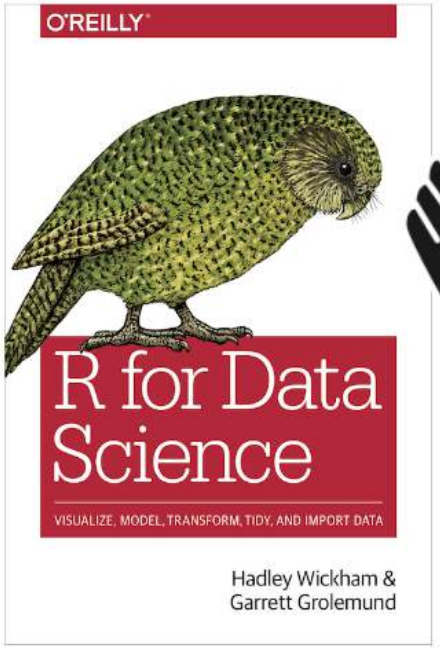
Teaching **Data Science**, **Statistical Thinking**, and **Collaboration** Using Team-Based Learning



Eric Vance
Dept. Applied Math

eCOTS Breakout Session
May 22, 2018

University of Colorado **Boulder**

Statistics + Data Science undergraduate majors should learn 9+ topics throughout curriculum

- Statistical Thinking & Data Acumen
- Statistical Theory & Methods
- Communication & Collaboration
- Domain Knowledge/ Areas of Application
- Ethics

- Computational Thinking
- Reproducible Workflows
- Mathematical foundations
- Programming/coding/hacking skills

Statistics + Data Science undergraduate majors should learn 9+ topics throughout curriculum

- Statistical Thinking & Data Acumen
- Statistical Theory & Methods
- Communication & Collaboration
- Domain Knowledge/ Areas of Application
- Ethics
- Computational Thinking
- Reproducible Workflows
- Mathematical foundations
- Programming/coding/hacking skills

Collaborate with experts who have complementary skills

Why learn Collaboration?

Why learn Collaboration?

Statisticians and Data Scientists rarely "own" the data they analyze

Important problems to be solved or decisions to be made typically originate with Domain Experts

For real-world impact, Statisticians and Data Scientists must **collaborate** with various Domain Experts to:

- understand and refine questions to be answered
- access and analyze the appropriate data
- communicate conclusions, recommendations, and findings

Why learn Collaboration?

Statisticians and Data Scientists rarely "own" the data they analyze

Important problems to be solved or decisions to be made typically originate with Domain Experts

For real-world impact, Statisticians and Data Scientists must **collaborate** with various Domain Experts to:

- understand and refine questions to be answered
- access and analyze the appropriate data
- communicate conclusions, recommendations, and findings

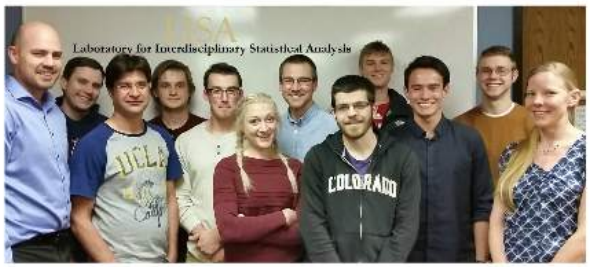
How and when to teach Collaboration?

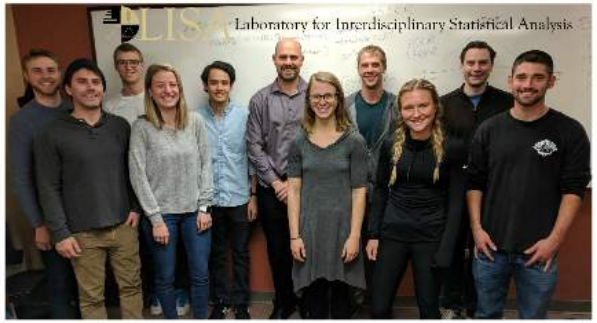
How and when to teach Collaboration?

