

This is the Way

*Integrating Open Data Science
Workflow & Software Carpentry into
the Statistical Ecology Classroom*



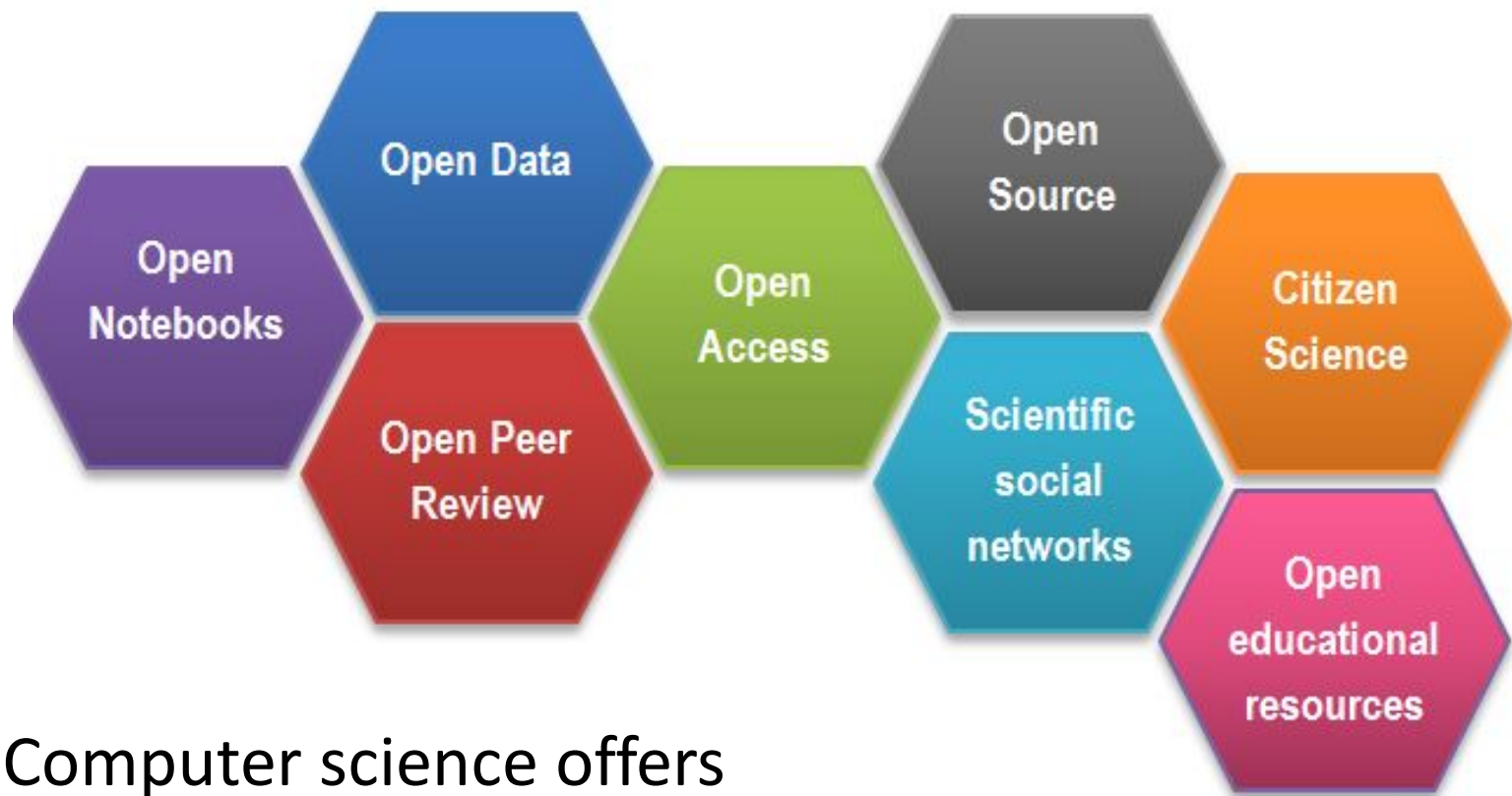
Gavin Fay

gfay@umassd.edu

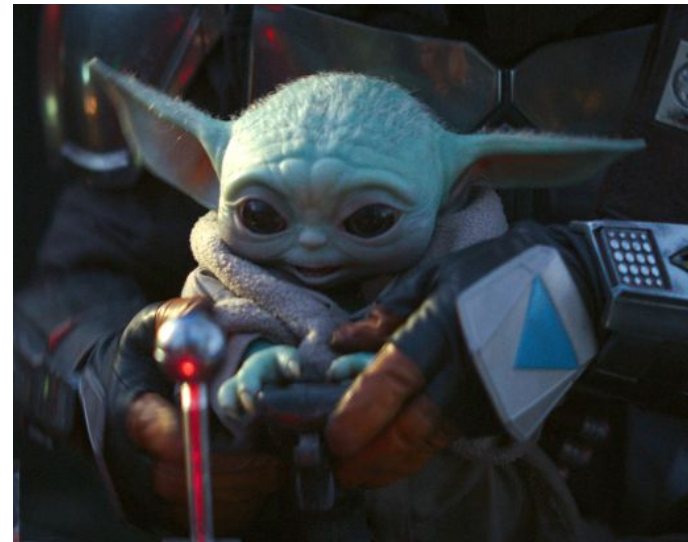
 @gavin_fay

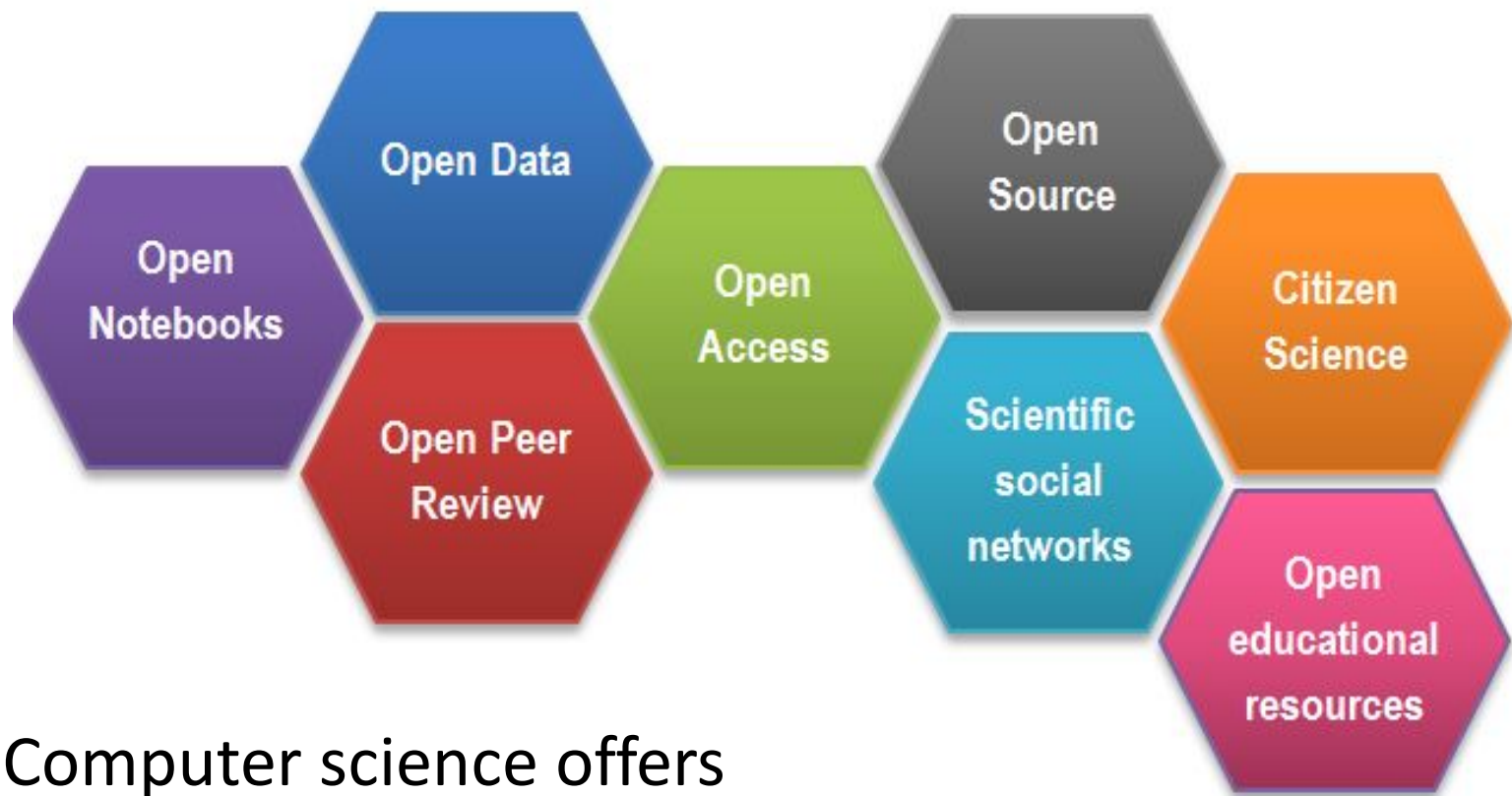
UMassD Teaching & Learning Conference

17 January 2020



Computer science offers many tools to facilitate shiny Open Science.





Computer science offers many tools to facilitate shiny Open Science.

But, few scientists are trained in their use.



Motivation #1



art: @allison_horst

[@julesquid](#)

nature
ecology & evolution

PERSPECTIVE

PUBLISHED: 23 MAY 2017 | VOLUME: 1 | ARTICLE NUMBER: 0160

Our path to better science in less time using open data science tools

Julia S. Stewart Lowndes^{1*}, Benjamin D. Best², Courtney Scarborough¹, Jamie C. Afflerbach¹,
Melanie R. Frazier¹, Casey C. O'Hara¹, Ning Jiang¹ and Benjamin S. Halpern^{1,3,4}

Source: Lowndes et al. 2017: ohi-science.org/betterscienceinlesstime

Motivation #2

PERSPECTIVE

Good enough practices in scientific computing

Greg Wilson^{1☯*}, **Jennifer Bryan**^{2☯}, **Karen Cranston**^{3☯}, **Justin Kitzes**^{4☯}, **Lex Nederbragt**^{5☯},
Tracy K. Teal^{6☯}

1 Software Carpentry Foundation, Austin, Texas, United States of America, **2** RStudio and Department of Statistics, University of British Columbia, Vancouver, British Columbia, Canada, **3** Department of Biology, Duke University, Durham, North Carolina, United States of America, **4** Energy and Resources Group, University of California, Berkeley, Berkeley, California, United States of America, **5** Centre for Ecological and Evolutionary Synthesis, University of Oslo, Oslo, Norway, **6** Data Carpentry, Davis, California, United States of America

☯ These authors contributed equally to this work.

