This is the Way

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

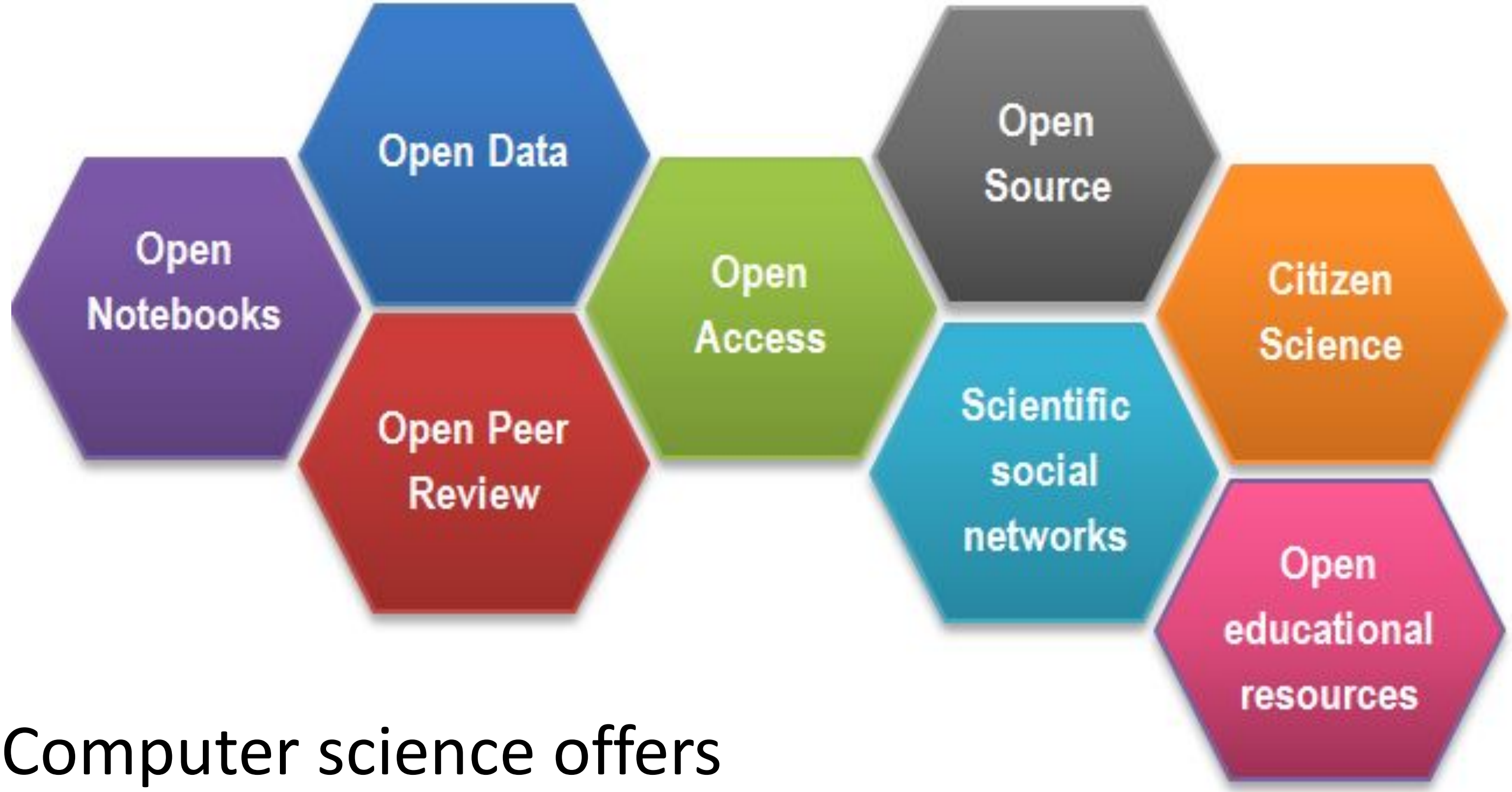
Gavin Fay

gfay@umassd.edu

@gavin_fay

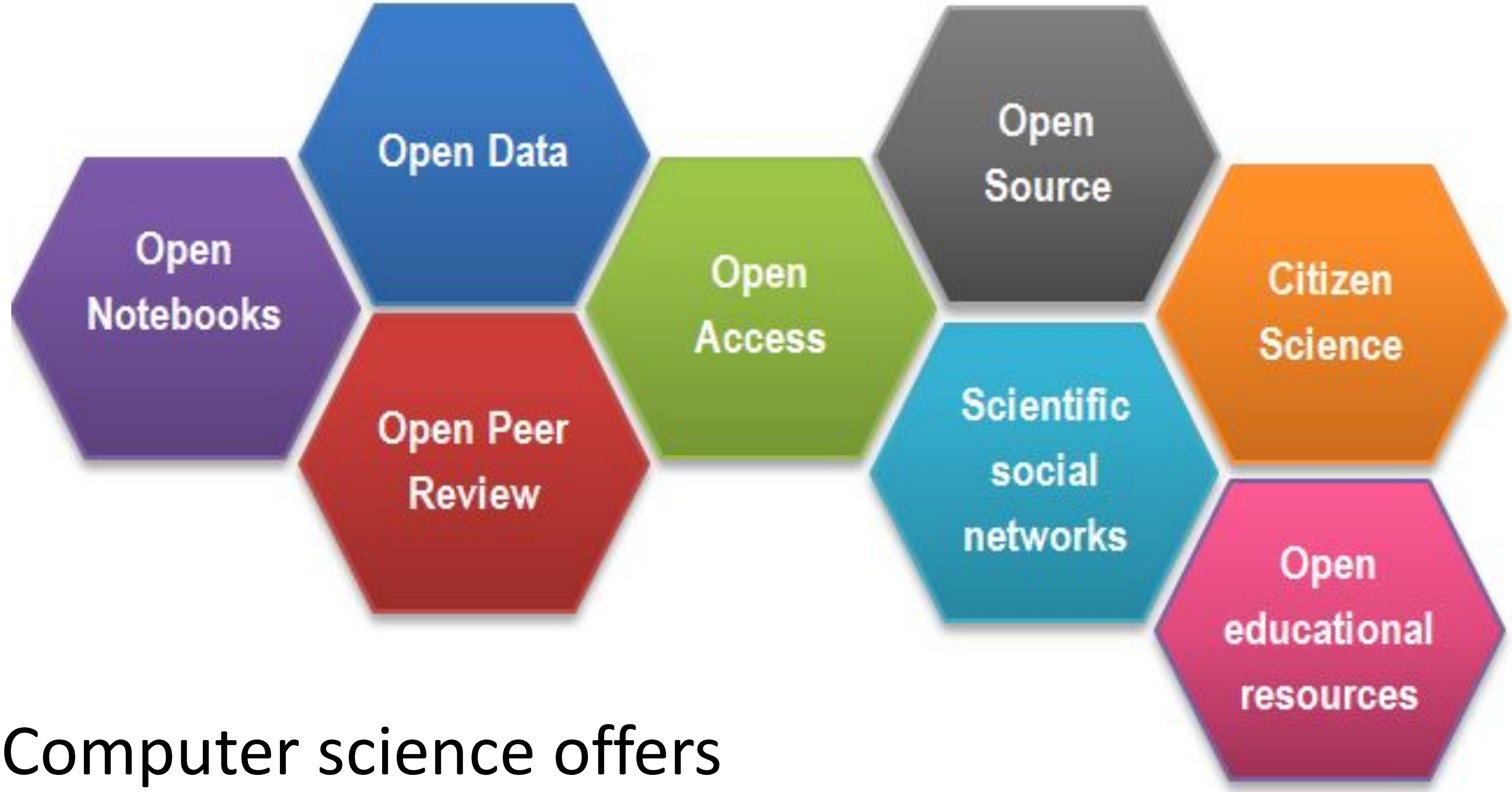
UMassD Teaching & Learning Conference

17 January 2020



Computer science offers many tools to facilitate shiny Open Science.

