R TRAINING AT WORK

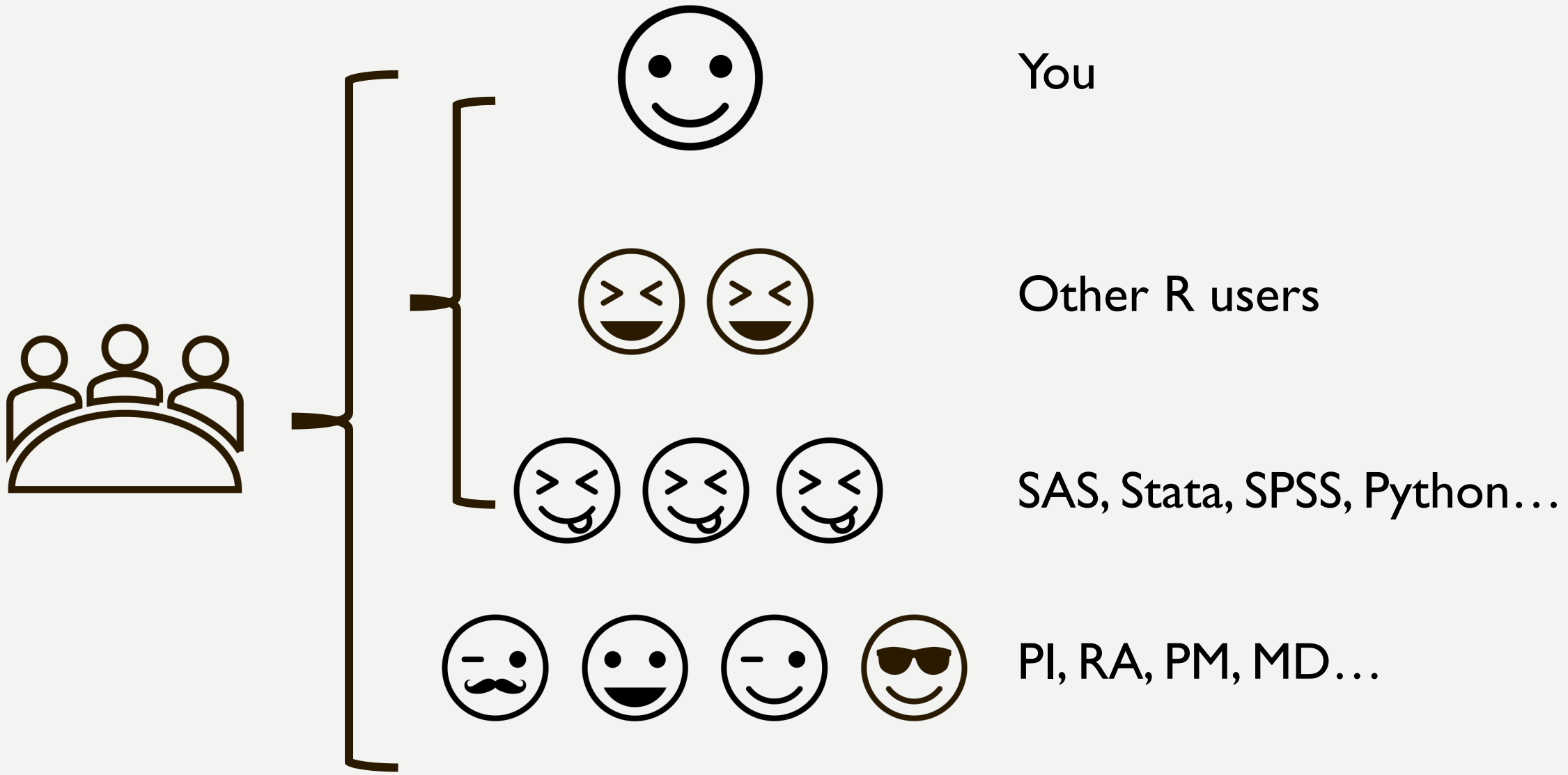
HOW TO BRING R TO YOUR RESEARCH TEAM

Hao Zhu

@haozhu233

Marcus Institute for Aging Research
Hebrew SeniorLife

HARVARD MEDICAL SCHOOL AFFILIATE


You
Other R users
SAS, Stata, SPSS, Python…
PI, RA, PM, MD…

"DIVERSITY" HAS GOOD SIDES BUT IT COMES WITH SOME PRICE

- We don't usually review code for each other
- Different workflows require different infrastructure and setup
- We can't utilize everyone's potential
- We can't learn advanced stuff from each other if we can't even understand the basics.