* gvwilson@software-carpentry.org

<https://doi.org/10.1371/journal.pcbi.1005510>

Motivation #2

PERSPECTIVE

Good enough practices in scientific computing

¹*, Jennifer Bryan², Karen Cranston³, Justin Kitzes⁴, Lex Nederbragt⁵,
⁶

Software Carpentry Foundation, Austin, Texas, United States of America, **2** RStudio and Department of
University of British Columbia, Vancouver, British Columbia, Canada, **3** Department of Biology,
Durham, North Carolina, United States of America, **4** Energy and Resources Group,
California, Berkeley, Berkeley, California, United States of America, **5** Centre for Ecological and
Evolutionary Synthesis, University of Oslo, Oslo, Norway, **6** Data Carpentry, Davis, California, United States

* and ⁶ contributed equally to this work.
<https://software-carpentry.org>

<https://doi.org/10.1371/journal.pcbi.1005510>

Approximations to ‘Best’ can still be OK



GF2020

TheCarpentries.org



We teach foundational coding and data science skills to researchers worldwide.



What we do

The Carpentries teaches foundational cod-



Who we are

Our diverse, global community includes



Get involved

See all the [ways you can engage](#) with The

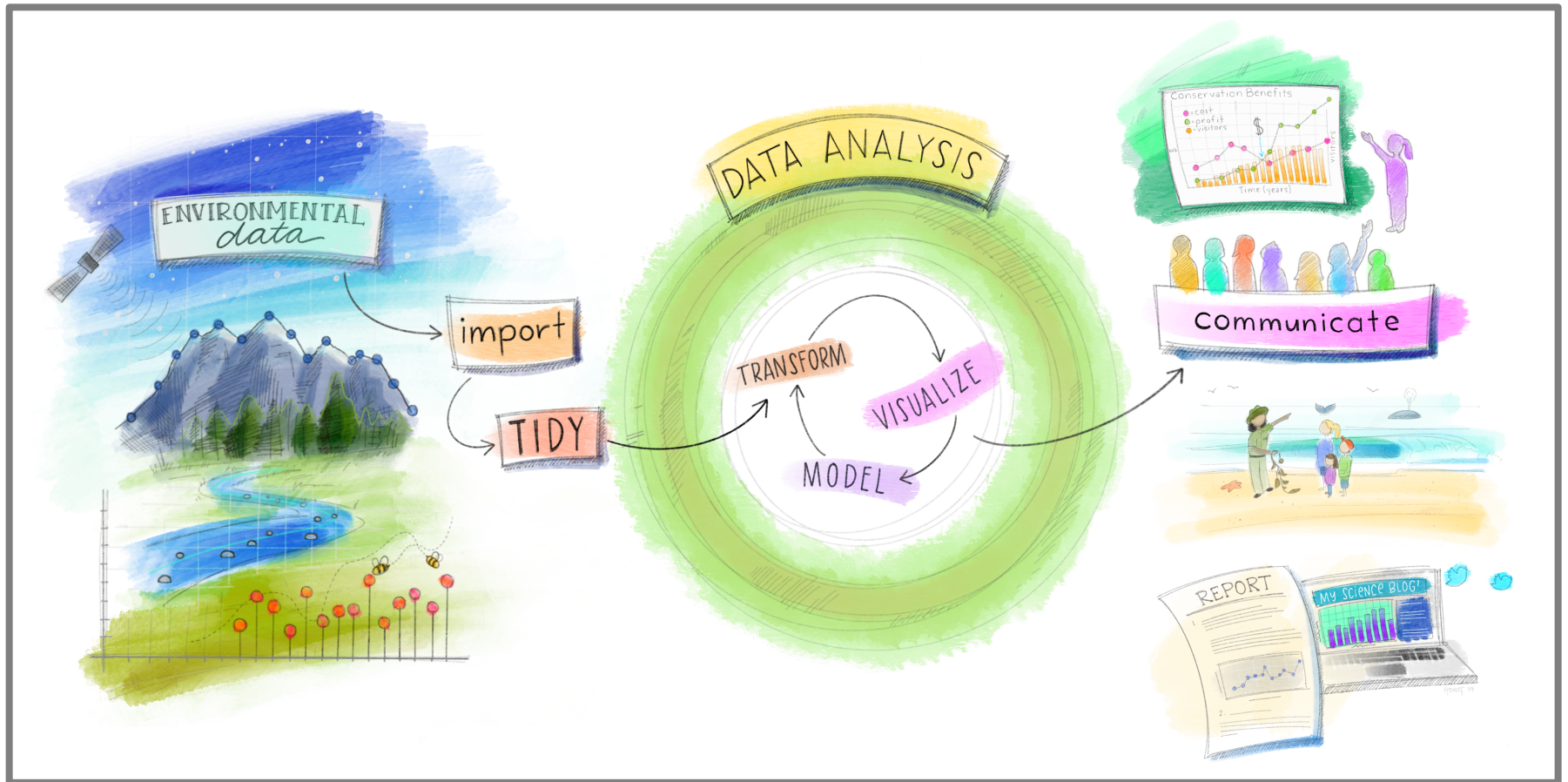
Why ?