FOSTERopenscience.eu



Open Notebooks

Open Data

Open Peer Review

Open Access

Open Source

Scientific social networks

Citizen Science

Open educational resources

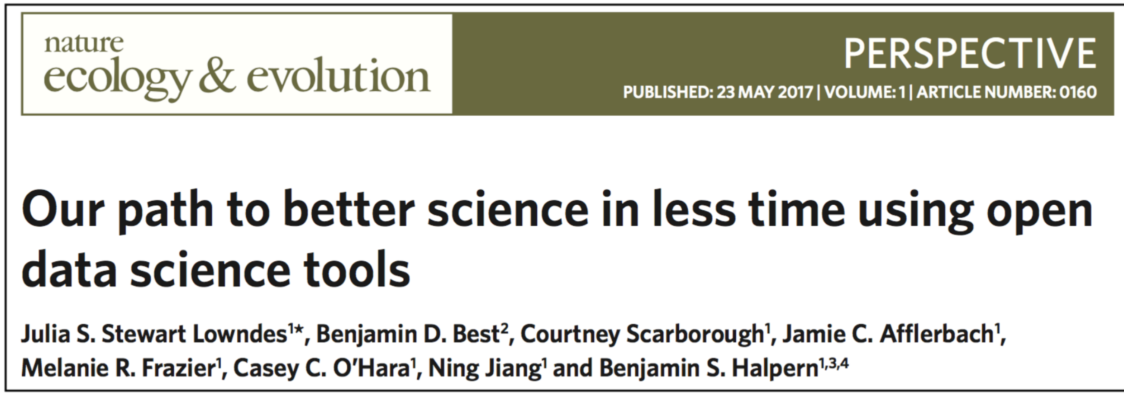
Computer science offers many tools to facilitate shiny Open Science.

But, few scientists are trained in their use.

FOSTERopenscience.eu

Motivation #1

art: @allison_horst



nature ecology & evolution

PERSPECTIVE

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

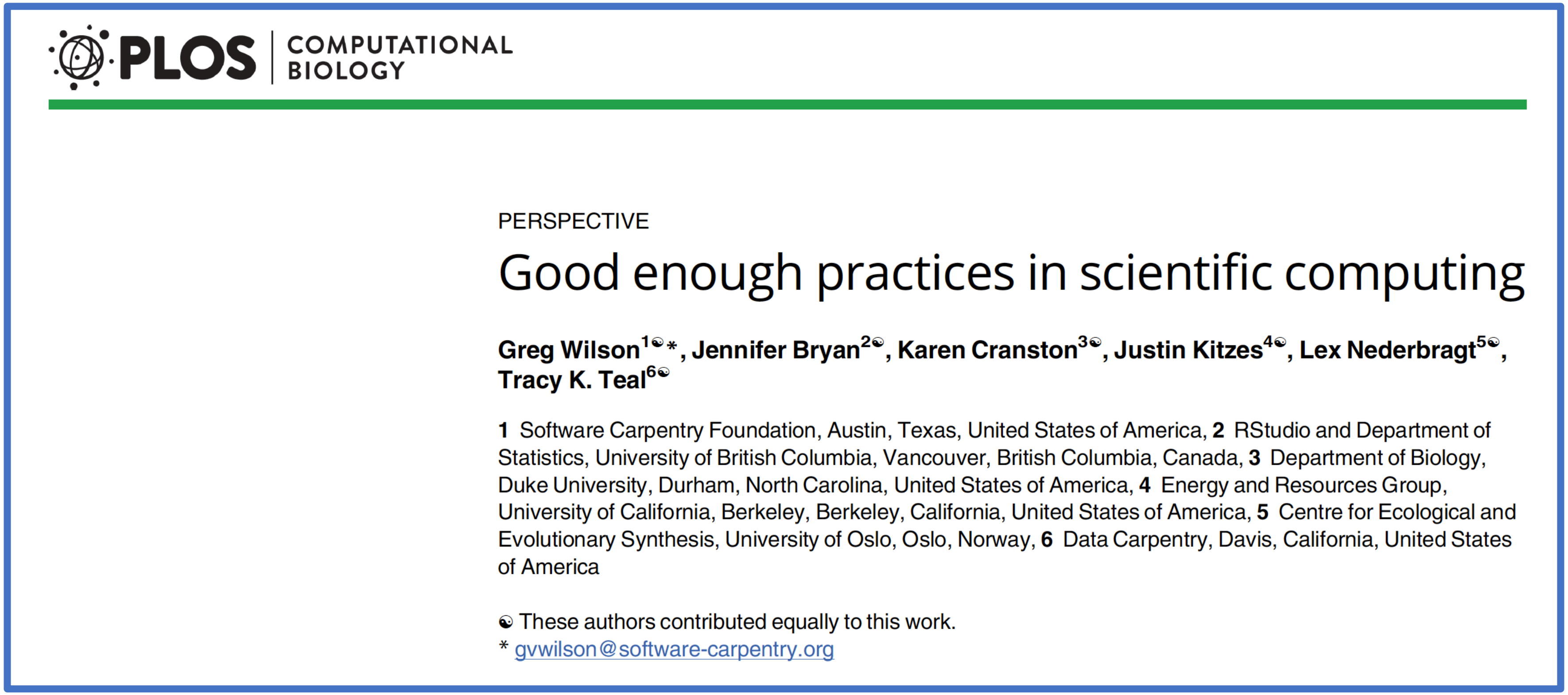
Our path to better science in less time using open data science tools

Julia S. Stewart Lowndes[1*], Benjamin D. Best[2], Courtney Scarborough[1], Jamie C. Afflerbach[1], Melanie R. Frazier[1], Casey C. O'Hara[1], Ning Jiang[1] and Benjamin S. Halpern[1,3,4]

@juliesquid

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

Motivation #2



PLOS | COMPUTATIONAL BIOLOGY

PERSPECTIVE

Good enough practices in scientific computing

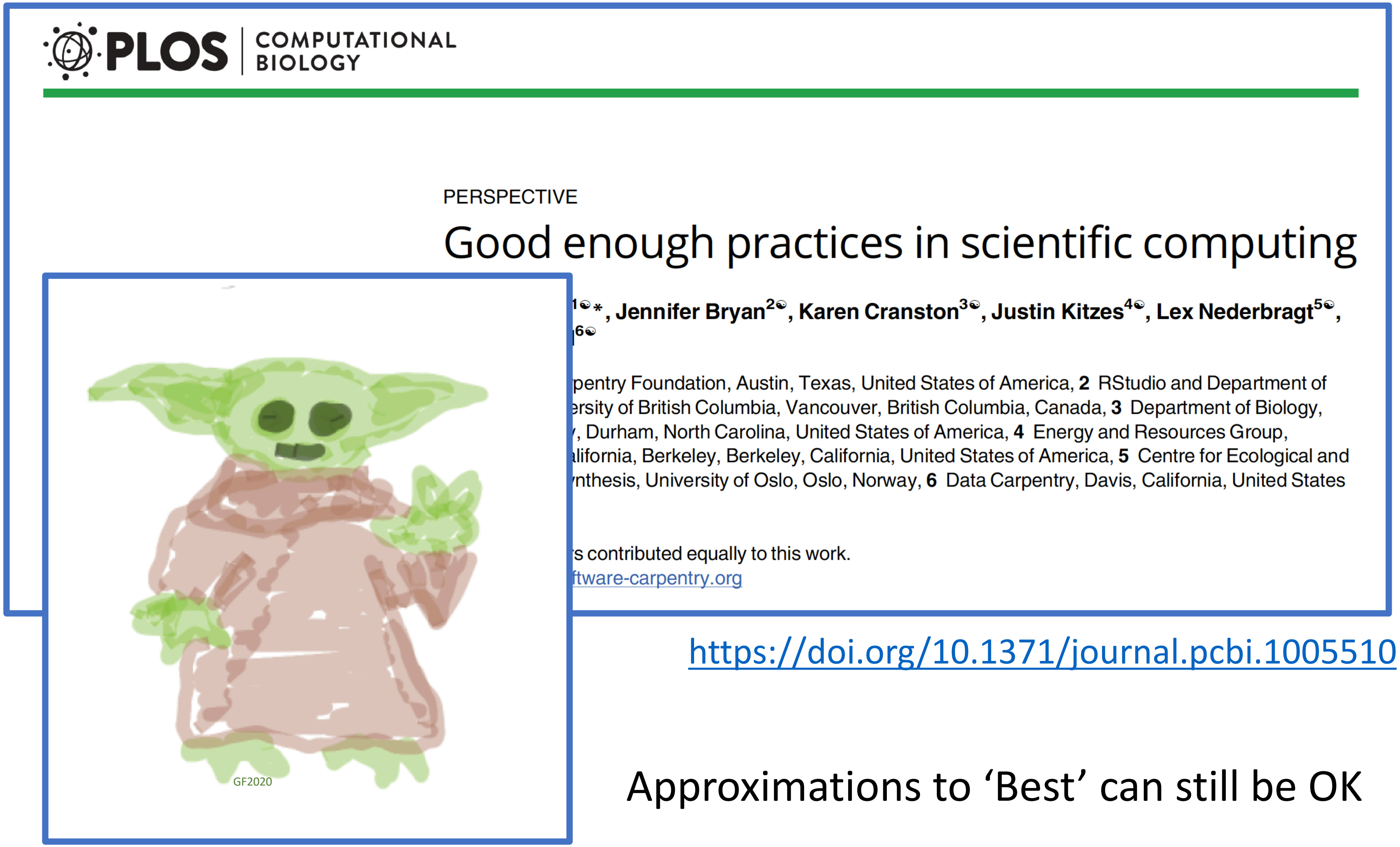
Greg Wilson[1☯]*, Jennifer Bryan[2☯], Karen Cranston[3☯], Justin Kitzes[4☯], Lex Nederbragt[5☯], Tracy K. Teal[6☯]