GOAL:
WE WANT TO TEACH PEOPLE R SO:

- A FEW MORE PEOPLE CAN START TO USE R AS THEIR PRIMARY DATA LANGUAGE
- MORE PEOPLE COULD AT LEAST READ R CODE

WHY

A LITTLE MOTIVATION

- Data input/output from various sources, including SAS, Stata and SPSS (via haven)
- State of art data manipulation tools, such as dplyr and tidyr
- State of art visualization tools, such as ggplot

A LITTLE MOTIVATION

- Straightforward reproducible reports using rmarkdown
- Dead easy web app framework, shiny
- Intuitive IDE, RStudio

...

- **Open and Welcoming Community**

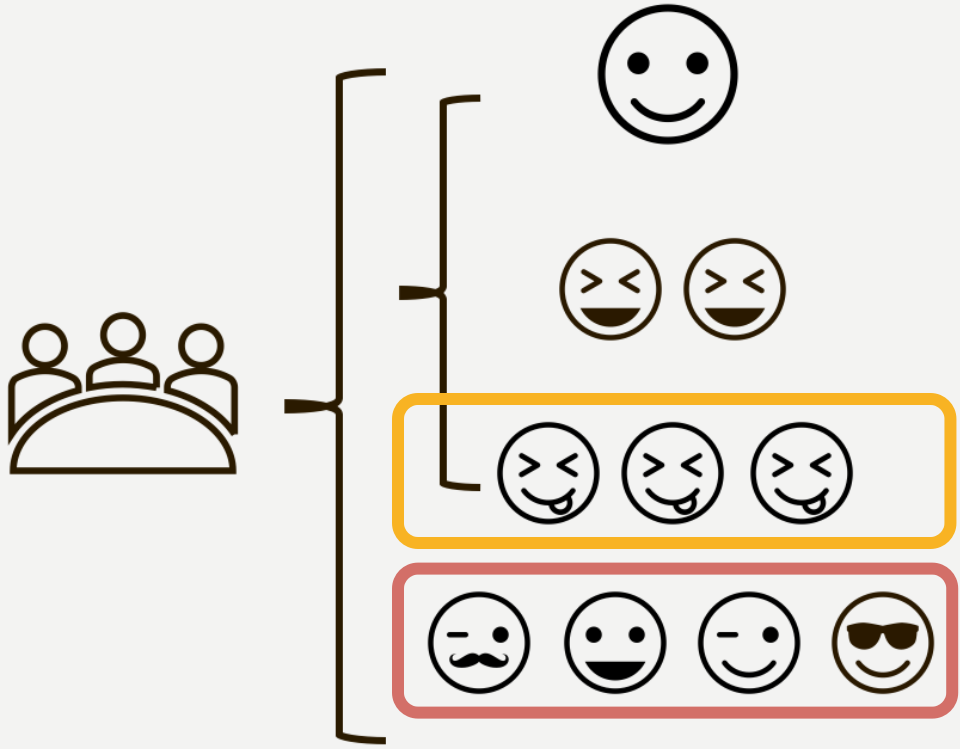
HOW

COULD WE ACHIEVE THIS GOAL?

1. UNDERSTAND YOUR TARGET AUDIENCE

- For Data Teams:
  - Strong background in statistics
  - Not familiar with R itself
- For Other Research Teams:
  - No background in statistics
  - Don't have need to do heavy statistical training

- Conclusion: Instead of an "Intro to Stats" class, our workshop should focus on the practical knowledge of how to use R.