- Reproducible Science!
- Full data science workflow
- Language & environment for statistical computing & graphics
- Open source & Free
- Lots of scientists use it!
- AMAZING online community
- Works well with other tools



Artwork by [@allison_horst](https://twitter.com/allison_horst)

Why ?



Why ?

- Reproducible Science!
- Full data science workflow
- Language & environment for statistical computing & graphics
- Open source & Free
- Lots of scientists use it!
- AMAZING online community
- Works well with other tools



Artwork by [@allison_horst](https://twitter.com/allison_horst)

MAR 536: Biological Statistics II

Statistical analysis for biological science graduate students

Computer labs: Intro to statistical analysis in R

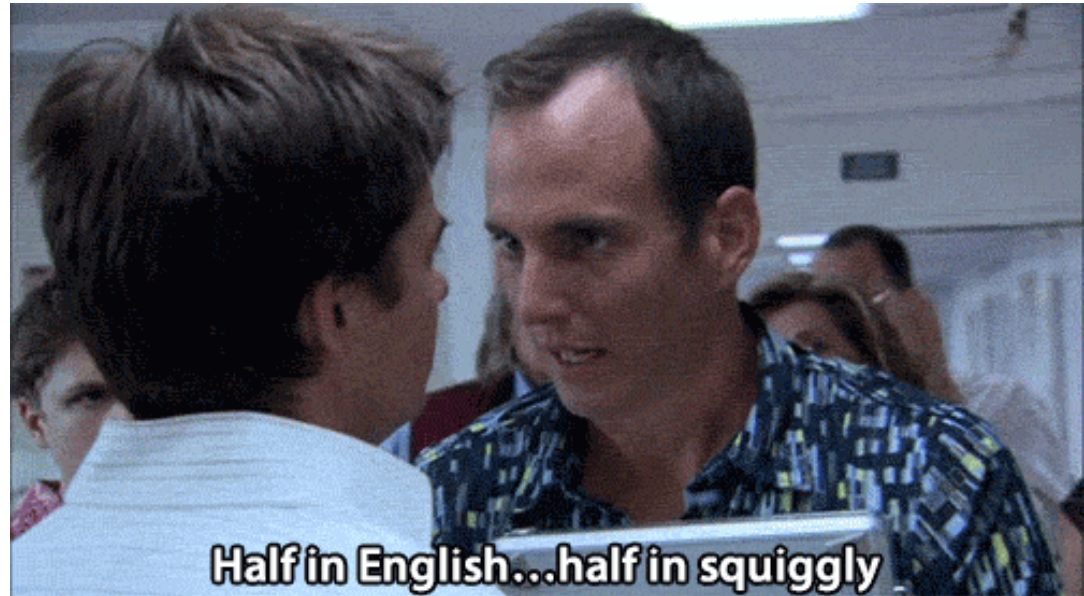
- Partial introduction of tidyverse functions in 2018

Extra credit for using R Markdown in assignments

- 1 student did this in 2017
- All but 1 student did it in 2018
- *Spring 2020 ???*

