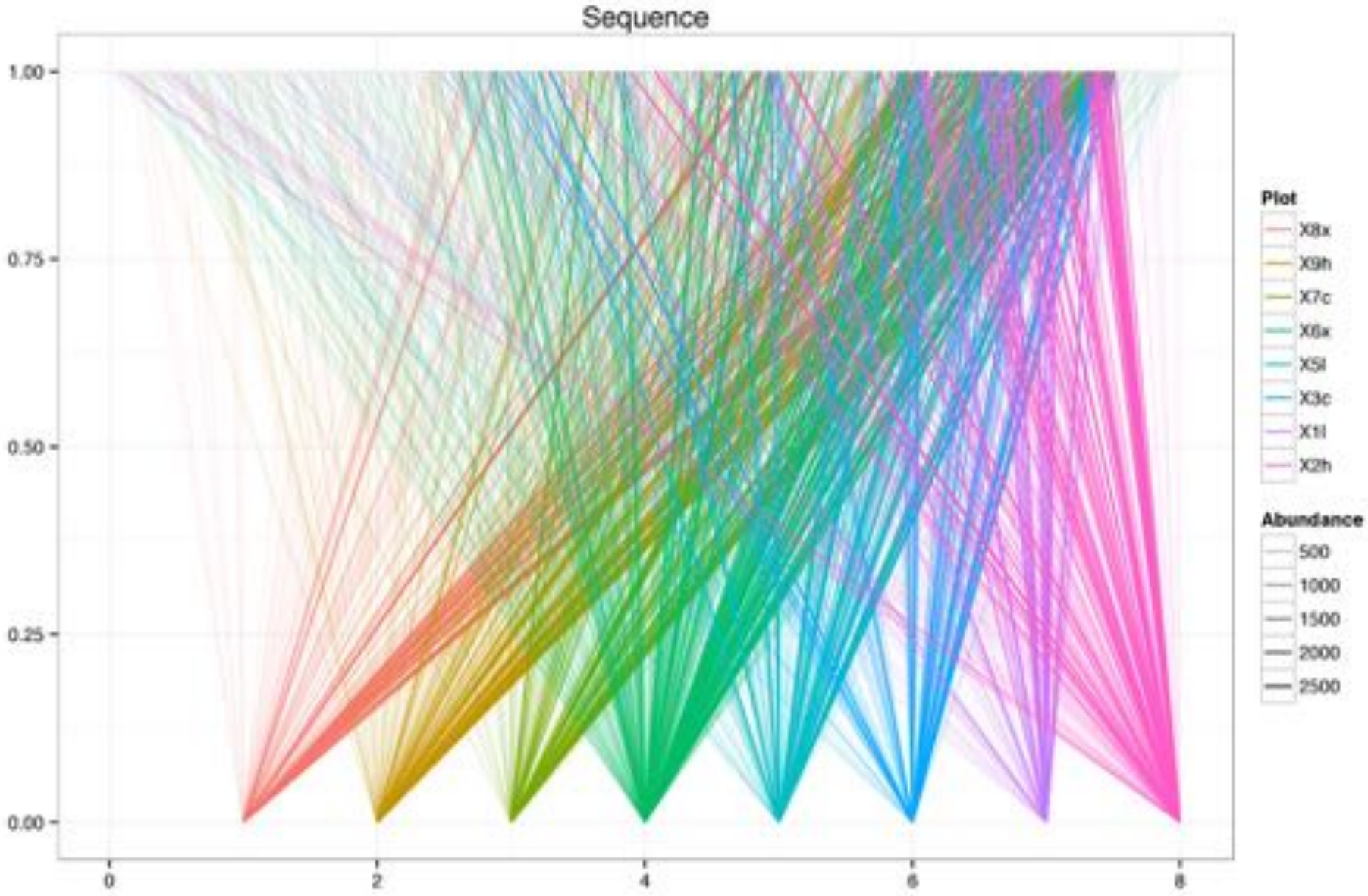

mean
0.2
0.1
0.0
-0.1
-0.2
-0.3
-0.5
0.0
0.5
maxTempChangeAnnualSample


mean
0.1
0.0
-0.1
-0.2
-0.3
-0.5
0.0
0.5
maxTempChangeAnnualSample

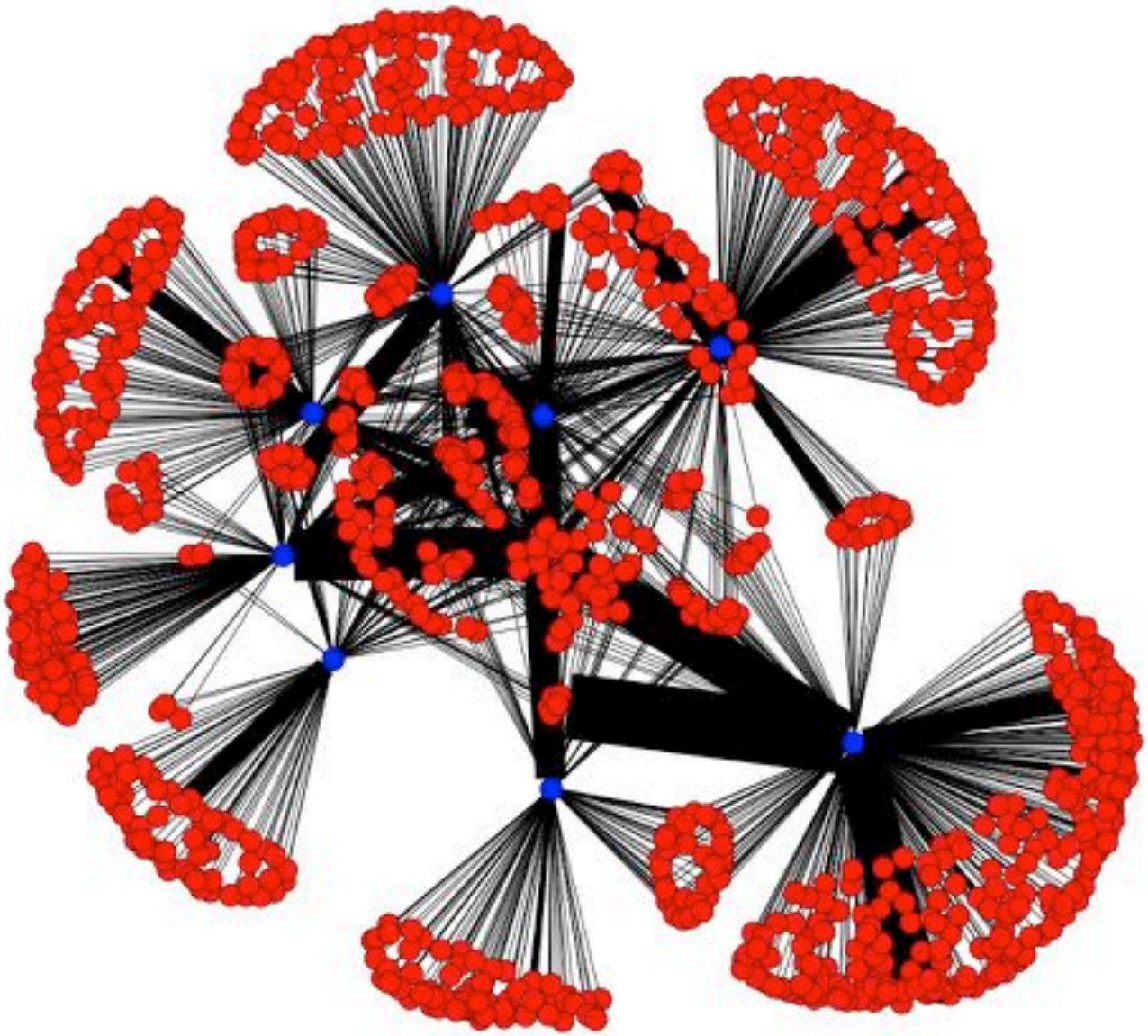

SYNTENIC ASSEMBLIES FOR CG15386
MD106 ATGCTTAGTAATCCCTACTTTAAGTCCGTTTTGTGGCTGATTGGCTTCGGAGGAATGGG
NEWC ATGCTTAGTAATCCTTACTTTAAATCCGTTTTGTGGCTGATTGGCTTCGGAGGAATGGG
W501 ATGCTTAGTAATCCCTACTTTAAGTCCGTTTTGTGGCTGATTGGCTTCGGAGGAATGGG
MD199 ATGCTTAGTAATCCCTACTTTAAGTCCGTTTTGTGGCTGATTGGCTTCGGAGGAATGGG
C1674 ATGCTTAGTAATCCCTACTTTAAGTCCGTTTTGTGGCTGATTGGCTTCGGAGGAATGGG
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