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



PLOS COMPUTATIONAL BIOLOGY

PERSPECTIVE

Good enough practices in scientific computing

[1]☯*, Jennifer Bryan[2]☯, Karen Cranston[3]☯, Justin Kitzes[4]☯, Lex Nederbragt[5]☯,
[6]☯

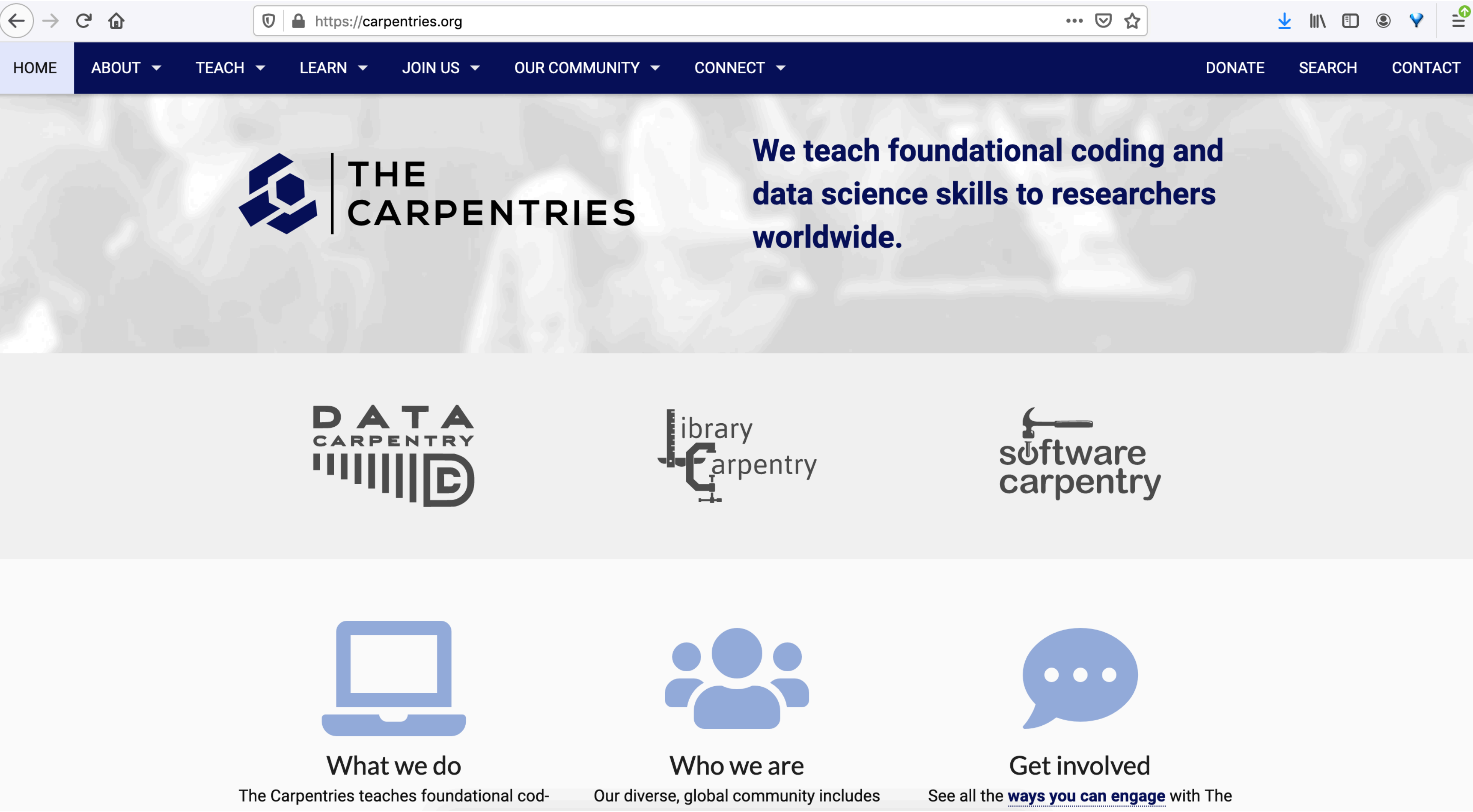
pentry Foundation, Austin, Texas, United States of America, **2** RStudio and Department of
ersity of British Columbia, Vancouver, British Columbia, Canada, **3** Department of Biology,
, Durham, North Carolina, United States of America, **4** Energy and Resources Group,
alifornia, Berkeley, Berkeley,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

s contributed equally to this work.
ftware-carpentry.org

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

Approximations to ‘Best’ can still be OK

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

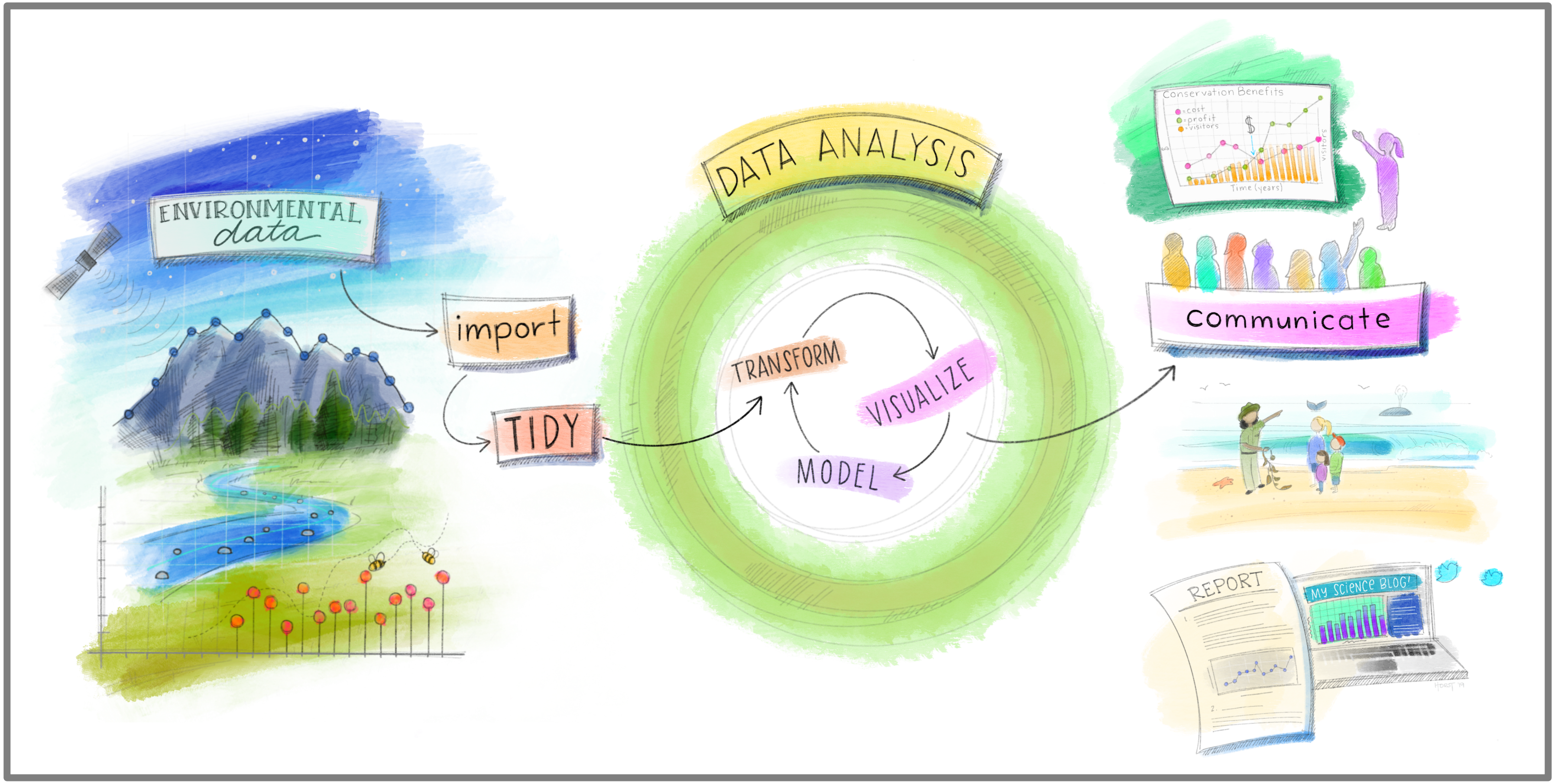
Why R ?