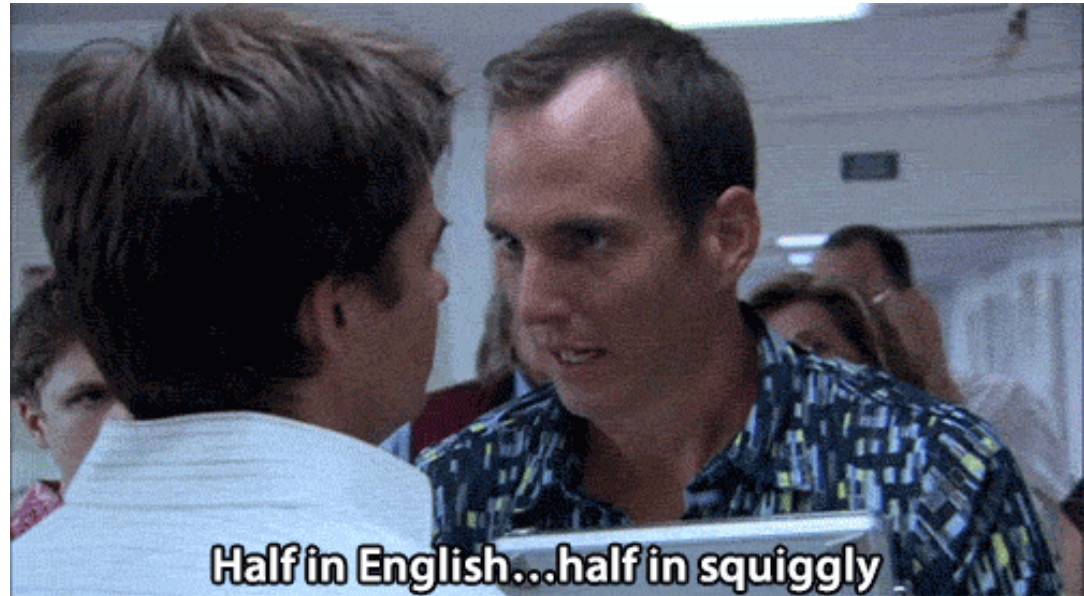


Readable code
that remains
consistent
across tasks



```
5  
6 fishdata %>%  
7   group_by(year, season, species) %>%  
8   summarise(avg_numbers = mean(abundance),  
9             max_depth = max(depth),  
10            avg_temp = mean(surftemp))  
11
```


Readable code
that remains
consistent
across tasks



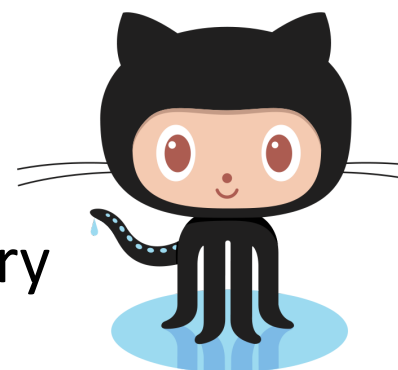
```
5  
6 fishdata %>% "and then"  
7   group_by(year, season, species) %>%  
8   summarise(avg_numbers = mean(abundance),  
9             max_depth = max(depth),  
10            avg_temp = mean(surftemp))  
11
```

MAR 580: Advanced Population Modeling

Fitting ecological models in R & Template Model Builder

2015

- separate lectures & computer labs
- many lab assignments
- course materials shared through github repository

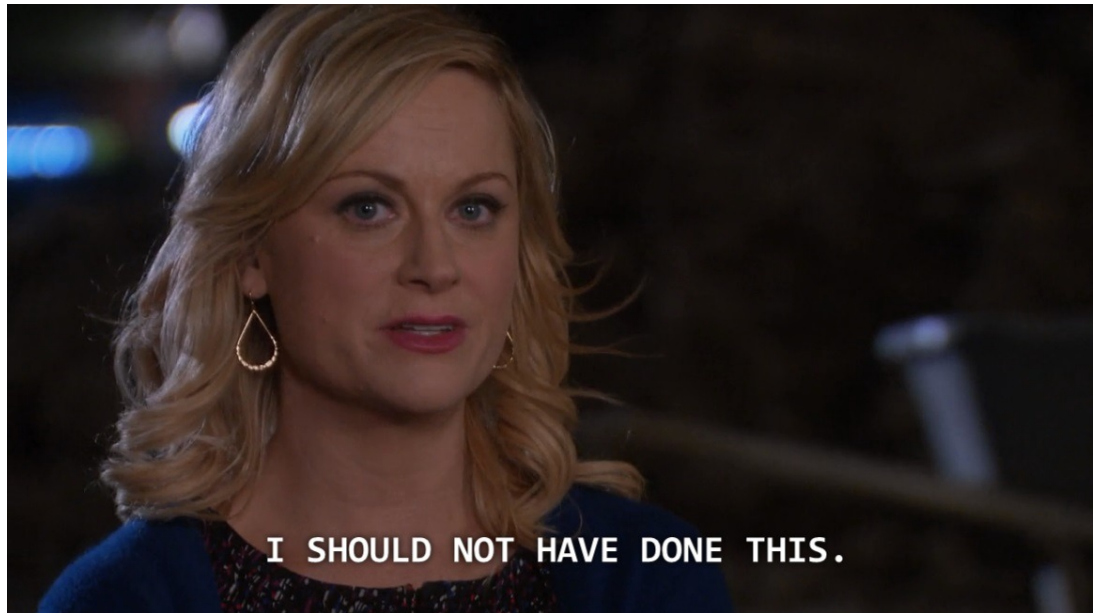


MAR 580: Advanced Population Modeling

Fitting ecological models in R & Template Model Builder

2015

- separate lectures & computer labs
- many lab assignments
- course materials shared through github repository



MAR 580: Advanced Population Modeling

Fitting ecological models in R & Template Model Builder

2015

- separate lectures & computer labs
- many lab assignments
- course materials shared through github repository

2019

- students using R Markdown for assignments
- live coding during mixed lab/lectures
- AirMedia to share student screens to class: debugging aid
- course materials shared via Google Drive

#quantfish woRkshops



#quantfish woRkshops

tutorials for beginner and intermediate R users

- Students & postdocs lead 1.5 hr sessions
- Live coding
- Learning R by doing useful things straight away
- Less is more
- Sharing of materials via GoogleDocs
- Materials version-controlled using git and github
- Feedback asked for (& acted on) often



A GoogleDoc for each workshop....

Search Drive

?