OTU_MAT.csv
Sequence
clusterDist_C
OTU7770
OTU4517
OTU3356
OTU2770
OTU2302
OTU1947
OTU1574
OTU1280
OTU1145
OTU941
OTU7734
OTU4492
OTU3368
OTU2733
OTU2318
OTU1957
OTU1640
OTU1352
OTU1125
OTU911
OTU7698
OTU4481
OTU3360
OTU2747
OTU2277
OTU1980
OTU1655
OTU1379
OTU1144
OTU957
OTU7768
OTU4499
OTU3366
OTU2761
OTU2320
OTU1933
OTU1584
OTU1388
OTU1141
OTU871
OTU7685
OTU4499
OTU3366
OTU2761
OTU2320
OTU1933
OTU1584
OTU1388
OTU1141
OTU871
OTU7789
OTU4481
OTU3360
OTU2747
OTU2277
OTU1980
OTU1655
OTU1379
OTU1144
OTU957
OTU7647
OTU4372
OTU3383
OTU2688
OTU2241
OTU1900
OTU1607
OTU1385
OTU1101
OTU901
OTU7768
OTU4499
OTU3366
OTU2761
OTU2320
OTU1933
OTU1584
OTU1388
OTU1141
OTU871
OTU7568
OTU4432
OTU3353
OTU2762
OTU2200
OTU1916
OTU1652
OTU1341
OTU1134
OTU944
OTU7734
OTU4492
OTU3368
OTU2733
OTU2318
OTU1957
OTU1640
OTU1352
OTU1125
OTU911
OTU7663
OTU4431
OTU3217
OTU2766
OTU2253
OTU1982
OTU1486
OTU1387
OTU1123
OTU870
OTU7770
OTU4517
OTU3356
OTU2770
OTU2302
OTU1947
OTU1574
OTU1280
OTU1145
OTU941
OTU7770
OTU4517
OTU3356
OTU2770
OTU2302
OTU1947
OTU1574
OTU1280
OTU1145
OTU941
OTU7745
OTU4529
OTU3330
OTU2568
OTU2239
OTU1752
OTU1458
OTU1292
OTU930
OTU859
OTU7742
OTU4520
OTU3306
OTU2769
OTU2256
OTU1907
OTU1534
OTU1377
OTU1142
OTU892
OTU7618
OTU4359
OTU3028
OTU2295
OTU2113
OTU1672
OTU1283
OTU1232
OTU1090
OTU798
OTU7662
OTU4320
OTU3379
OTU2697
OTU2287
OTU1990
OTU1559
OTU1269
OTU1099
OTU902
OTU7685
OTU4499
OTU3366
OTU2761
OTU2320
OTU1933
OTU1584
OTU1388
OTU1141
OTU871
OTU7788
OTU4495
OTU3347
OTU2767
OTU2255
OTU1944
OTU1628
OTU1355
OTU1146
OTU942
OTU7782
OTU4432
OTU3353
OTU2762
OTU2200
OTU1916
OTU1652
OTU1341
OTU1134
OTU944
OTU7774
OTU4527
OTU3374
OTU2763
OTU2287
OTU1990
OTU1559
OTU1269
OTU1099
OTU902
OTU7742
OTU4520
OTU3306
OTU2769
OTU2256
OTU1907
OTU1534
OTU1377
OTU1142
OTU892
OTU7774
OTU4527
OTU3374
OTU2763
OTU2287
OTU1990
OTU1559
OTU1269
OTU1099
OTU902
OTU7647
OTU4372
OTU3383
OTU2688
OTU2241
OTU1900
OTU1607
OTU1385
OTU1101
OTU901
OTU7734
OTU4527
OTU3374
OTU2763
OTU2287
OTU1990
OTU1559
OTU1269
OTU1099
OTU902
OTU7509
OTU4522
OTU3364
OTU2739
OTU2322
OTU1984
OTU1663
OTU1375
OTU1150
OTU919
OTU7770
OTU4517
OTU3356
OTU2770
OTU2302
OTU1947
OTU1574
OTU1280
OTU1145
OTU941
OTU7509
OTU4522
OTU3364
OTU2739
OTU2322
OTU1984
OTU1663
OTU1375
OTU1150
OTU919
OTU7668
OTU4259
OTU3367
OTU2676
OTU2186
OTU1948
OTU1632
OTU1281
OTU1116
OTU926

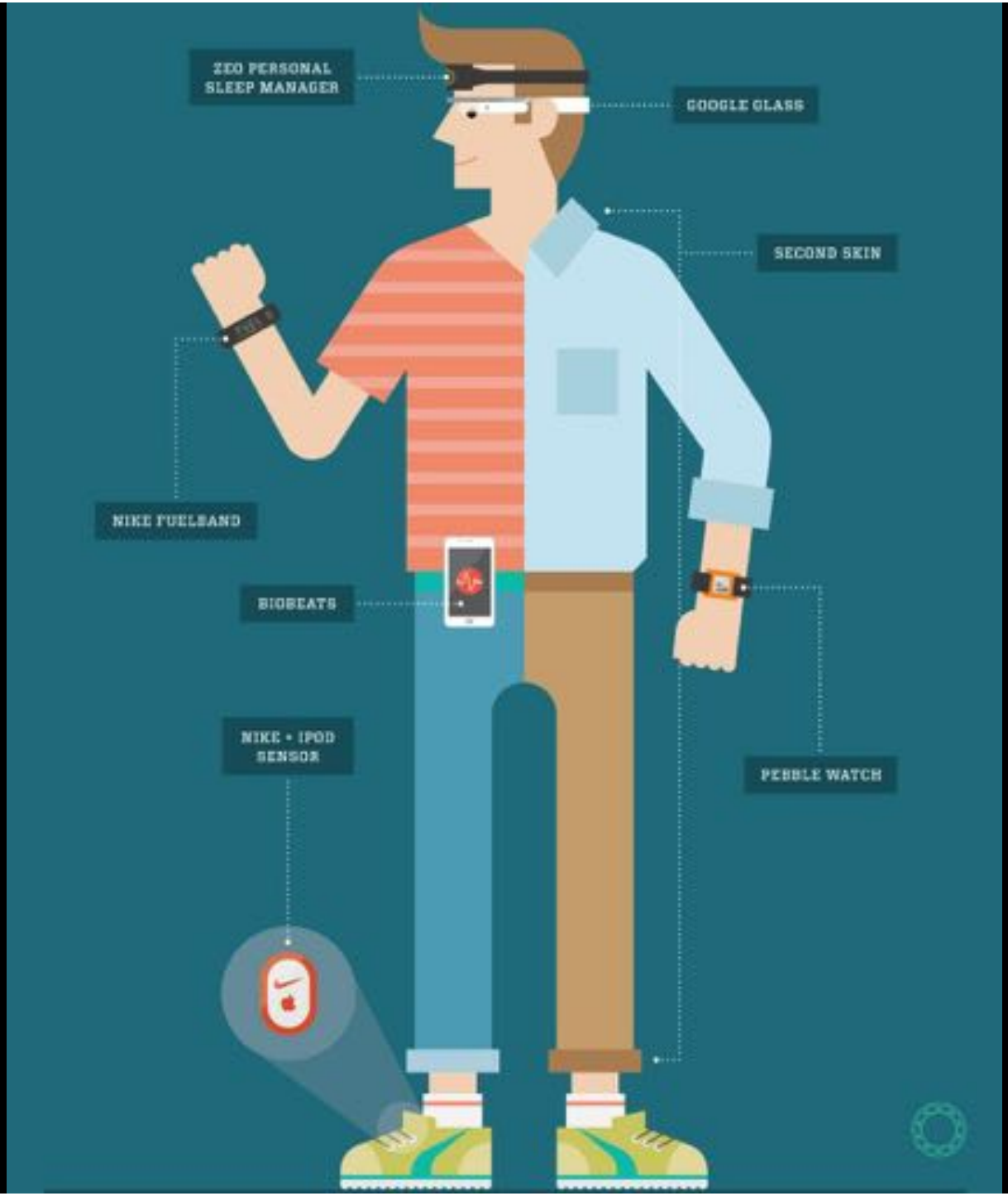



Sequence
1.00
0.75
0.50
0.25
0.00
0
2
4
6
8
Plot
X8x
X9h
X7c
X6x
X5l
X3c
X1l
X2h
Abundance
500
1000
1500
2000
2500