I teach a capstone course “Statistical Collaboration” for seniors (and graduate students)

Students learn fundamentals of Collaboration, collaborate with domain experts on LISA research projects & learn new statistical methods via projects

LISA
Laboratory for Interdisciplinary Statistical Analysis





Poll #1

How important is teaching undergrads collaboration?

A. Collaboration should be taught throughout the curriculum

B. A course focused on Collaboration is a necessity

C. We should add a module on Collaboration to a current course

D. It would be nice to mention Collaboration in a course if there is time

E. Not enough to displace anything currently taught

Teaching “Intro to Data Science” to 1st-years

17 students (mostly Applied Math majors)

3-credit 2000-level course to become 4 cr. next year

Learning Objectives:

To develop the technical and professional skills necessary to analyze data as a member of a team

- Understanding fundamental statistical concepts
- Visualizing and exploring data
- Importing and tidying datasets
- Programming effectively in R
- Building basic statistical models
- Collaborating with teammates to discover & communicate interesting findings & recommendations based on data
- Mastering reproducible workflows

Teaching "Intro to Data Science" to 1st-years

17 students (mostly Applied Math majors)

3-credit 2000-level course to become 4 cr. next year

Learning Objectives:

To develop the technical and professional skills necessary to analyze data as a member of a team

- Understanding fundamental statistical concepts
- Visualizing and exploring data
- Importing and tidying datasets
- Programming effectively in R
- Building basic statistical models
- Collaborating with teammates to discover & communicate interesting findings & recommendations based on data
- Mastering reproducible workflows

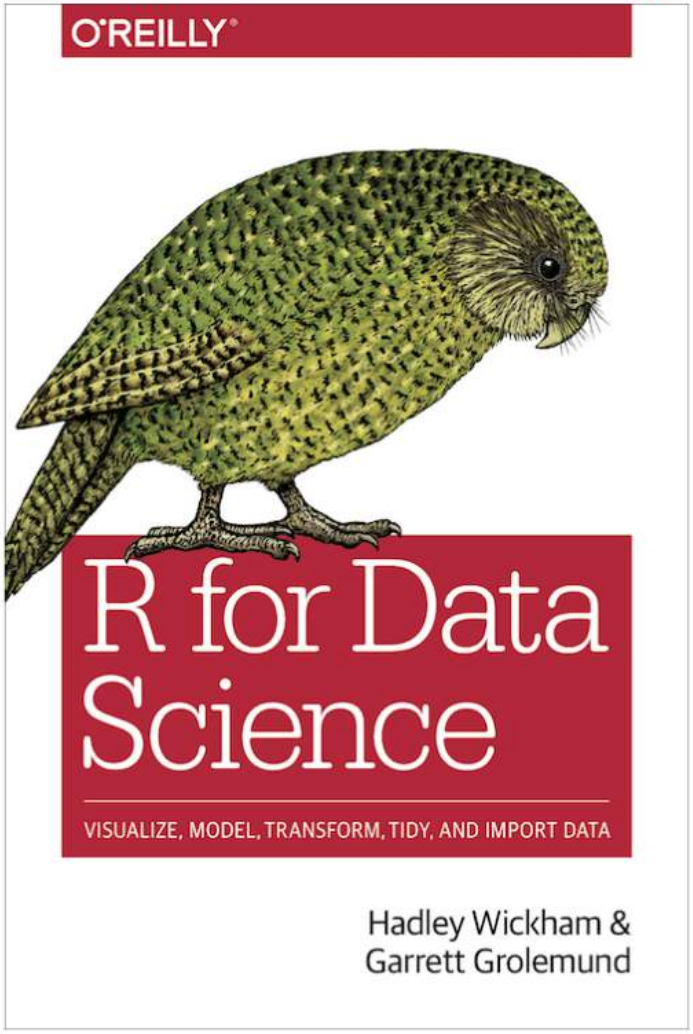
"Learn R to do interesting and useful things with data"