⚙

⋮

My Drive > quantfish

🔗

👤

👁

🗑

⋮

📁

ℹ

📅 31

💡

📝

Name ↑	Owner	Last modified	File size
📁 data	me	Oct 2, 2018 me	—
📁 Quantfish_worKshop_07_mapping_ggplot	me	5:09 PM me	—
📄 Quantfish worRkshop 1 👤	me	Feb 13, 2019 me	—
📄 Quantfish worRkshop 02: data basics 👤	me	Feb 13, 2019 me	—
📄 Quantfish worRkshop 03: data wrangling 👤	me	Feb 13, 2019 me	—
📄 Quantfish worRkshop 04: data transformation 👤	me	Dec 5, 2018 me	—
📄 Quantfish worRkshop 06: Data Visualization 👤	me	Feb 13, 2019 me	—
📄 Quantfish worRkshop 08: Ask an expert group work 👤	me	Mar 19, 2019 me	—
📄 Quantfish worRkshop 09: mapping with ggplot2 👤	me	Apr 23, 2019 me	—

quantfishwoRkshop Introduction to R

November 13 2018

Rstudio interface

<https://rstudio.cloud/project/134344>

Materials (continuously updated)

https://github.com/thefaylab/quantfishR_01_introR

R for data science (online book): <https://r4ds.had.co.nz/>

Sign-in sheet (name, institution, email)

- Gavin Fay (UMassD), gfay@umassd.edu
- Margot Wilderman (UMassD), mwilderman@umassd.edu
- Danielle Lavoie (UMassD), davoie1@umassd.edu
- Megan Winton (UMassD), mwinton@umassd.edu
- Renee Halloran (UMassD), rhalloran@umassd.edu
- Greg DeCelles (MDMF), gregory.decelles@mass.gov
- Nicole Danaher-Garcia (UMassD), ngarcia3@umassd.edu
- Flynn Casey (UMassD), fcasey1@umassd.edu
- Seth Larson (UMassD), slarson1@umassd.edu
- Jonathan Cummings (UMassD), jcumming@umassd.edu
- Debra Duarte (UMassD), dduarte@umassd.edu
- Susan Inis (UMassD), sinis@umassd.edu
- Cheng Liu (UMassD), cliu2@umassd.edu
- Alex Hansell (UMassD), ahansell@umassd.edu
- Brooke Wright (UMassD), brooke.wright@umassd.edu
- Tammy Silva (UMassD), tsilva4@umassd.edu
- Robert Wildermuth (UMassD), rwildermuth@umassd.edu
- Dave Martins (MA DMF), dave.martins@mass.gov
- Marjorie Lysakatos (UMassD), mlysakatos@umassd.edu
- Harriet Booth (MA DMF), harriet.booth@mass.gov

Topics you would like to see

- More on tidyverse
- Ggplot
- Spatial analysis
- Time series analysis

Standard environment: RStudio & RStudio Cloud

quantfish - Google Drive | Quantfish worKshop 1 - Google | RStudio Cloud | GitHub - thefaylab/quantfishR

https://rstudio.cloud/project/134344

Studio Cloud

Spaces

Your Workspace

New Space

Learn

Guide

What's New

Primers

Cheat Sheets

Feedback and Questions

Info

Terms and Conditions

System Status

Your Workspace / quantfishR_01_introR Gavin Fay

TEMPORARY PROJECT

Save a Permanent Copy

Bob boblaw

File Edit Code View Plots Session Build Debug Profile Tools Help

Go to file/function

Addins

R 3.5.3

```
18 # here we create a new variable 'means' that contains the mean abundance
19 mean_abundance <- summarise(grouped_data, means=mean(abundance))
20 mean_abundance <- summarise(grouped_data, means=mean(abundance, na.rm = TRUE)) #t
21
22 # print our new summarised object
23 mean_abundance
24
25 ## plot the mean abundances over time, by season (color) & species (panels)
26 ggplot(mean_abundance, aes(x=year, y=means, color=season)) + #sets up plot, maps
27   geom_point() + #produces the scatterplot
28   facet_wrap(~comname, scales = "free") + #adds panels (scales argument makes y
29   geom_smooth(method="loess") #adds the smoother/trend line
30
31
32 #####
33 ## map the abundance of species
34 #####
35
36
```

45:1 (Top Level) R Script

Console

```
/cloud/project/
6 2001 SPRING ATLANTIC COD 27.4
7 2001 SPRING BLACK SEA BASS 16.0
8 2001 SPRING SCUP 43.2
9 2001 SPRING SILVER HAKE 1031.
10 2001 SPRING SPINY DOGFISH 548.
# ... with 135 more rows
>
> ## plot the mean abundances over time, by season (color) & species (panels)
> ggplot(mean_abundance, aes(x=year, y=means, color=season)) + #sets up plot, maps variables to plot aesthetics
+ geom_point() + #produces the scatterplot
+ facet_wrap(~comname, scales = "free") + #adds panels (scales argument makes y axes separate by species)
+ geom_smooth(method="loess") #adds the smoother/trend line
>
```

Environment

History

Connections

Import Dataset

Global Environment

Data

Variable	Observations
grouped_data	43865 obs. of 11 variables
mean_abundance	145 obs. of 4 variables
mydata	43865 obs. of 11 variables