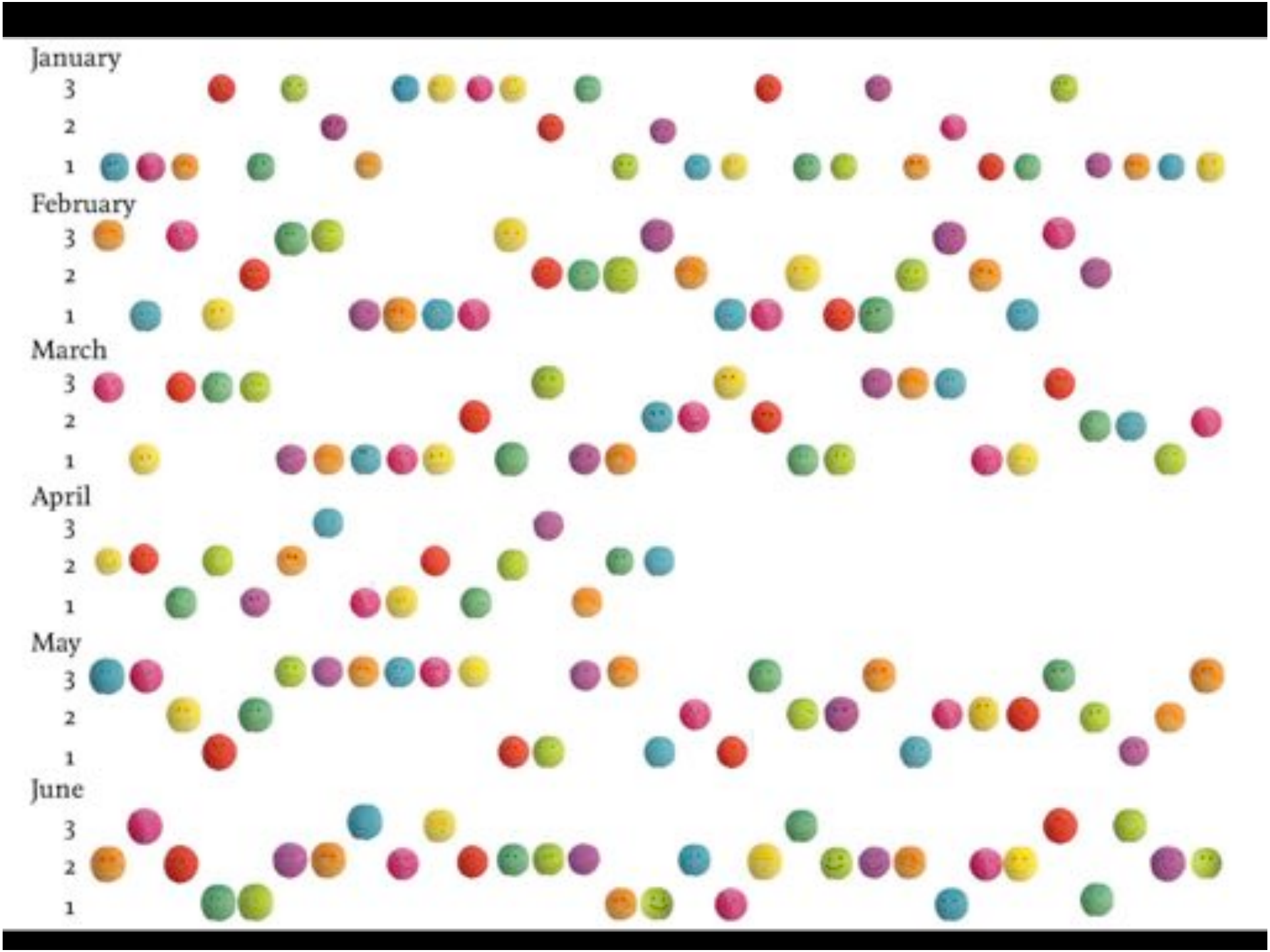
clusterDist_0.09



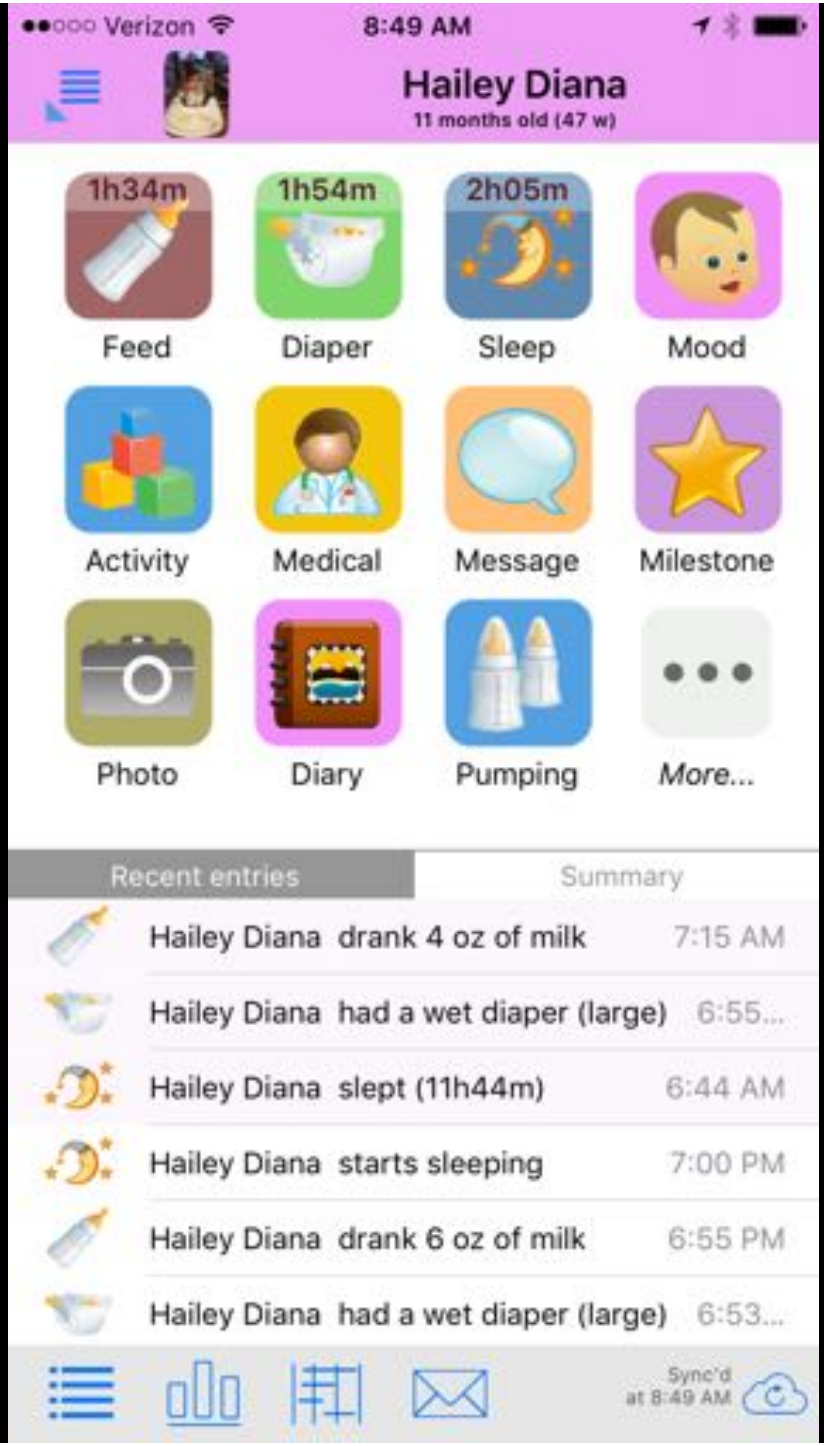

8:44

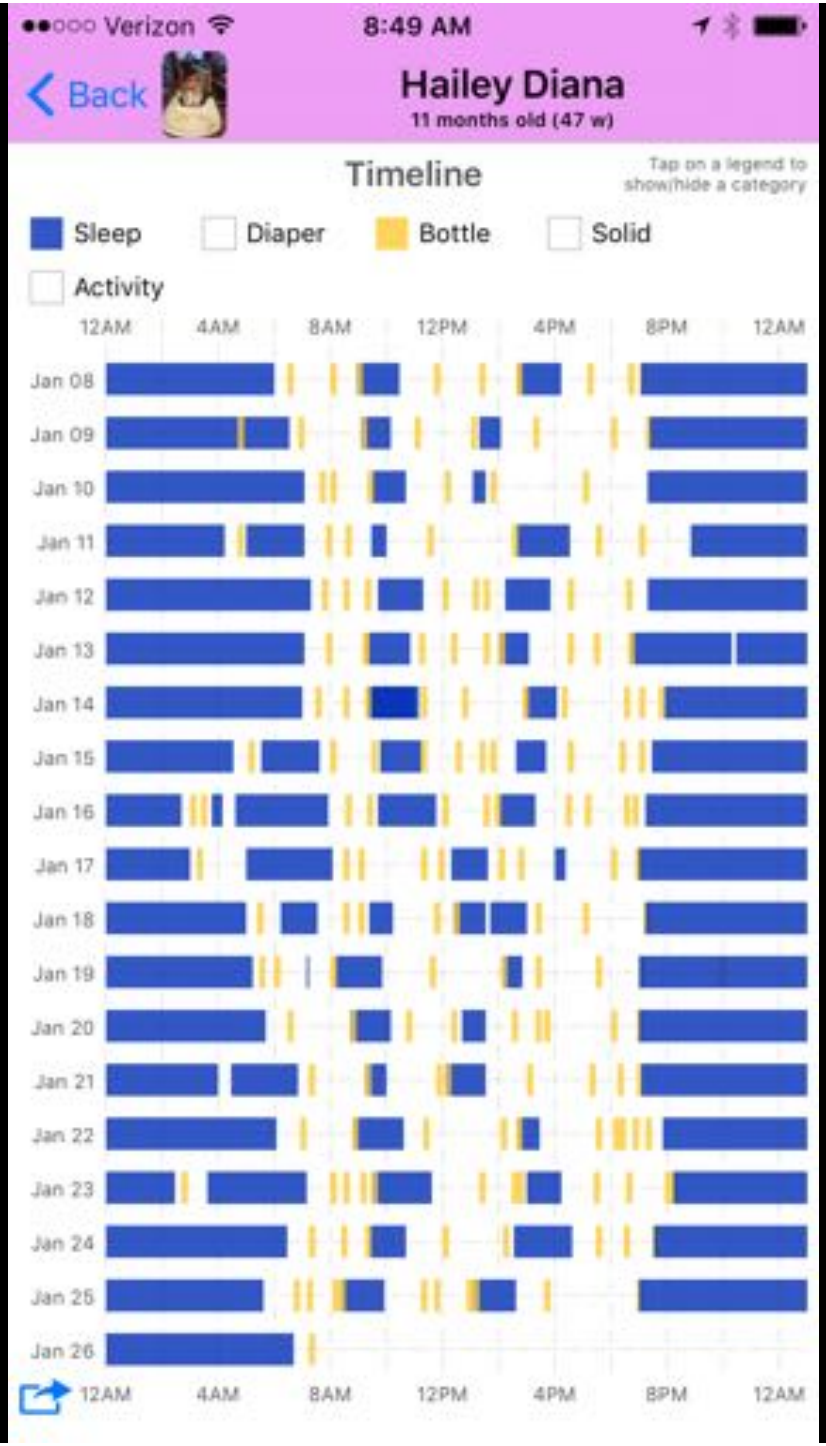

ZEO PERSONAL SLEEP MANAGER
GOOGLE GLASS
SECOND SKIN
NIKE FUELBAND
BIOBEATS
NIKE + IPOD SENSOR
PEBBLE WATCH


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


Pepcid


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
Tap on a legend to show/hide a category
Sleep
Diaper
Bottle
Solid
Activity
12AM
4AM
8AM
12PM
4PM
8PM
12AM
Jan 08
Jan 09
Jan 10
Jan 11
Jan 12
Jan 13
Jan 14
Jan 15
Jan 16
Jan 17
Jan 18
Jan 19
Jan 20
Jan 21
Jan 22
Jan 23
Jan 24
Jan 25
Jan 26
12AM
4AM
8AM
12PM
4PM
8PM
12AM