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

Artwork by @allison_horst

RStudio

Why R ?



Artwork by @allison_horst

RStudio

Why R ?

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

Artwork by @allison_horst

RStudio

MAR 536: Biological Statistics II

Statistical analysis for biological science graduate students

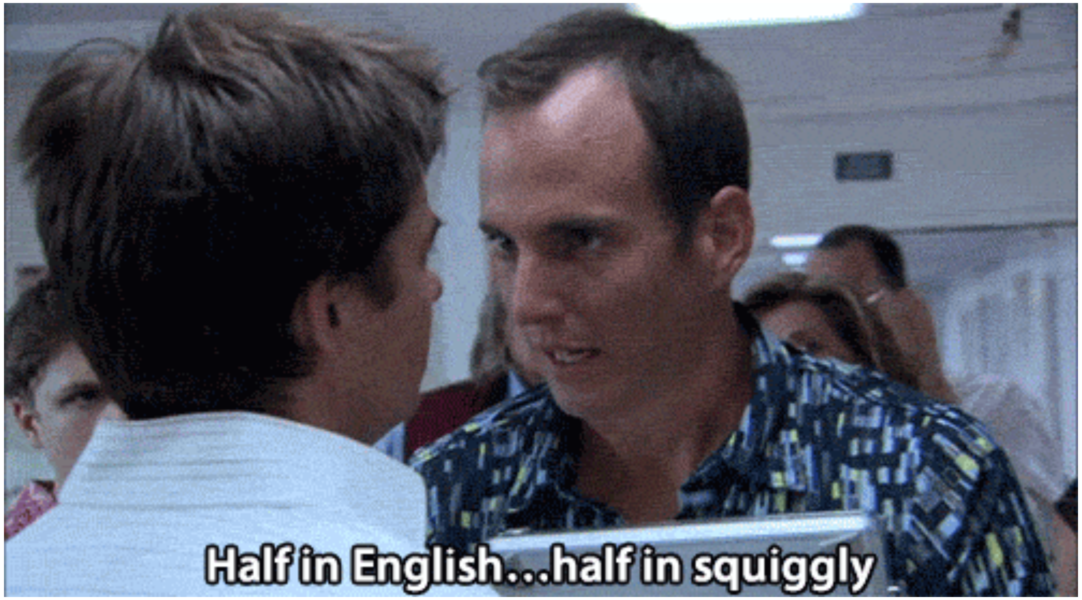
Computer labs: Intro to statistical analysis in R

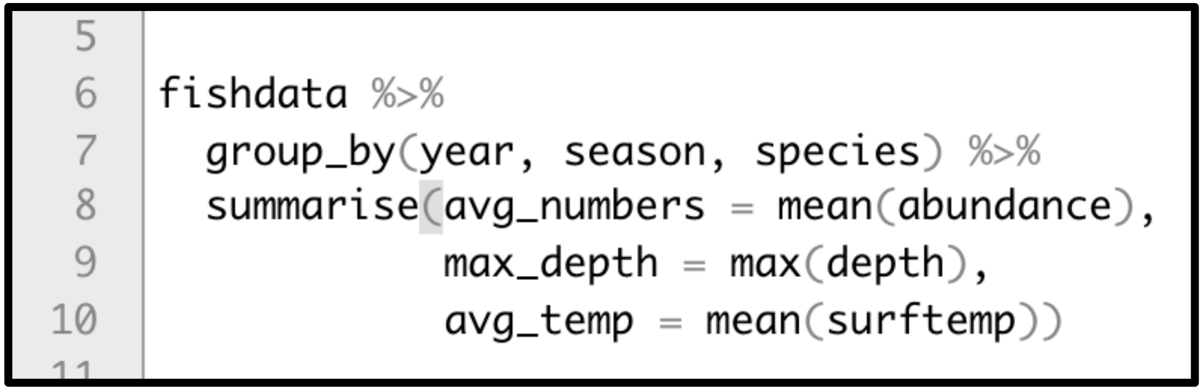
- Partial introduction of tidyverse functions in 2018

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

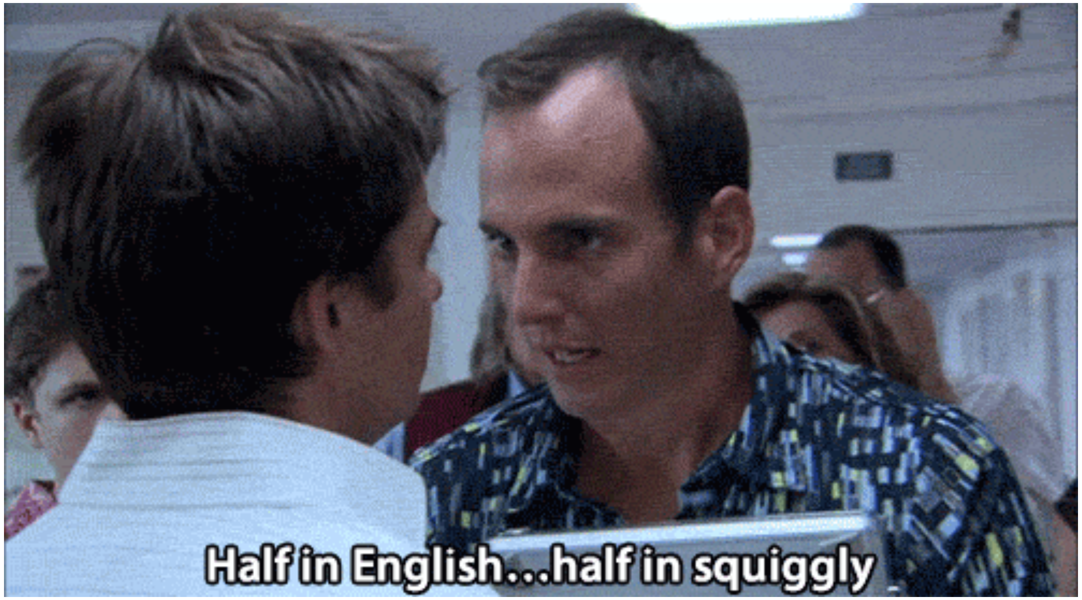




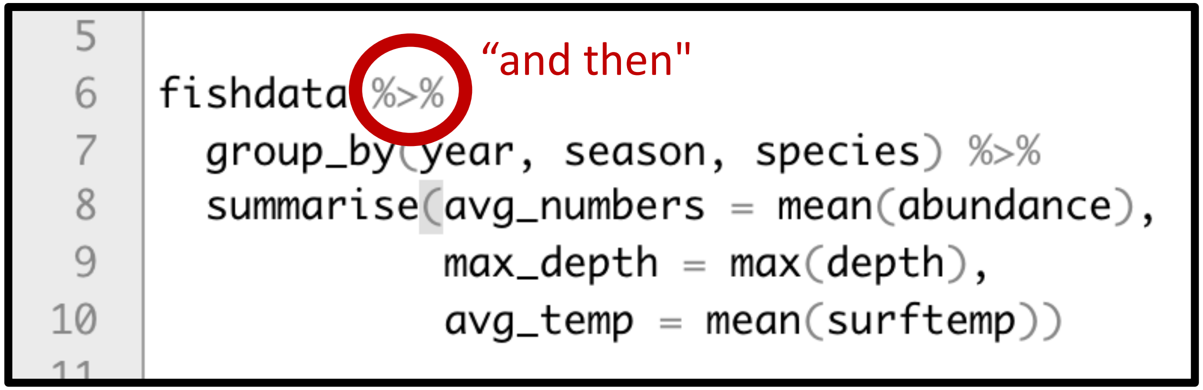
```
fishdata %>%
  group_by(year, season, species) %>%
  summarise(avg_numbers = mean(abundance),
            max_depth = max(depth),
            avg_temp = mean(surftemp))
```