Files

Plots

Packages

Help

Viewer

Zoom

Export

Publish

ATLANTIC COD

BLACK SEA BASS

SCUP

SILVER HAKE

SPINY DOGFISH

season

FALL

SPRING

means

year

Code online in github repository

Auto-updated during workshop

The screenshot shows a web browser with two tabs: 'Quantfish worRkshop 1 - Google' and 'GitHub - thefaylab/quantfishR_01_introR'. The address bar shows the URL 'https://github.com/thefaylab/quantfishR_01_introR'. The GitHub navigation bar is visible with links for 'Why GitHub?', 'Enterprise', 'Explore', 'Marketplace', 'Pricing', a search bar, and 'Sign in'/'Sign up' buttons. The repository page for 'thefaylab/quantfishR_01_introR' is displayed, showing '2' watches, '0' stars, and '0' forks. A 'Code' button is highlighted in the repository navigation. A 'Join GitHub today' banner is present, stating 'GitHub is home to over 40 million developers working together to host and review code, manage projects, and build software together.' Below the banner, the repository name 'Introduction to R quantfishworkshop' is shown. A summary bar indicates '34 commits', '1 branch', '0 packages', '0 releases', and '2 contributors'. A 'Branch: master' dropdown and a 'New pull request' button are visible. A 'Find file' button and a 'Clone or download' button are also present. A list of files is shown, including '01_Intro.pptx', 'IntroR.R', 'README.md', 'neus_bts.csv', and 'neus_bts.xlsx', each with a description of the commit and the date of the last commit. The latest commit is '80cf65e on Nov 15, 2018'.

thefaylab / quantfishR_01_introR

Watch 2 Star 0 Fork 0

<> Code Issues 0 Pull requests 0 Projects 0 Security Insights

Join GitHub today Dismiss

GitHub is home to over 40 million developers working together to host and review code, manage projects, and build software together.

[Sign up](#)

Introduction to R quantfishworkshop

34 commits 1 branch 0 packages 0 releases 2 contributors

Branch: master New pull request Find file Clone or download

gavinfay final R script Latest commit 80cf65e on Nov 15, 2018

01_Intro.pptx	adds intro slides	last year
IntroR.R	final R script	last year
README.md	update new commands	last year
neus_bts.csv	add datafiles	last year
neus_bts.xlsx	add datafiles	last year

What's next?

More conversion of MAR 536 R labs to the tidyverse.
Course Management using R Studio Cloud / github
rstudio::conf

Tuesday Feb 25

Special Seminar at UMassD-SMAST

“R and teamwork for better science in less time”

Dr. Julia Stewart Lowndes, NCEAS



Thank you!

gfoy@umassd.edu

thefaylab.com

 [@gavin_fay](https://twitter.com/gavin_fay)

To be added to
#quantfish email list:
anovak@umassd.edu

These slides:

bit.ly/fay_tlearnconf2020



“This is the best! The kind of intro content I’ve been looking for. I’m really happy this exists!”

any tips for how
to not panic / give
up
when things
don't work?

Larger text
on screen

THIS IS THE
BEST! THE KIND
OF INTRO CONTENT
I'VE BEEN LOOKING FOR
I'M REALLY HAPPY
THIS EXISTS!
THANKS!

Very easy to follow
and practical &
useful functions

Maybe we ~~we~~ need 1.5hrs
It seems like there is
more material than can
fit into the time.

If possible, let us
know beforehand which
packages we need
(my computer is very
slow)

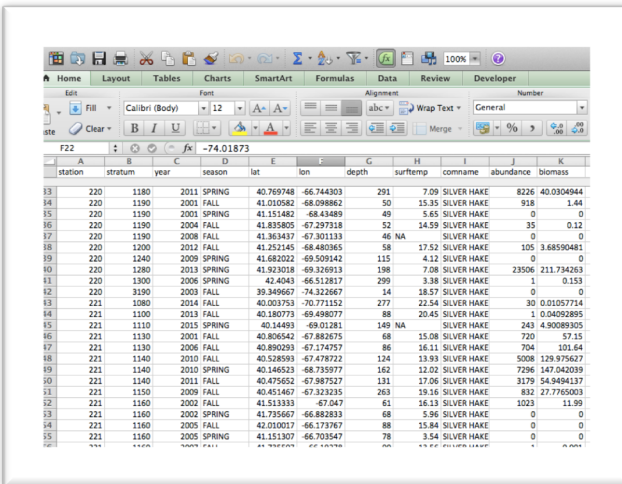
Very helpful to have
the code written as
we go

I liked
I liked the
approach & making
people feel
comfortable & not
intimidated.
• I liked the first
data files easy for
us to understand.

“... let us know beforehand which packages we need (my computer is very slow)”