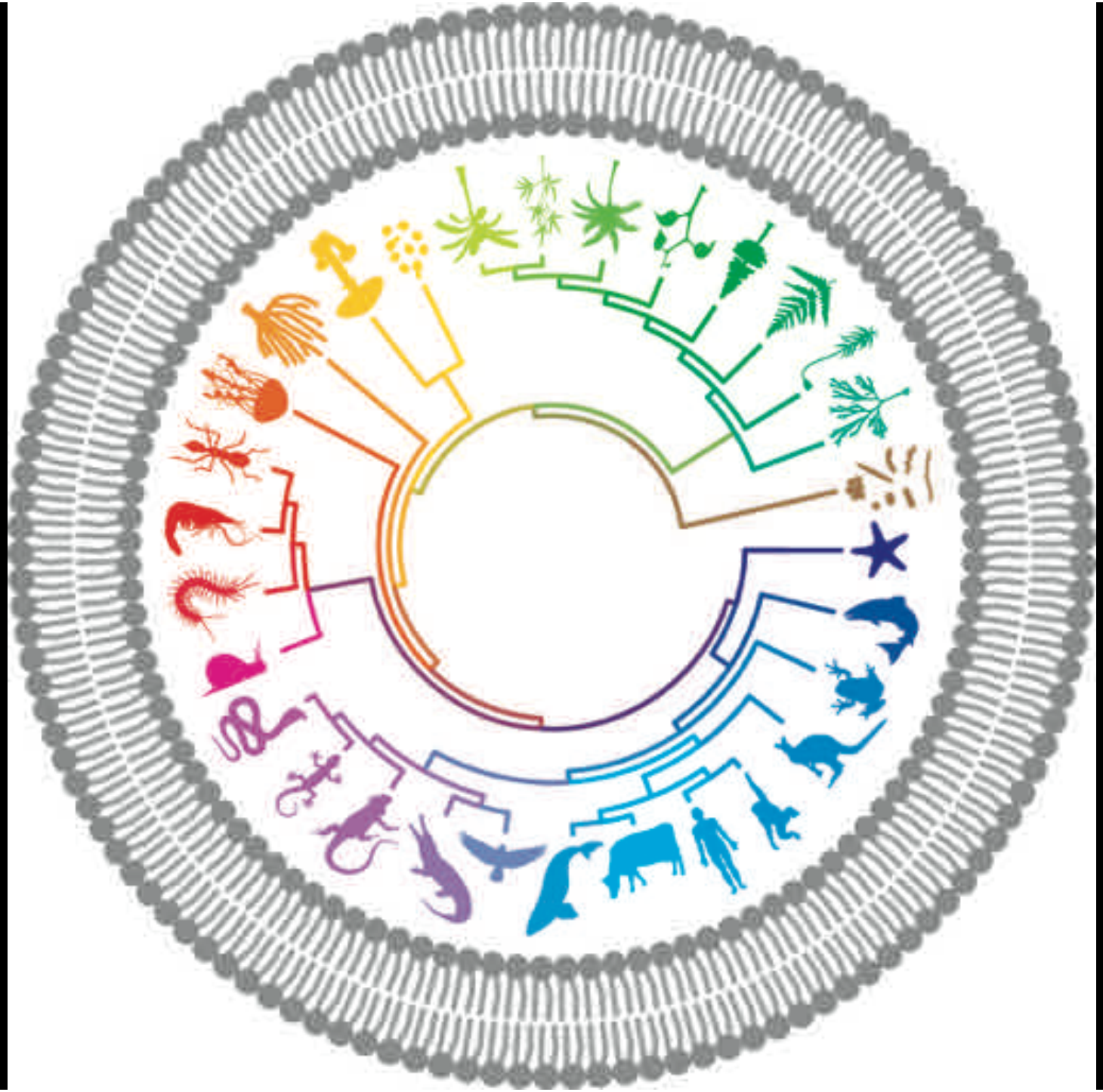
We Are Awash in Data

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

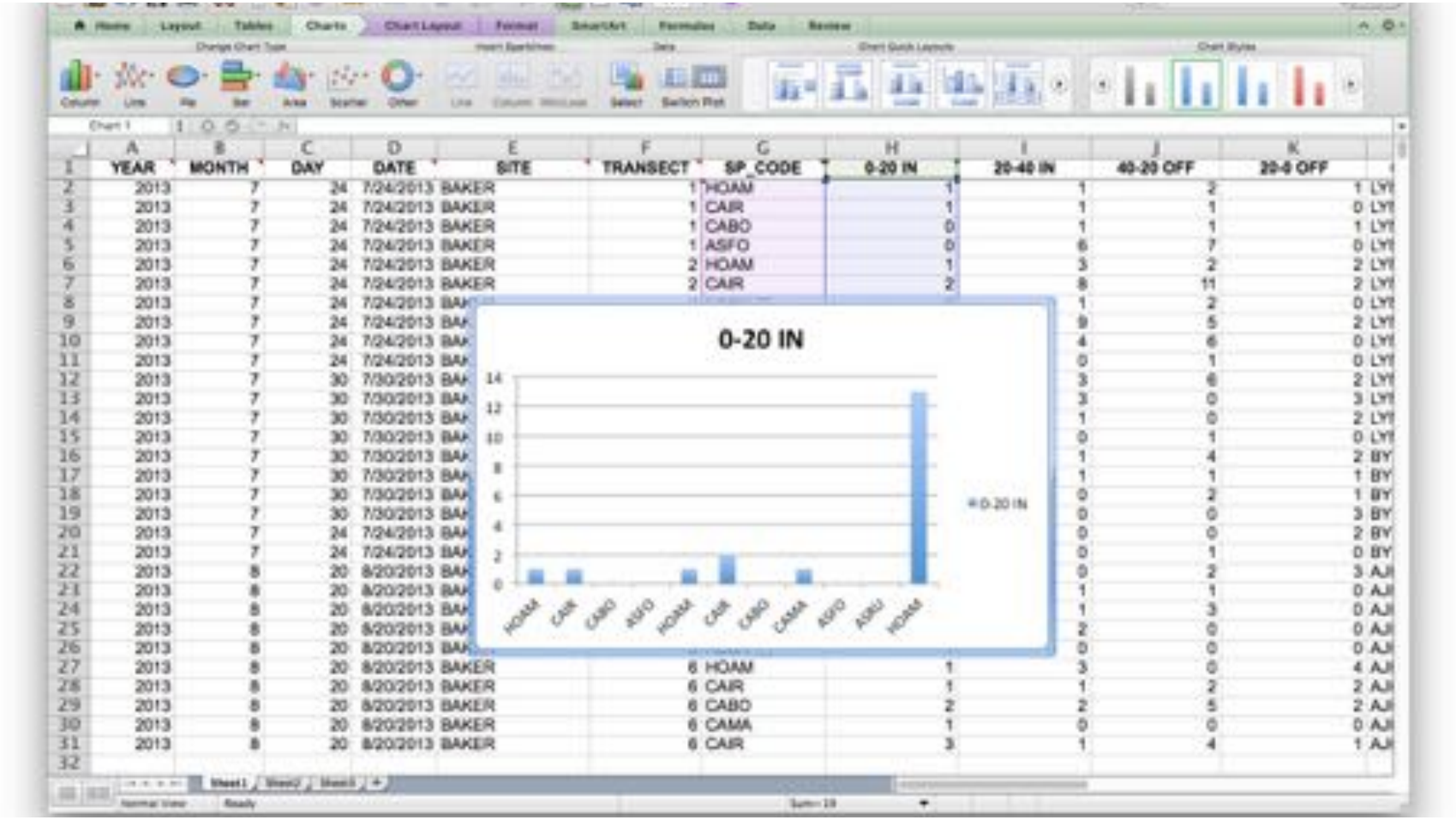
Data is at the Center of Biology


Linked
Different
human
Immune
Critical
Learning
Pipeline
Semantic
translational
Genes
genomic
Prioritization
Cohort
Biomedical
Bioinformatics
neurodegenerative
Health
Repository
model
Screening
Expression
Novel
Framework
Adverse
Modeling
Epistatic
network
Mining
informatics
Proteomics
Cancer
Integrative
Based
Discovery
Multinet
Interactions
analysis
Pharmacokinetics
Disease
Knowledge
Quantifying
Medicine
Preprocessing
Approach
Prediction
Finding
tool
Bioinformatic
risk
Questionnaire
Multiple
bioinformatics
Integrating
Genome
Sources
health
Enhance
Application
Drug
Research
patients
Translational
Data
Factor
data
gene
SNPs
Genomic
Demographics
Cell
Study
using
Base
Omic
Bayesian
Association
Pharmacogenomics
Gene
Free
Young
Identification
Automatic
Automated
American
Towards
expression
Integrated
Integration
Role
Alcohol
Large
Analysis
Using
Model
Sequencing
Frames
Space
Patients
Acute
Genetic
Management
Risk
Interaction
Bootstrapping
Scale
identification
EHR
Genetics
Extraction
Dependence
Network-Based
Clinical
Prediction-based
Generation
Treatment
Potential
SemPoD
Ontology
within
Text
Information-Based
Numerical
Annotation
Method
Question
Vector
Personalized
Networks
clinical
Patient
History
research
HIV
multiple
Response
Heart
Studies




ss Lever
to Start
Fail
sfer
OK
onset
Part No.
U-DTW-1

Classical Tools Not Up to the Task