HT @dataandme

Readable code that remains consistent across tasks



tidyverse



```
fishdata %>%
  group_by(year, season, species) %>%
  summarise(avg_numbers = mean(abundance),
            max_depth = max(depth),
            avg_temp = mean(surftemp))
```

“and then"

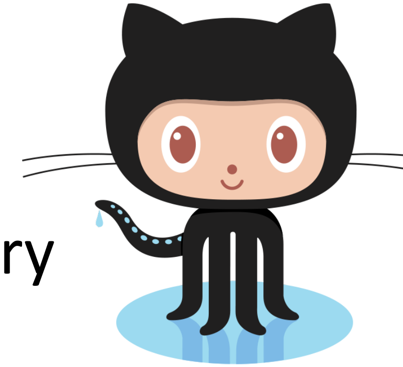
HT @dataandme

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

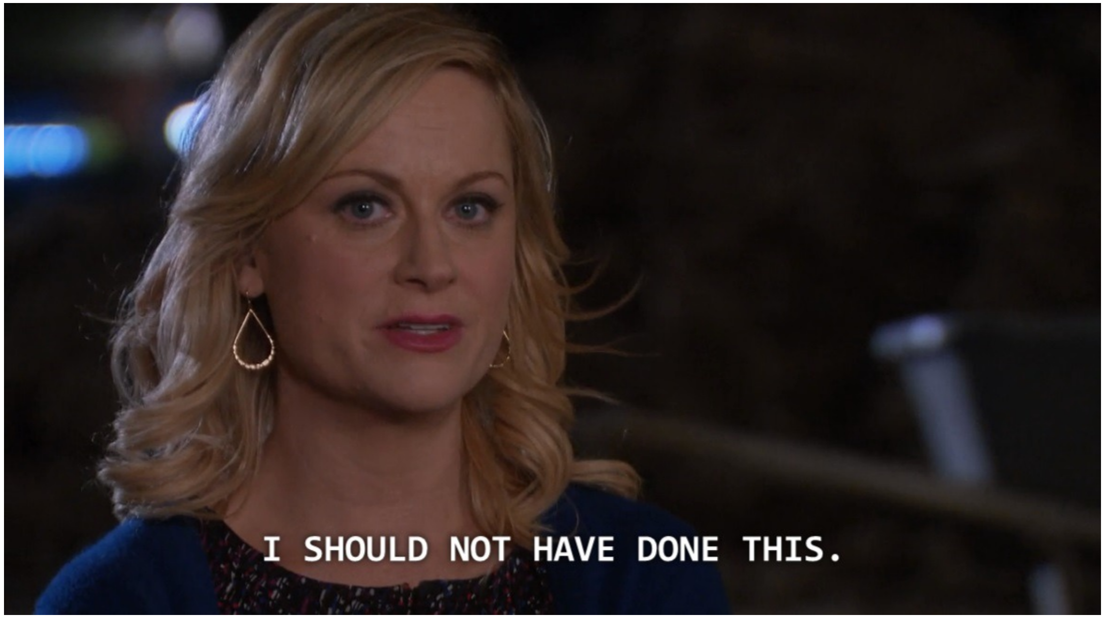


MAR 580: Advanced Population Modeling

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

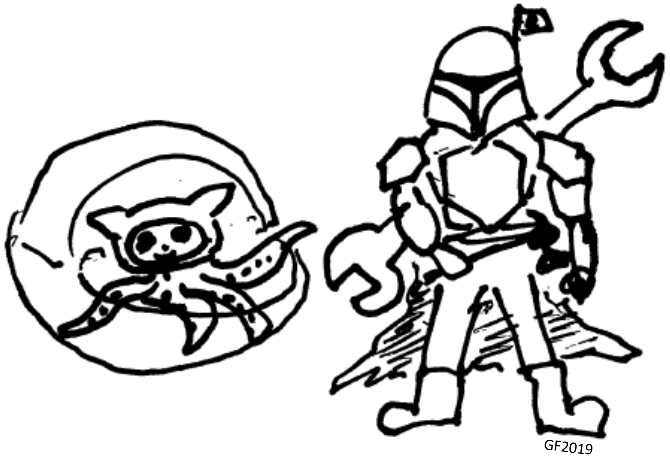
- students using R Markdown for assignments
- live coding during mixed lab/lectures
- AirMedia to share student screens to class: debugging aid
- course materials shared via Google Drive