Course split into 7 modules over 15 weeks

Each module was 3-5 chapters in ***R for Data Science***

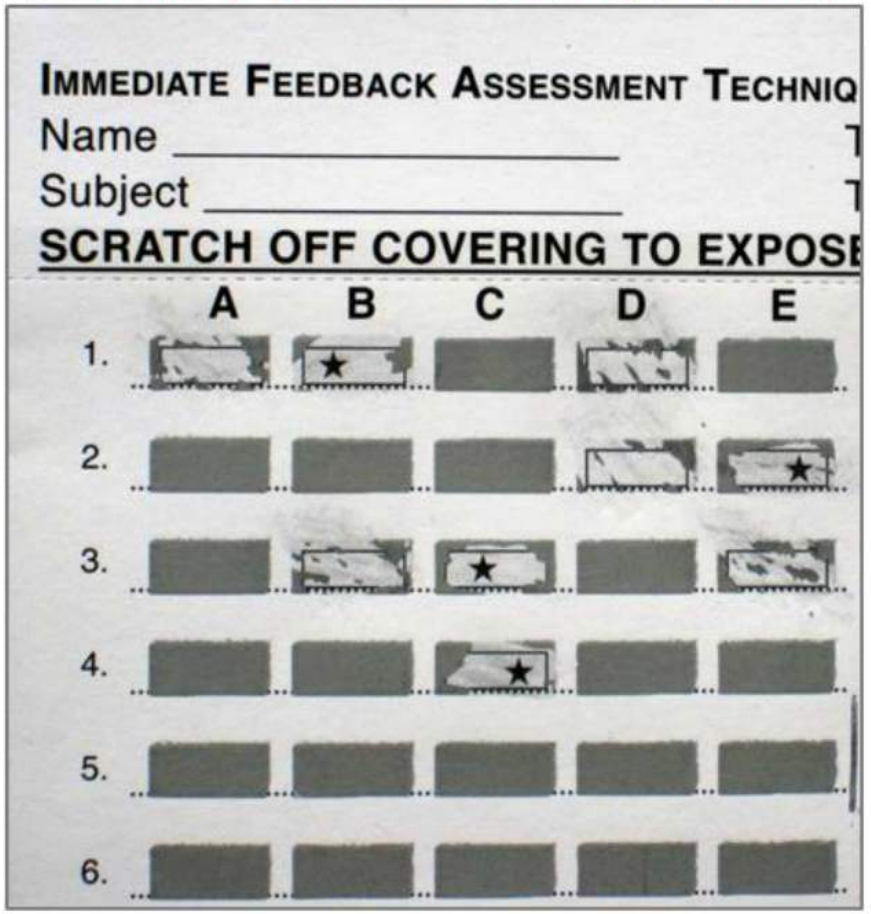
Students read the book and do the exercises then take an *individual* and *team* test on main concepts

0. **Workflows**: Ch 1, 26, 27, 30
1. **Data Visualization**: Ch 2-4
2. **EDA**: Ch 5-8
3. **Importing and Tidying Data** Ch 9-12
4. **Relational Data, Strings, Factors, Dates and Times**: Ch 13-16
5. **Programming in R** (pipes, functions, vectors, iteration): Ch 17-21
6. **Statistical Modeling**: Ch 22-25