2. UNDERSTAND THE CHALLENGES

- One of the biggest challenges for workspace training is that people are really **busy**.
  - No protected time for study
  - Sometimes it seems to be not very effect to recommend working learners a textbook such as *R for data science*.
- If people don't use R very often, they will forgot it eventually and need to be trained again.

- Conclusion: Our training materials need to be concise and re-accessible.

RESULTS

Data Science with R Workshop Series

Biostatistics and Data Science Group at Marcus Institute

This site includes materials for the Data Science with R Workshop Series hosted by Marcus Institute. All the materials are developed by the Biostats and Data Science group at Marcus Institute. We aim to provide trainings to our own group numbers and individuals who are interested in doing data science works using R.

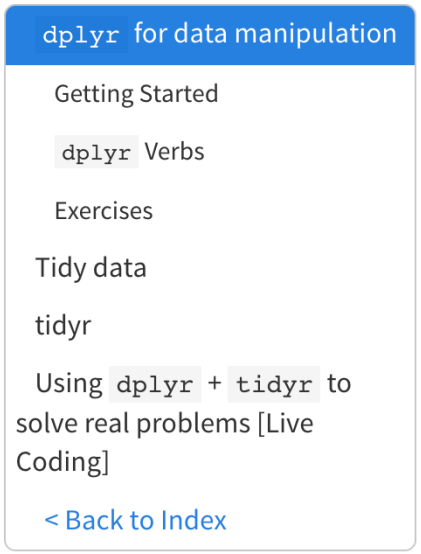
- Chapter 1 - R Programming Fundamentals
- Chapter 2 - Read Data into R
- Chapter 3 - Data Transformation using Tidyverse
- Chapter 4 - Data Visualization using ggplot2 & plotly
- Chapter 5 - Statistical Models in R
- Chapter 6 - R Markdown Reports and Tables in Reports
- Chapter 7 - Creating Interactive Web Apps using R Shiny

Chapter 1 - 3 are fundamentals, which you are not recommended to skip. Chapter 4 - 8 don't have to be read in order, although consuming them one by one is still recommended.

All pages are generated using R Markdown. If you want to see the source R Markdown files, just replace the `.html` with `.Rmd` in the URL.

Material Accessible from: http://bit.ly/marcusdata
Github Source: https://github.com/hebrewseniorlife/data_science_workshops (Or http://bit.ly/rtraining_github)

RESULTS



Chapter 3 - Data Transformation using Tidyverse

Mutate, Summarize, Gather & Spread - Time for some R magic spells

Hao Zhu

2019-04-05 (updated 2019-06-24)

< Back to Index

`Tidyverse` is a collection of 8 awesome R packages, including:

- `dplyr` : Data manipulation (Chapter 3)
- `tidyr` : Data Long/Wide Transformation (Chapter 3)
- `purrr` : List-based manipulation (Chapter 3)
- `tibble` : Improved version of data.frame (Chapter 1)
- `readr` : Improved data reading methods (Chapter 2)
- `stringr` : String Operations using regular expression
- `forcats` : Categorical Analyses
- `ggplot2` : Data Visualizations with a set of grammar (Chapter 5)

In this chapter, we will talk about data transformation and will focus on `dplyr` and `tidyr`. We will also give a basic introduction to list-based analyses using `purrr`. As we continue forward in this training, we will discuss the

Material Accessible from: http://bit.ly/marcusdata
Github Source: https://github.com/hebrewseniorlife/data_science_workshops (Or http://bit.ly/rtraining_github)