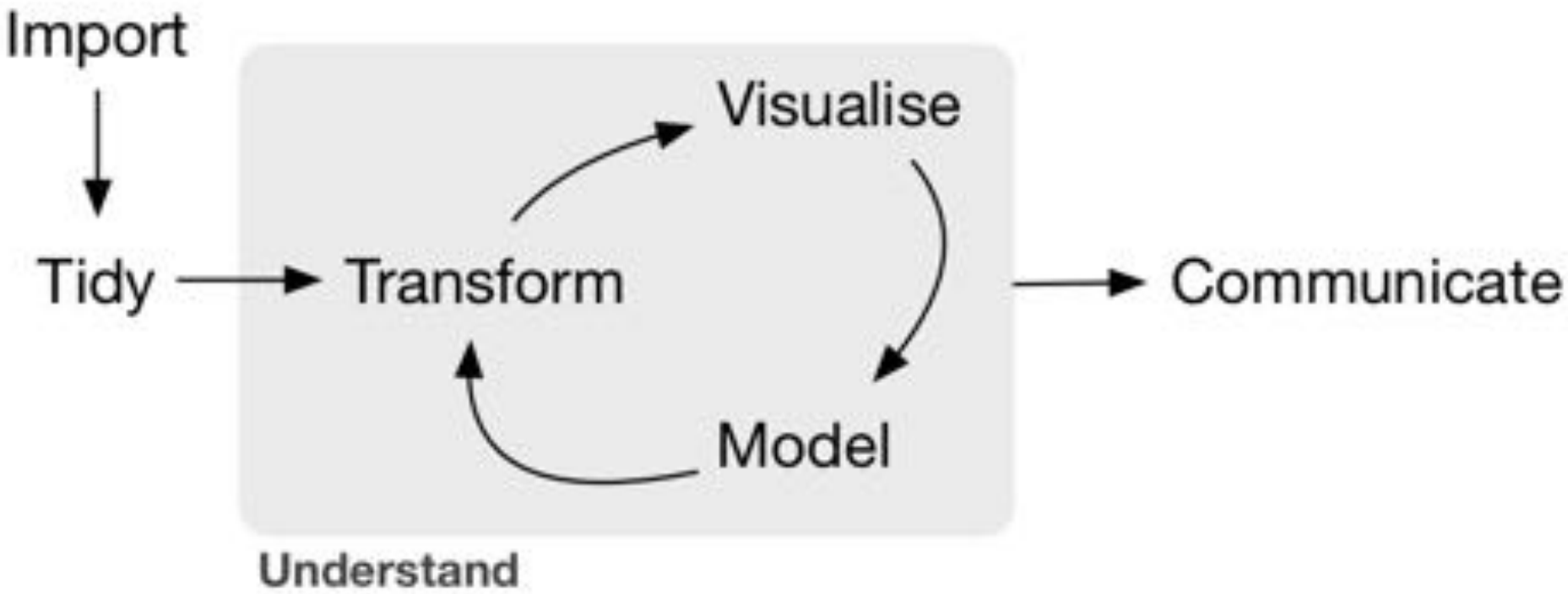
	A	B	C	D	E	F	G	H	I	J	K	
1	YEAR	MONTH	DAY	DATE	SITE	TRANSECT	SP_CODE	0-20 IN	20-40 IN	40-20 OFF	20-0 OFF	
2	2013	7	24	7/24/2013	BAKER	1	HOAM	1	1	2	1	LYI
3	2013	7	24	7/24/2013	BAKER	1	CAIR	1	1	1	0	LYI
4	2013	7	24	7/24/2013	BAKER	1	CABO	0	1	1	1	LYI
5	2013	7	24	7/24/2013	BAKER	1	ASFO	0	6	7	0	LYI
6	2013	7	24	7/24/2013	BAKER	2	HOAM	1	3	2	2	LYI
7	2013	7	24	7/24/2013	BAKER	2	CAIR	2	8	11	2	LYI
8	2013	7	24	7/24/2013	BA				1	2	0	LYI
9	2013	7	24	7/24/2013	BA				9	5	2	LYI
10	2013	7	24	7/24/2013	BA				4	6	0	LYI
11	2013	7	24	7/24/2013	BA				0	1	0	LYI
12	2013	7	30	7/30/2013	BA				3	6	2	LYI
13	2013	7	30	7/30/2013	BA				3	0	3	LYI
14	2013	7	30	7/30/2013	BA				1	0	2	LYI
15	2013	7	30	7/30/2013	BA				0	1	0	LYI
16	2013	7	30	7/30/2013	BA				1	4	2	BY
17	2013	7	30	7/30/2013	BA				1	1	1	BY
18	2013	7	30	7/30/2013	BA				0	2	1	BY
19	2013	7	30	7/30/2013	BA				0	0	3	BY
20	2013	7	24	7/24/2013	BA				0	0	2	BY
21	2013	7	24	7/24/2013	BA				0	1	0	BY
22	2013	8	20	8/20/2013	BA				0	2	3	AJI
23	2013	8	20	8/20/2013	BA				1	1	0	AJI
24	2013	8	20	8/20/2013	BA				1	3	0	AJI
25	2013	8	20	8/20/2013	BA				2	0	0	AJI
26	2013	8	20	8/20/2013	BA				0	0	0	AJI
27	2013	8	20	8/20/2013	BAKER	6	HOAM	1	3	0	4	AJI
28	2013	8	20	8/20/2013	BAKER	6	CAIR	1	1	2	2	AJI
29	2013	8	20	8/20/2013	BAKER	6	CABO	2	2	5	2	AJI
30	2013	8	20	8/20/2013	BAKER	6	CAMA	1	0	0	0	AJI
31	2013	8	20	8/20/2013	BAKER	6	CAIR	3	1	4	1	AJI
32												

Classical Tools Not Up to the Task

Introduction to Data Science for Biology

Who Are You?