#quantfish woRkshops

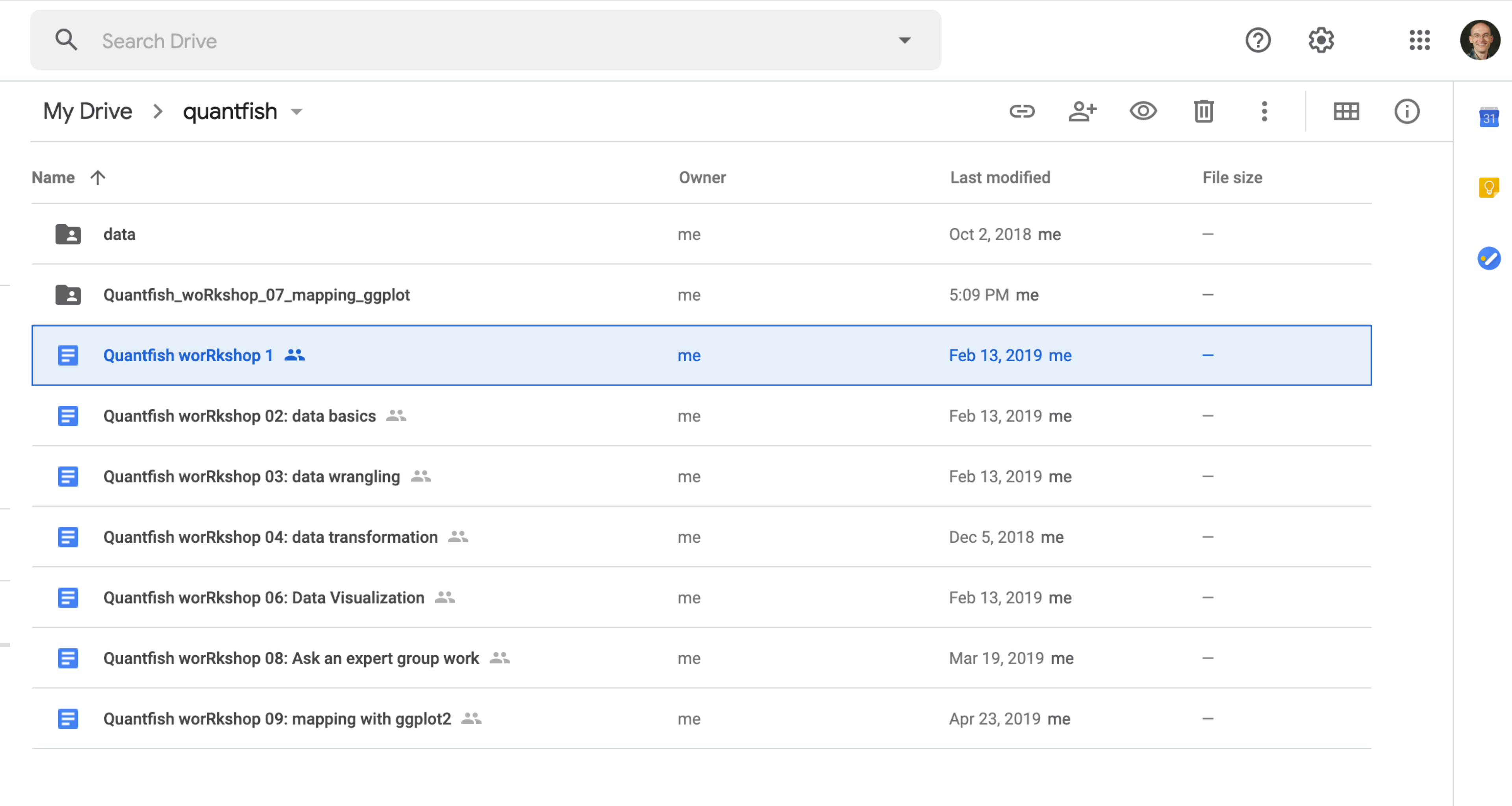
#quantfish woRkshops

tutorials for beginner and intermediate R users

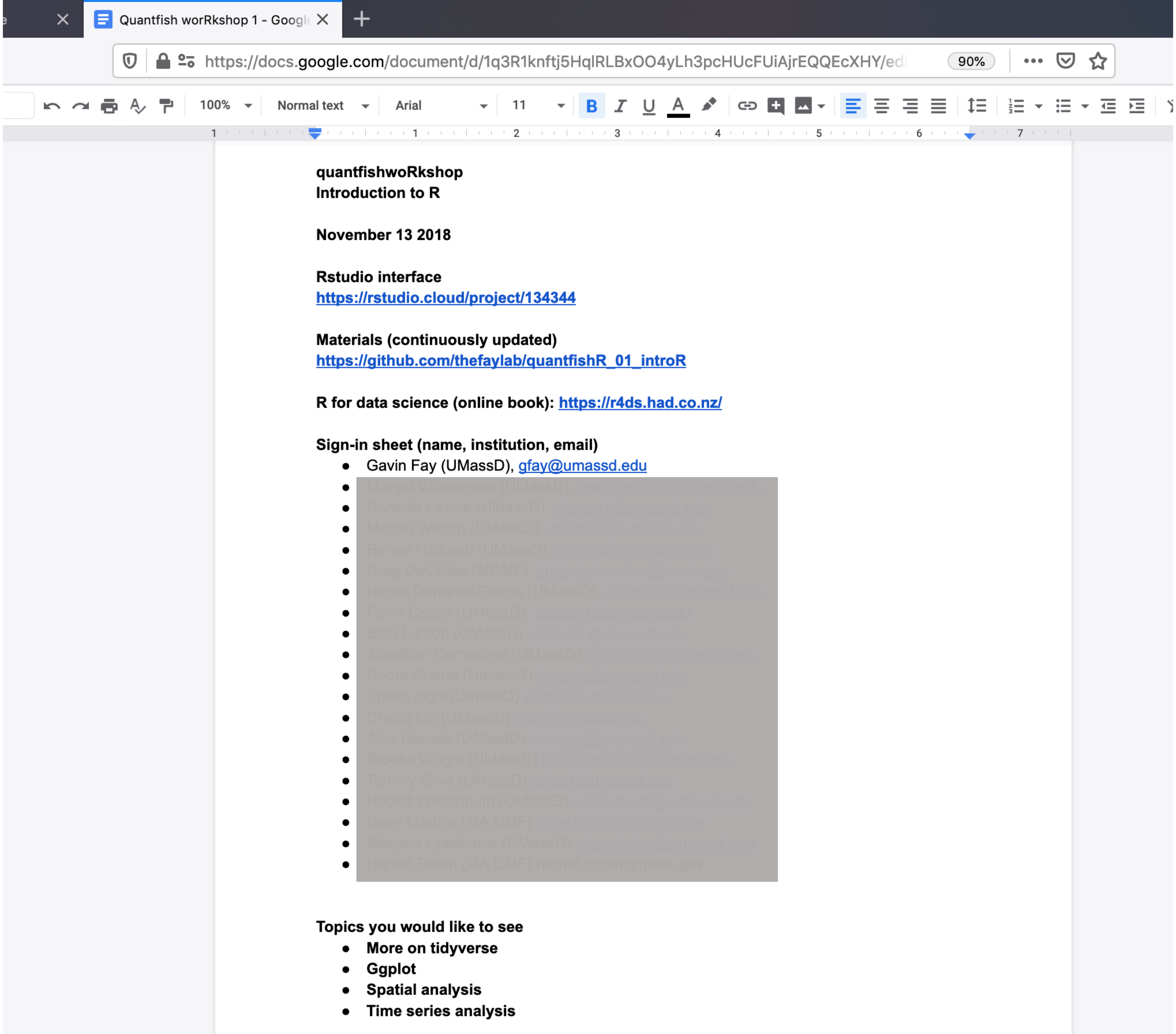
- Students & postdocs lead 1.5 hr sessions
- Live coding
- Learning R by doing useful things straight away
- Less is more
- Sharing of materials via GoogleDocs
- Materials version-controlled using git and github
- Feedback asked for (& acted on) often



A GoogleDoc for each workshop….



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	–


Quantfish worRkshop 1 - Googl
https://docs.google.com/document/d/1q3R1knftj5HqlRLBxOO4yLh3pcHUcFUiAjrEQQEcXHY/ed
90%
100%
Normal text
Arial
11

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/

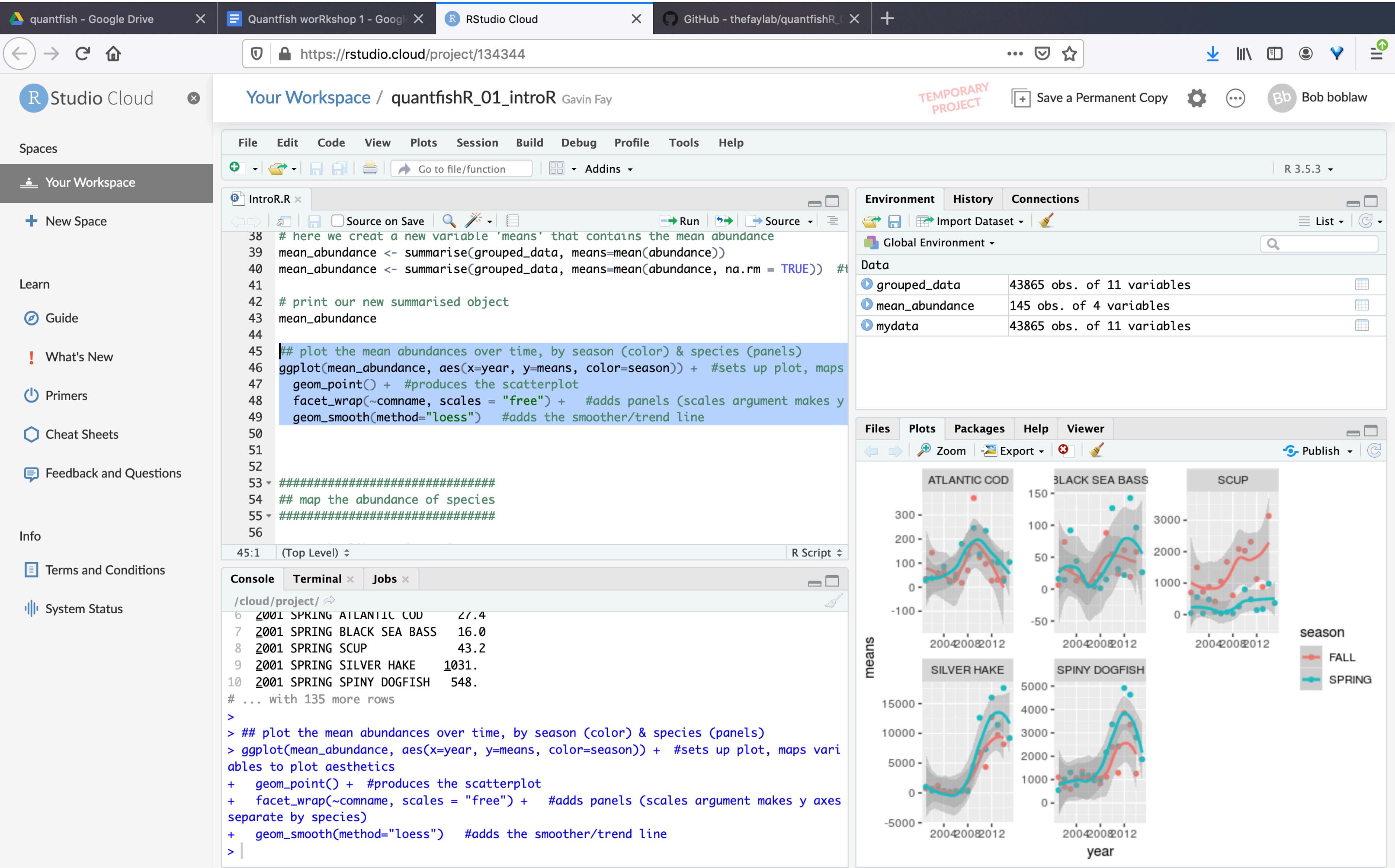
Sign-in sheet (name, institution, email)