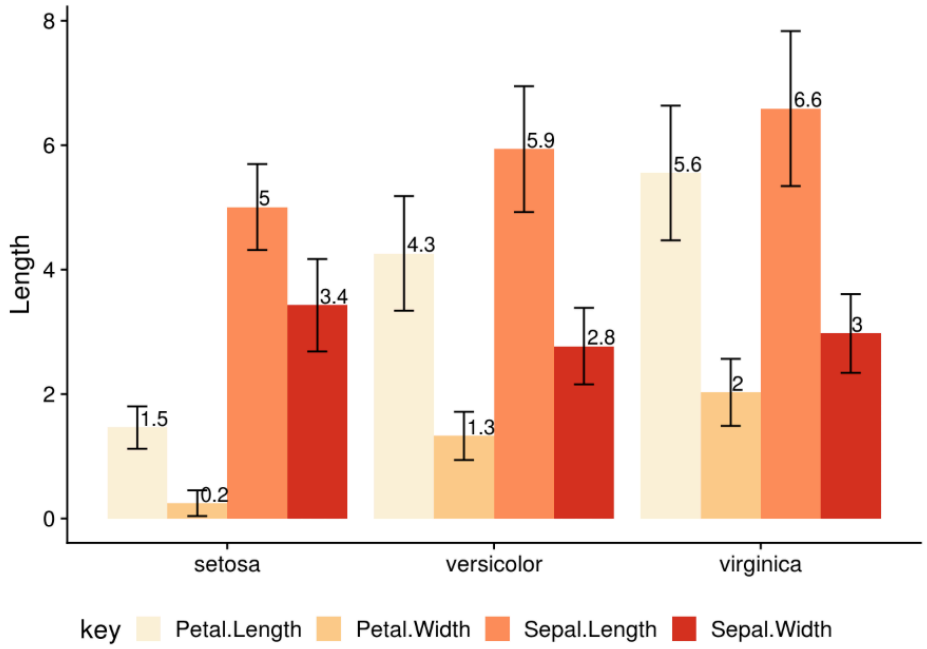
RESULTS

- ggplot2
  - Grammar of Graphics Fundementals
  - Scale
  - Coorinate
  - Facets
  - Theme
  - Position dodge/jitter
  - Multi-plot panel
  - Save/Export plot
  - **Common Figures**
    - Bar plot with error bar
    - Forest Plot