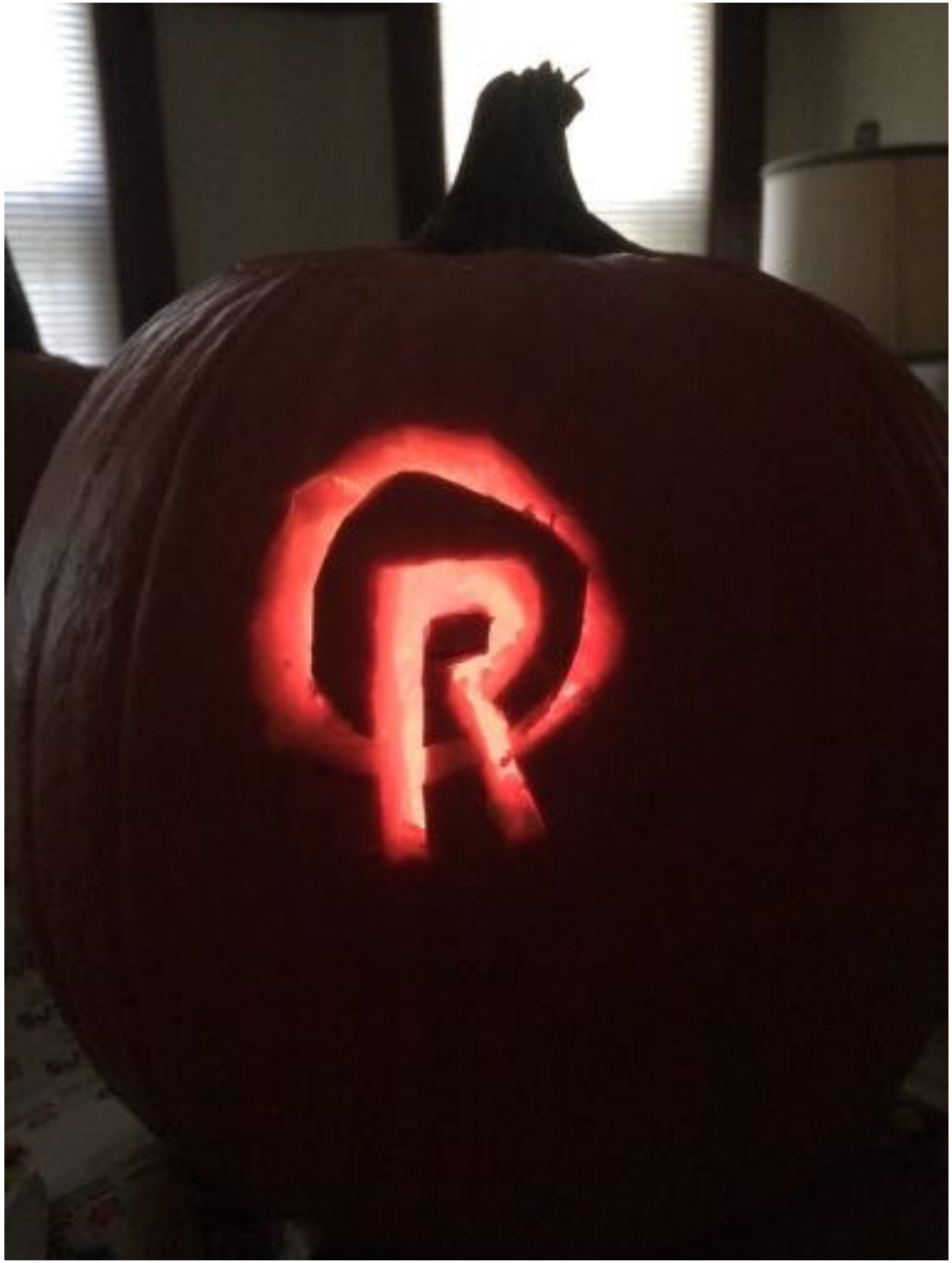
Our Semester


Import
Tidy
Transform
Visualise
Model
Communicate
Understand

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	247/24/2013		BAKER		1HOAM	1	1	2
2013	7	247/24/2013		BAKER		1CAIR	1	1	1
2013	7	247/24/2013		BAKER		1CABO	0	1	1
2013	7	247/24/2013		BAKER		1ASFO	0	6	7
2013	7	247/24/2013		BAKER		2HOAM	1	3	2
2013	7	247/24/2013		BAKER		2CAIR	2	8	11
2013	7	247/24/2013		BAKER		2CABO	0	1	2
2013	7	247/24/2013		BAKER		2CAMA	1	9	5
2013	7	247/24/2013		BAKER		2ASFO	0	4	6
2013	7	247/24/2013		BAKER		2ASRU	0	0	1

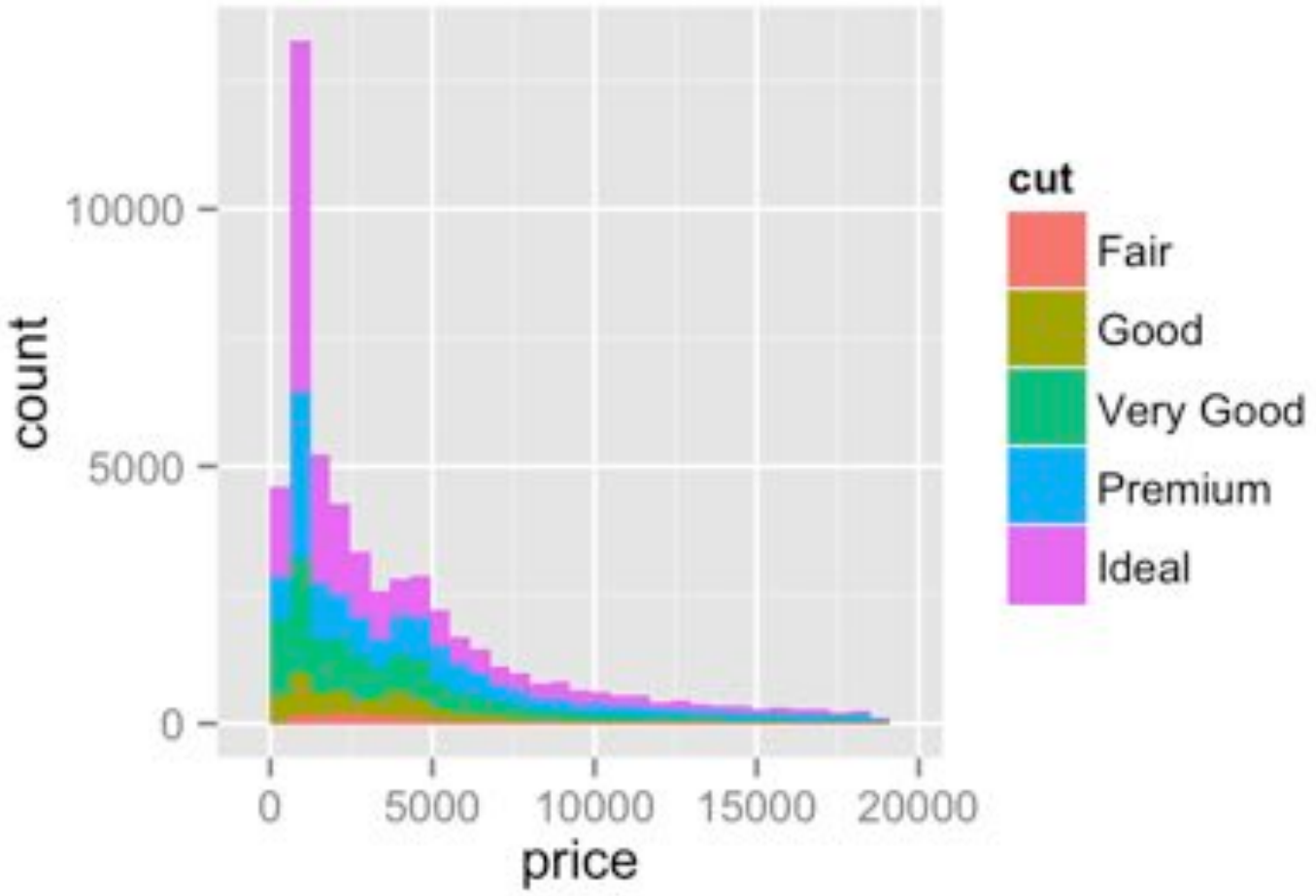
Learn common programming language(s) associated with data science



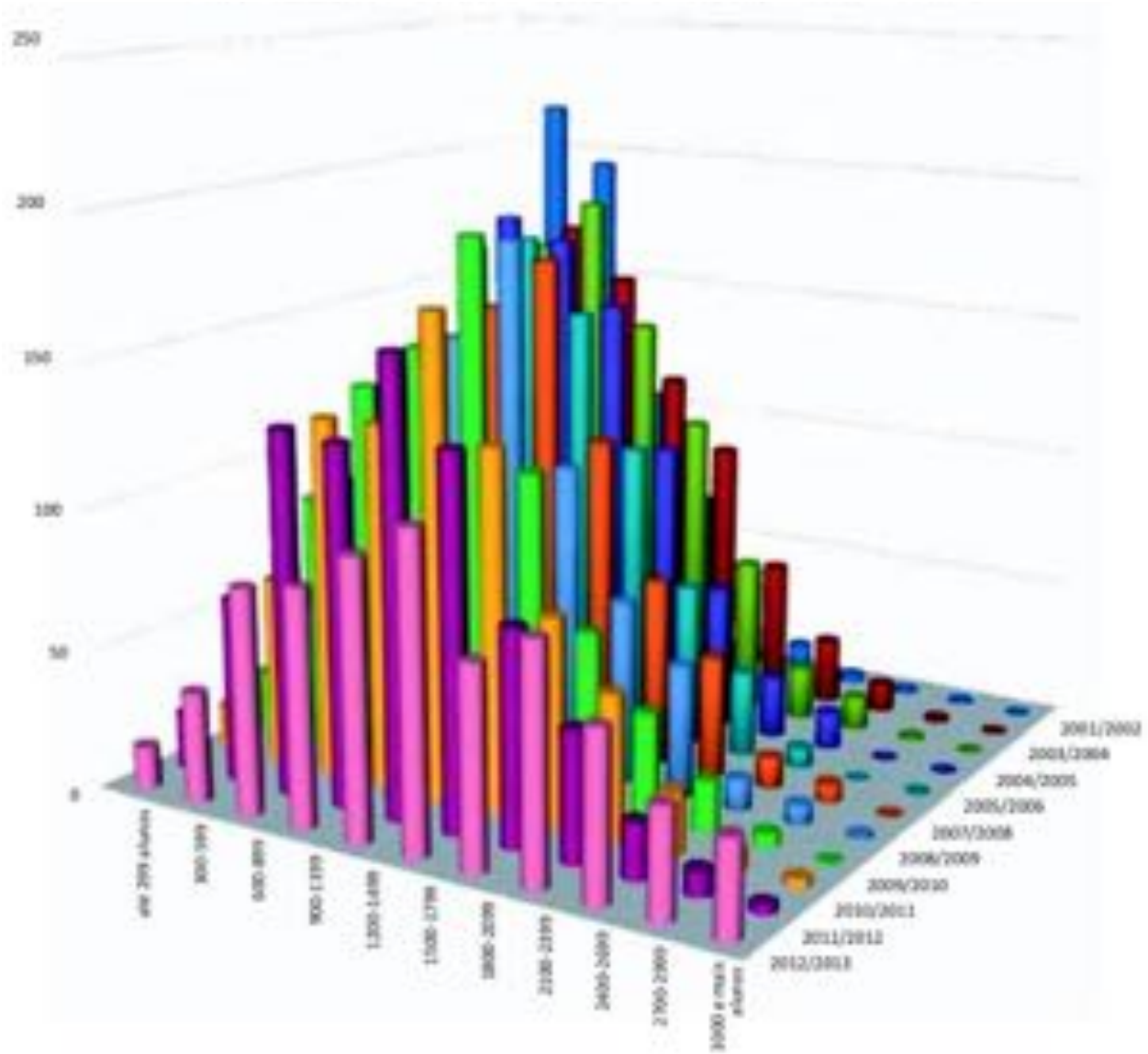

R
4CATS


ONE DOES NOT SIMPLY
PASTE CODE INTO [R]