- Gavin Fay (UMassD), gfay@umassd.edu
- [redacted]

Topics you would like to see

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

Standard environment: RStudio & RStudio Cloud



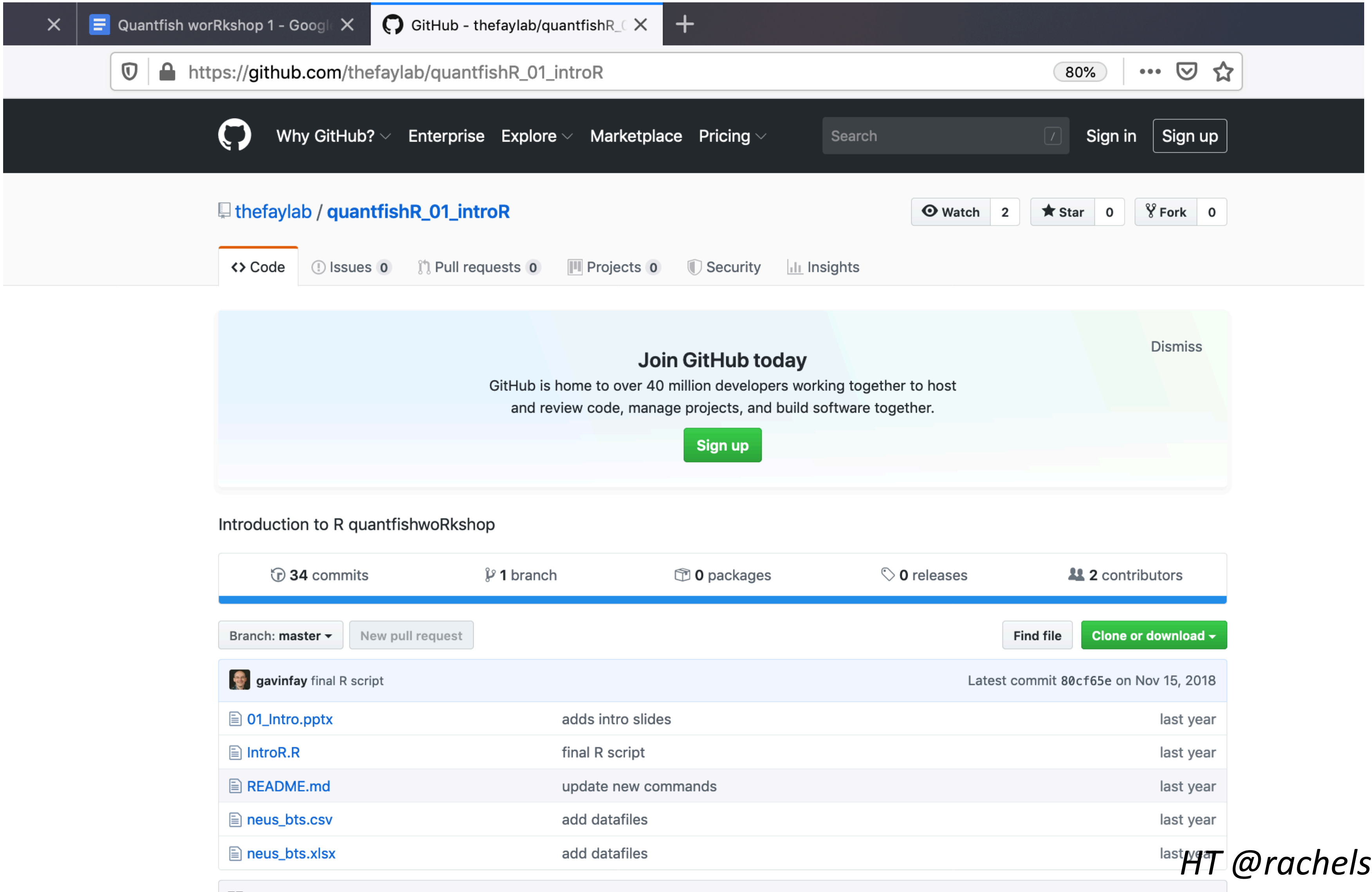
```
# here we creat a new variable 'means' that contains the mean abundance
mean_abundance <- summarise(grouped_data, means=mean(abundance))
mean_abundance <- summarise(grouped_data, means=mean(abundance, na.rm = TRUE))  #t

# print our new summarised object
mean_abundance

## 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
  geom_point() +  #produces the scatterplot
  facet_wrap(~comname, scales = "free") +   #adds panels (scales argument makes y
  geom_smooth(method="loess")   #adds the smoother/trend line



################################
## map the abundance of species
################################
```

Data

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

```
 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 vari
ables 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
> 
```

Code online in github repository
Auto-updated during workshop



Join GitHub today

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

Introduction to R quantfishwoRkshop

34 commits · 1 branch · 0 packages · 0 releases · 2 contributors

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

HT @rachelss

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!