Common Figures

Bar plot with error bar

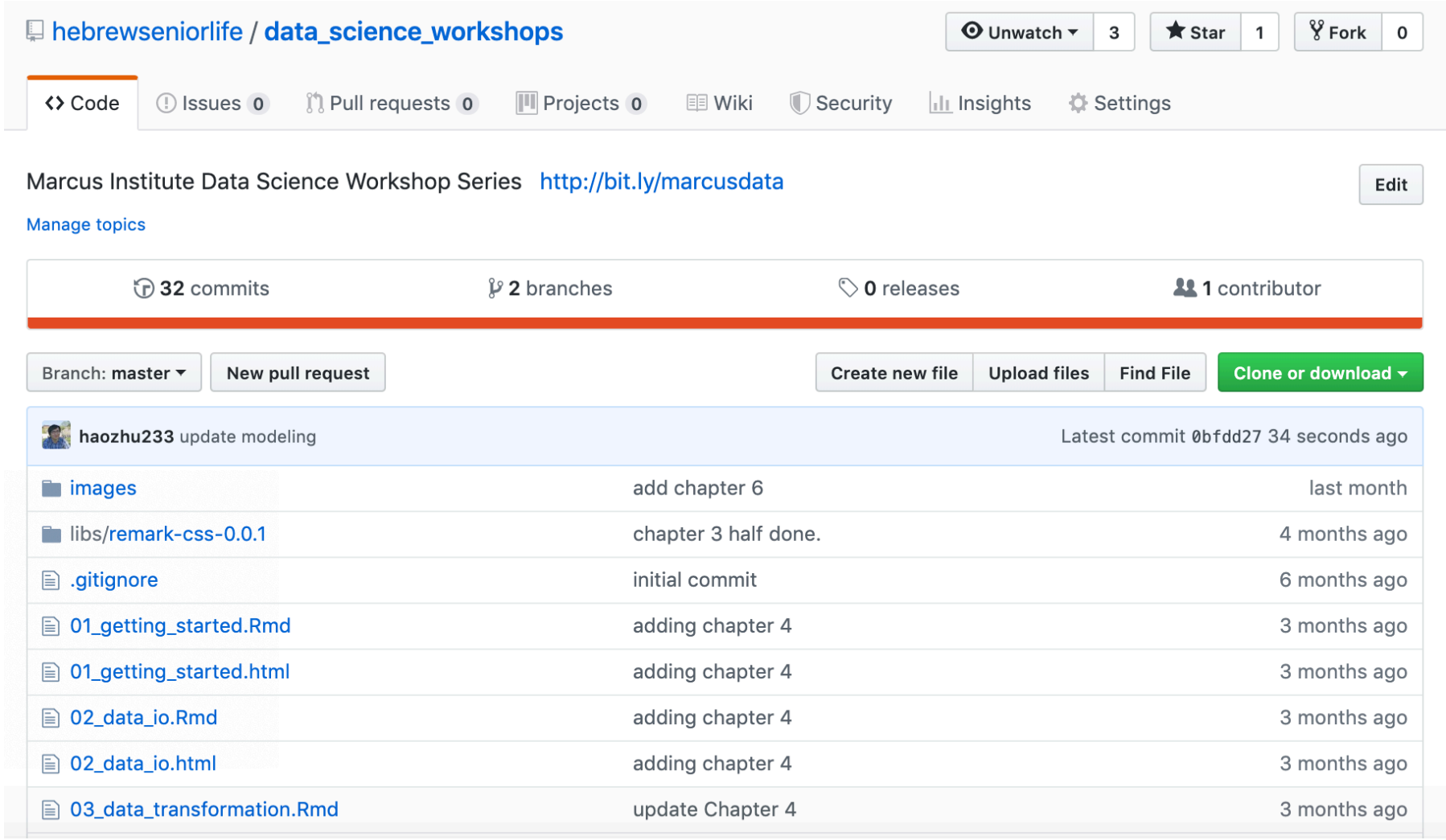
```
ggplot(iris_sum, aes(x = Species, y = mean)) +
  geom_bar(aes(fill = key), stat = "identity", position = "dodge") +
  geom_errorbar(aes(ymin = low, ymax = high, group = key),
                position = position_dodge(0.9), width = 0.3) +
  geom_text(aes(label = round(mean, 1), group = key),
            position = position_dodge(0.9), vjust = -0.1, hjust = -0.1) +
  scale_fill_brewer(palette="OrRd") +
  labs(x = NULL, y = "Length") +
  theme(legend.position = "bottom")
```



Material Accessible from: http://bit.ly/marcusdata

Github Source: https://github.com/hebrewseniorlife/data_science_workshops (Or http://bit.ly/rtraining_github)

RESULTS



Material Accessible from: http://bit.ly/marcusdata
Github Source: https://github.com/hebrewseniorlife/data_science_workshops (Or http://bit.ly/rtraining_github)

WORKSHOP SERIES (8HR IN TOTAL)

Biostats Core	8 1hr sessions during our weekly group meeting
Field Team	4 2hr sessions happening on different days within 3 weeks

FEEDBACK

"I had prior experience with R, but I tend to program in a very SAS-influenced way. The modules about neater ways of data manipulation, tidyverse, and also R Markdown were eye-opening. I'll try to implement in practice what I learnt during the workshop."

"The training workshop is very thorough. It covers many aspects of R. "

"The training materials not only gives detailed explanation about different functions, they also provide examples about how to use them. I found the workshop extremely helpful!"

THOUGHTS

Teaching at work is different from teaching at school

Helping your colleagues build confidence on using the tool is the key to success

You should think about using RStudio Server or RStudio Cloud to do training

Live coding is great!

The teaching workshop from the Software Carpentry is very helpful!

ACKNOWLEDGEMENT

- Alyssa Dufour, for all the editorial work on course materials!
- Tom Travison, for all the leadership and support
- Biostats and Data Science Core at Marcus Institute, for all the course feedback
- Funding Sources:
  - Boston Pepper Center (P30AG031679, S Bhasin)
  - Core Development Project Grant (T Travison)

THANK YOU!

ANY QUESTIONS?