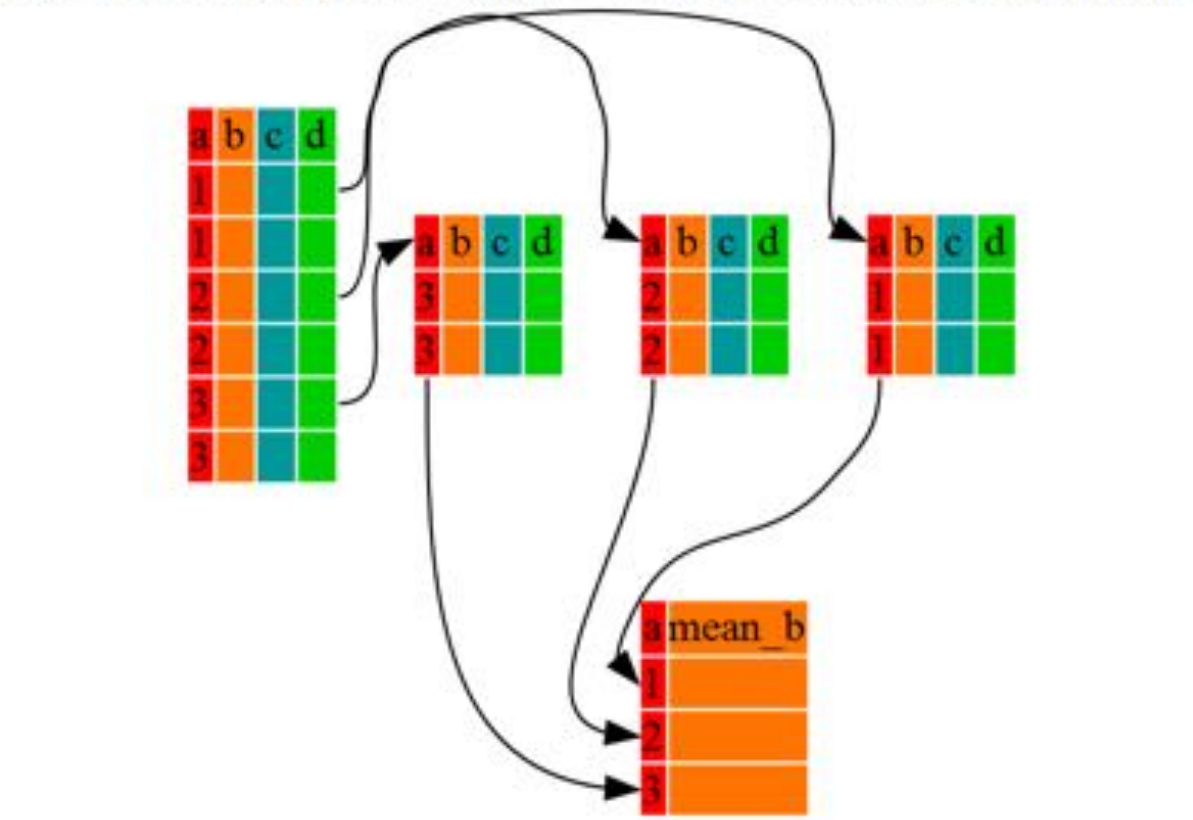
Build a vocabulary of visualization tools that enable students to see what their data means



This is How I Know I Failed You



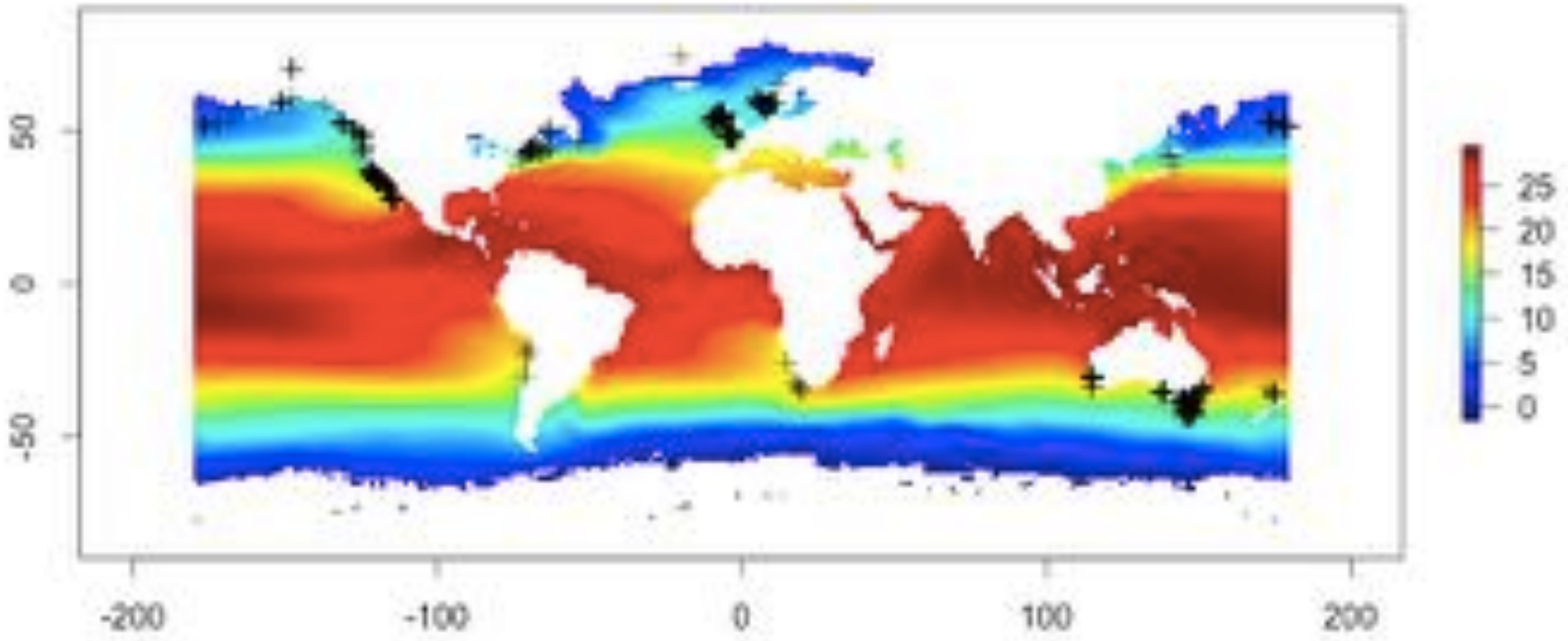
Develop an understanding of how to manipulate data for the purposes of seeing useful patterns