gfay@umassd.edu

thefaylab.com

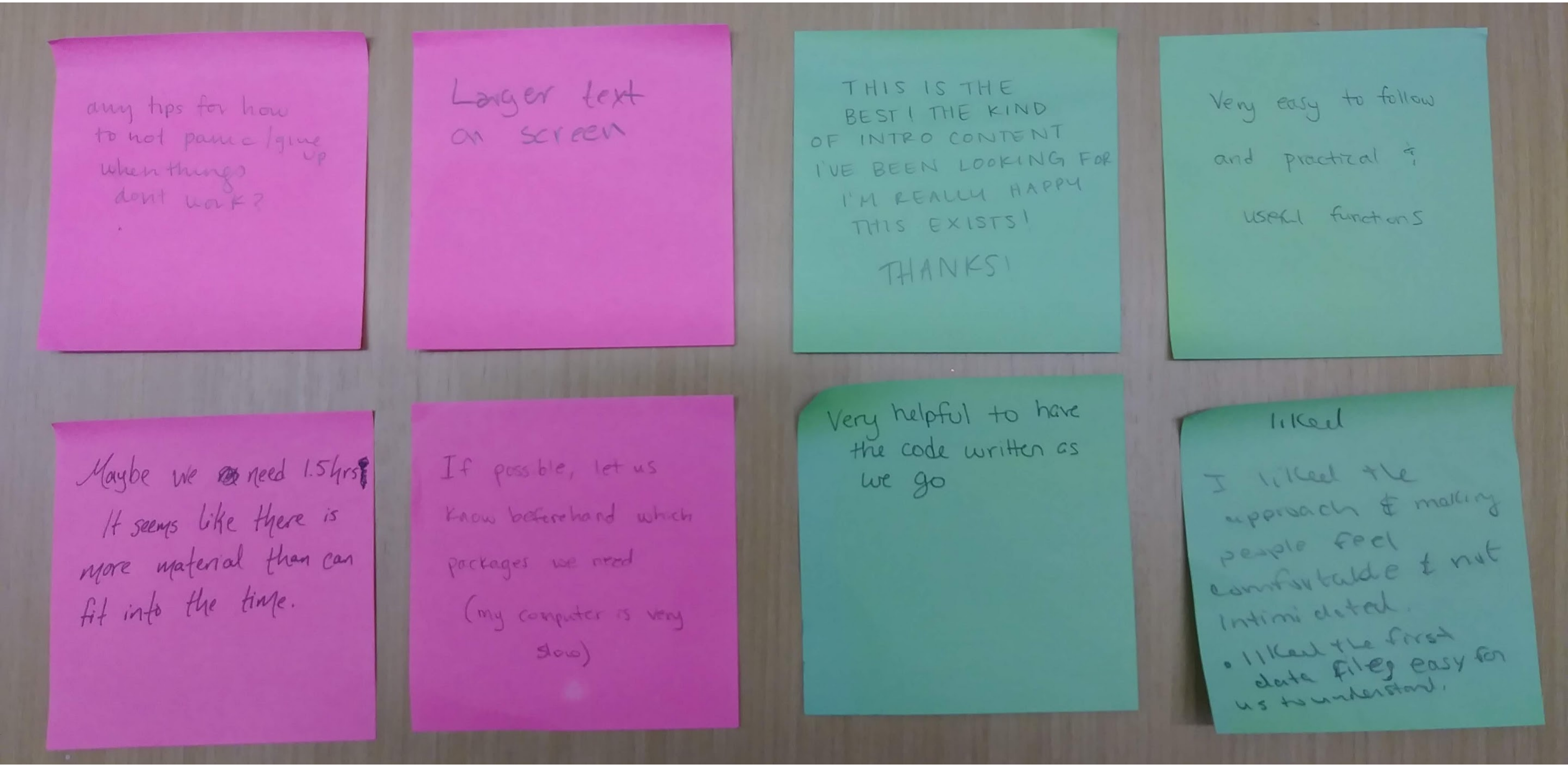
@gavin_fay

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 dont 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 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

liked

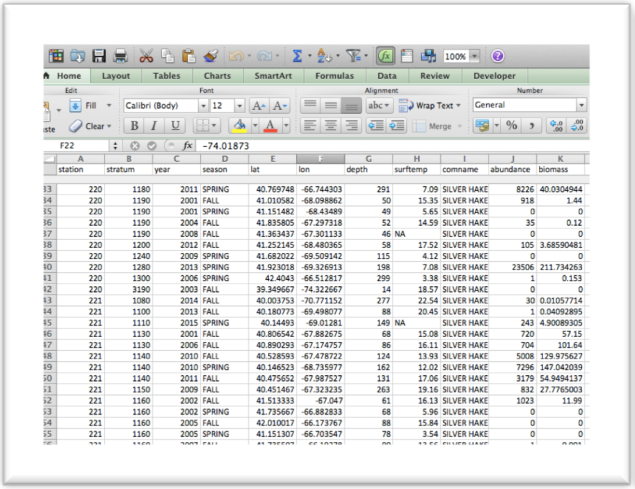
I liked the approach & making people feel comfortable & not intimidated.

- I liked the first data files easy for us to understand.

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

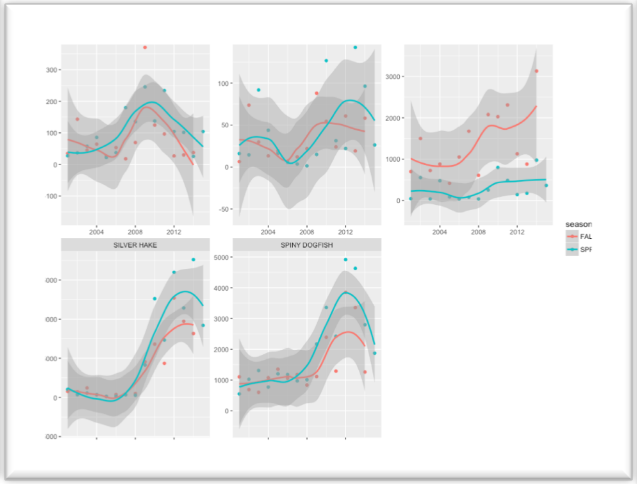
Intro to R: Analyze US fish data

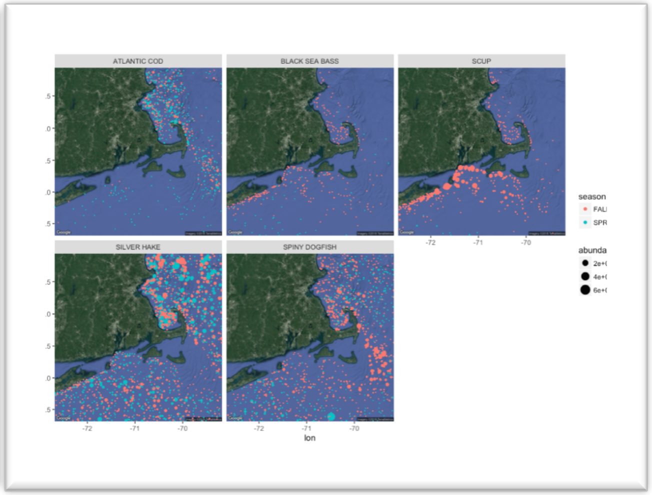
- Take data from spreadsheet to visualization
- Data wrangling
- Summarizing data by species over time
- Mapping of fish distributions



F22 *fx* -74.01873

	A	B	C	D	E	F	G	H	I	J	K
	station	stratum	year	season	lat	lon	depth	surftemp	comname	abundance	biomass
33	220	1180	2011	SPRING	40.769748	-66.744303	291	7.09	SILVER HAKE	8226	40.0304944
34	220	1190	2001	FALL	41.010582	-68.098862	50	15.35	SILVER HAKE	918	1.44
35	220	1190	2001	SPRING	41.151482	-68.43489	49	5.65	SILVER HAKE	0	0
36	220	1190	2004	FALL	41.835805	-67.297318	52	14.59	SILVER HAKE	35	0.12
37	220	1190	2008	FALL	41.363437	-67.301133	46	NA	SILVER HAKE	0	0
38	220	1200	2012	FALL	41.252145	-68.480365	58	17.52	SILVER HAKE	105	3.68590481
39	220	1240	2009	SPRING	41.682022	-69.509142	115	4.12	SILVER HAKE	0	0
40	220	1280	2013	SPRING	41.923018	-69.326913	198	7.08	SILVER HAKE	23506	211.734263
41	220	1300	2006	SPRING	42.4043	-66.512817	299	3.38	SILVER HAKE	1	0.153
42	220	3190	2003	FALL	39.349667	-74.322667	14	18.57	SILVER HAKE	0	0
43	221	1080	2014	FALL	40.003753	-70.771152	277	22.54	SILVER HAKE	30	0.01057714
44	221	1100	2013	FALL	40.180773	-69.498077	88	20.45	SILVER HAKE	1	0.04092895
45	221	1110	2015	SPRING	40.14493	-69.01281	149	NA	SILVER HAKE	243	4.90089305
46	221	1130	2001	FALL	40.806542	-67.882675	68	15.08	SILVER HAKE	720	57.15
47	221	1130	2006	FALL	40.890293	-67.174757	86	16.11	SILVER HAKE	704	101.64
48	221	1140	2010	FALL	40.528593	-67.478722	124	13.93	SILVER HAKE	5008	129.975627
49	221	1140	2010	SPRING	40.146523	-68.735977	162	12.02	SILVER HAKE	7296	147.042039
50	221	1140	2011	FALL	40.475652	-67.987527	131	17.06	SILVER HAKE	3179	54.9494137
51	221	1150	2009	FALL	40.451467	-67.323235	263	19.16	SILVER HAKE	832	27.7765003
52	221	1160	2002	FALL	41.513333	-67.047	61	16.13	SILVER HAKE	1023	11.99
53	221	1160	2002	SPRING	41.735667	-66.882833	68	5.96	SILVER HAKE	0	0
54	221	1160	2005	FALL	42.010017	-66.173767	88	15.84	SILVER HAKE	0	0
55	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.)
lubri
dplyr
tidy
%>%
ggpl
readr
for
CODE
# HEADERS
italics
LaTe
TEXT
FIGUR
PANDOC
KNITR
FAIRY DUST
YAML
rmarkdown
RECIPE
1. ADD TEXT
2. ADD CODE
3. KNIT
4. (magic)
5. CELEBRATE PERCEIVED WIZARDRY
.Rmd
.md
knitr
pandoc
Horst '19

Artwork by @allison_horst

https://vimeo.com/178485416