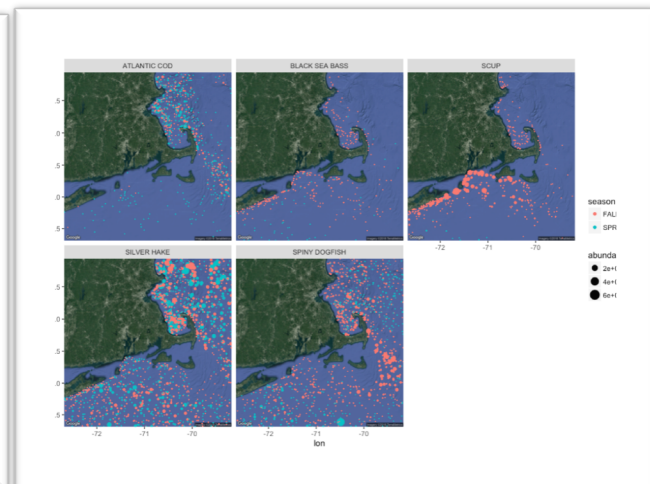
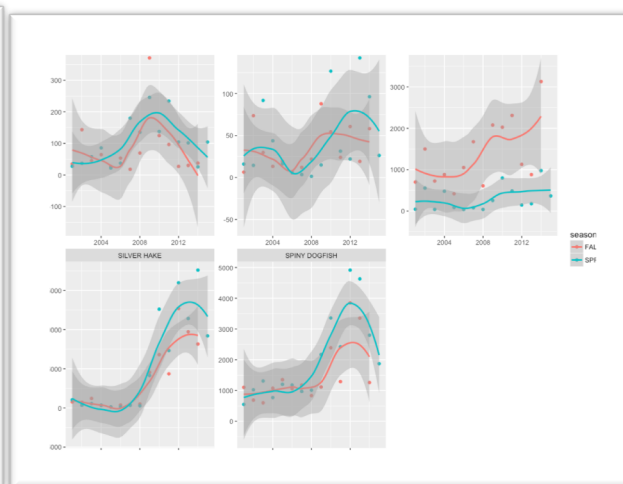
Intro to R: Analyze US fish data

- Take data from spreadsheet to visualization
- Data wrangling
- Summarizing data by species over time
- Mapping of fish distributions



A screenshot of an Excel spreadsheet with a green ribbon at the top. The spreadsheet contains a table of fish data with columns for station, stratum, year, season, lat, lon, depth, surftemp, comname, abundance, and biomass. The data is organized into rows, with some rows highlighted in yellow. The formula bar at the top shows the formula =74.01873.

station	stratum	year	season	lat	lon	depth	surftemp	comname	abundance	biomass
13	220	1180	2011 SPRING	40.769748	-66.744303	291	7.09	SILVER HAKE	8226	40.0304944
14	220	1190	2001 FALL	41.010582	-68.098862	50	15.35	SILVER HAKE	918	1.44
15	220	1190	2001 SPRING	41.151482	-68.43489	49	5.65	SILVER HAKE	0	0
16	220	1190	2004 FALL	41.815805	-67.297318	52	14.59	SILVER HAKE	35	0.12
17	220	1190	2008 FALL	41.363437	-67.301133	46	NA	SILVER HAKE	0	0
18	220	1200	2012 FALL	41.252145	-68.480365	58	17.52	SILVER HAKE	105	3.68590481
19	220	1240	2009 SPRING	41.682022	-69.509142	115	4.12	SILVER HAKE	0	0
20	220	1280	2013 SPRING	41.923018	-69.326913	198	7.08	SILVER HAKE	23506	211.734263
21	220	1300	2006 SPRING	42.4043	-66.512817	299	3.38	SILVER HAKE	1	0.153
22	220	3190	2003 FALL	39.349667	-74.322667	14	18.57	SILVER HAKE	0	0
23	221	1080	2014 FALL	40.003753	-70.771152	277	22.54	SILVER HAKE	30	0.01057714
24	221	1100	2013 FALL	40.180773	-69.498077	88	20.45	SILVER HAKE	1	0.04092895
25	221	1110	2015 SPRING	40.14493	-69.01281	149	NA	SILVER HAKE	243	4.90089305
26	221	1130	2001 FALL	40.806542	-67.882675	68	15.08	SILVER HAKE	720	57.15
27	221	1130	2006 FALL	40.890293	-67.174757	86	16.11	SILVER HAKE	704	101.64
28	221	1140	2010 FALL	40.528593	-67.478722	124	13.93	SILVER HAKE	5008	129.975627
29	221	1140	2010 SPRING	40.146513	-68.719577	162	12.02	SILVER HAKE	7296	147.042039
30	221	1140	2011 FALL	40.475652	-67.987527	131	17.06	SILVER HAKE	3179	54.9494137
31	221	1150	2009 FALL	40.451467	-67.323235	263	19.16	SILVER HAKE	832	27.7765003
32	221	1160	2002 FALL	41.513333	-67.047	61	16.13	SILVER HAKE	1023	11.59
33	221	1160	2002 SPRING	41.735667	-66.882833	68	5.96	SILVER HAKE	0	0
34	221	1160	2005 FALL	42.010017	-66.173767	88	15.84	SILVER HAKE	0	0
35	221	1160	2005 SPRING	41.151307	-66.703547	78	3.54	SILVER HAKE	0	0



My courses that use R



MAR 536: Biological Statistics II

- *statistical analysis for biological science graduate students*

MAR 580: Advanced Population Modeling

- *fitting ecological models to data*

Quantfish WoRkshops

- *tutorials for beginner and intermediate R users*

MAR 338: Ecological and Environmental Data Analysis in R

- *coming 2021 ?*

Rmarkdown

TEXT.CODE.OUTPUT.
(GET IT TOGETHER, PEOPLE.)



Artwork by [@allison_horst](https://twitter.com/allison_horst)

<https://vimeo.com/178485416>