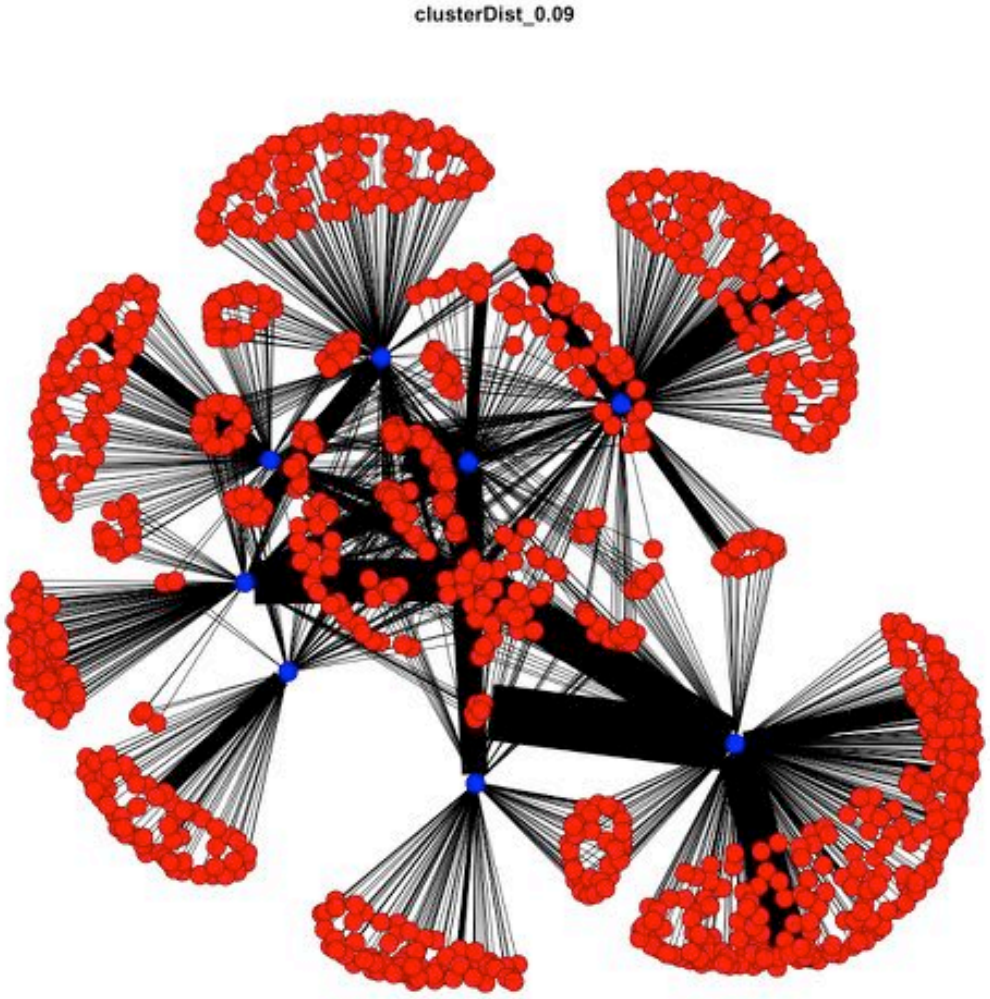
http://swcarpentry.github.io/r-novice-gapminder/


TRIED LEARNING R
DIDN'T INSTALL DPLYR

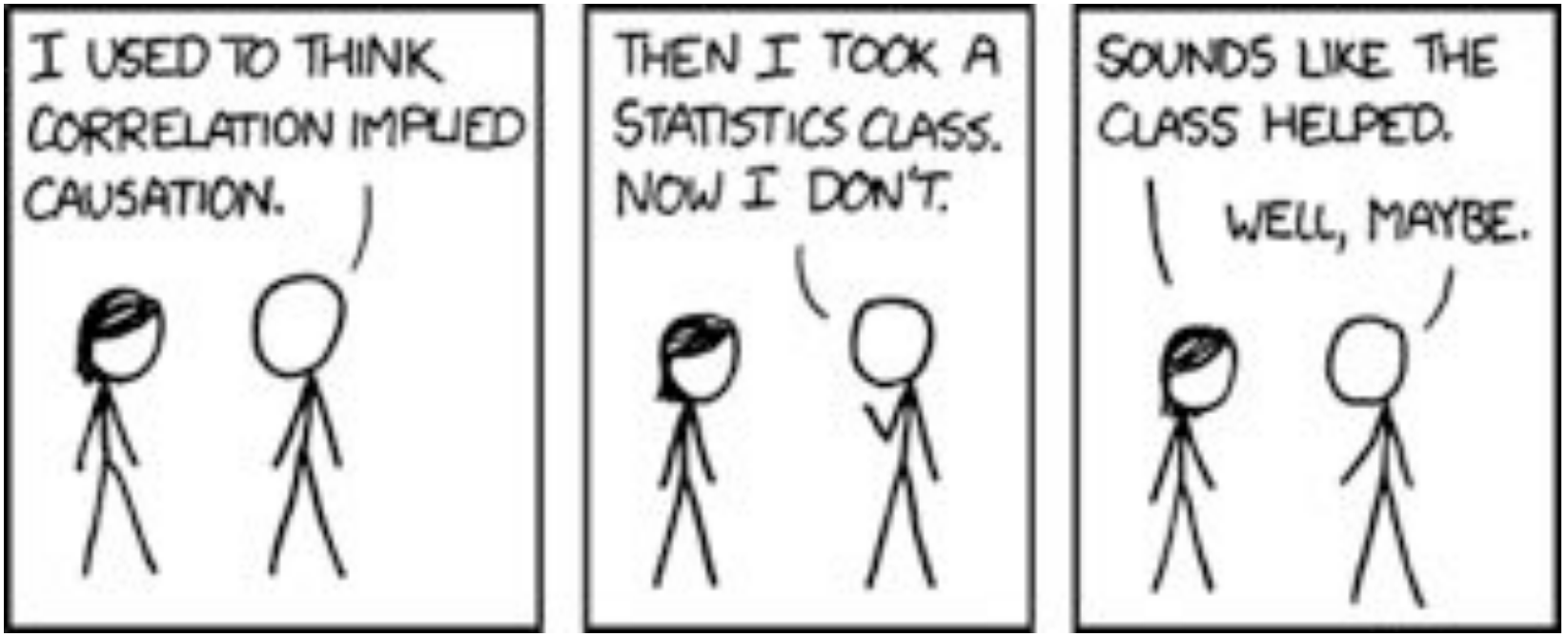
Understand how to unify data from disparate sources to build a larger picture of biological phenomena



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

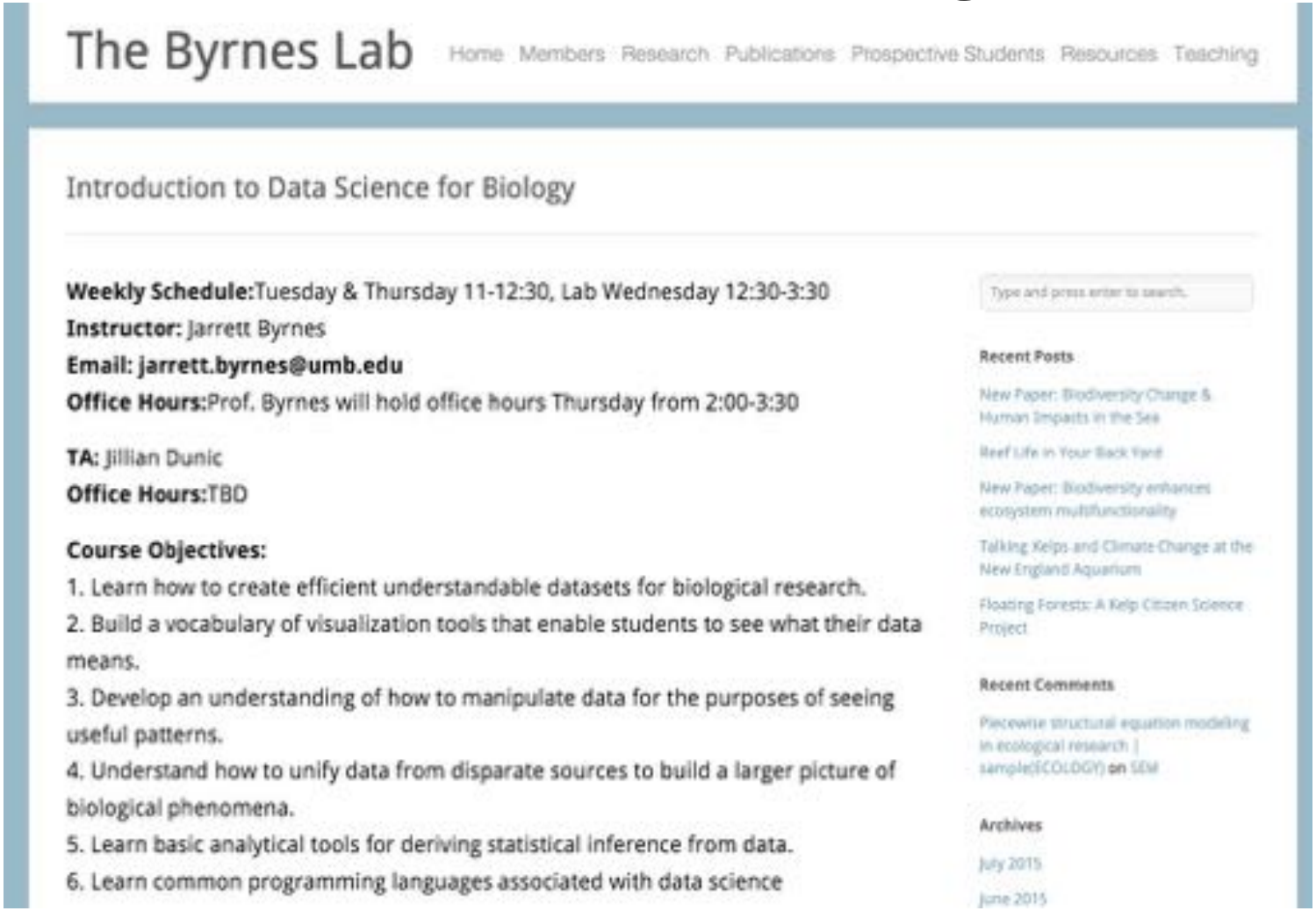


Learn basic analytical tools for deriving statistical inference from data



This Class

Course Web Page



http://byrneslab.net/teaching/data_science/

"Text"book & Weekly Readings

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

Lab

- Coding!
- TA: Jillian Dunic
- 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

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?

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

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

(And listed to the Not So Standard Deviations Podcast)