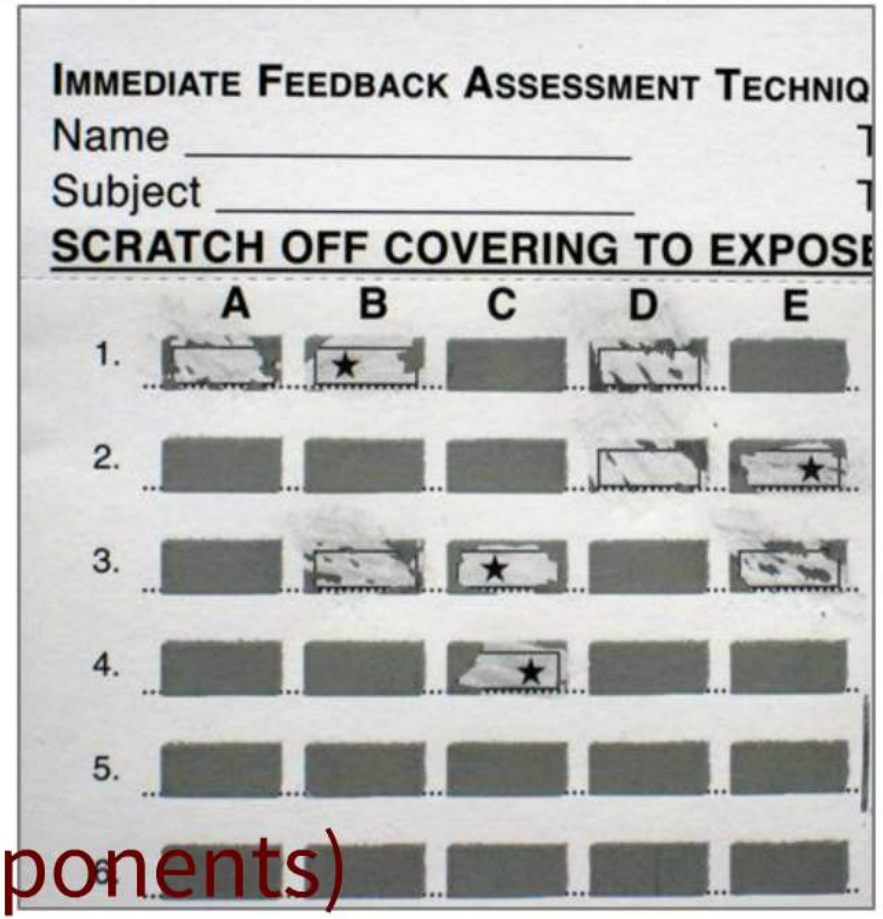
Each module repeats the same structure

- Students read the book and do the exercises
- *Individual* readiness assurance m.c. test on Class 1
- Exact same *team* test with immediate feedback



Each module repeats the same structure

- Students read the book and do the exercises
- *Individual* readiness assurance m.c. test on Class 1
- Exact same *team* test with immediate feedback
- Appeals
- Clarifying lecture
- Application exercises during Classes 1-6
- Weekly lab assignments (individual and team components)



Students learn Statistical Thinking through in-class application exercises and weekly lab assignments

- Correlation v. causation & confounding variables
- Sampling, populations, inference
- Simpson's Paradox, Bayes Rule, conditional probabilities
- Comparisons of groups or values
- Simulated p-values
- Testing hypotheses via permutations

Poll #2

What do you think would be the BIGGEST problem with this structure?

A. Too much material to cover in just one semester

B. Students won't read the book before each module

C. Difficult to create in-class exercises and labs that teach statistical thinking skills while practicing programming skills

D. Some students will be slackers and ruin team cohesion (unfair to high-performing students)

E. Some teams much higher performing than others

Team-Based Learning (TBL) combines the best aspects of flipped-classroom, problem-based learning with small-group learning

Team-Based Learning™ Collaborative

www.teambasedlearning.org

Team-Based Learning (TBL) combines the best aspects of flipped-classroom, problem-based learning with small-group learning

TBL keys to effective flipped-classroom, problem-based learning:

- Students come to class prepared to apply course content to solve problems

Students read the book because there will be a test

They also learn from their peers during the team test

- Application exercises must be well designed

Significant, **S**ame problem, **S**pecific choice, **S**imultaneous reporting

Example warm-up Application Exercises

Team Application Exercises

Your data science team has now been hired by the NYC Port Authority to analyze flights departing NYC in 2013. You searched for and found and imported datasets showing all the flights departing NYC in 2013 ("flights"); a list of all North American airports, their codes and locations ("airports"); a dataset of weather variables at the three NYC airports by hour and day for 2013 ("weather"); a dataset of commercial airline planes active in 2013 ("planes"); and a small dataset of commercial airline names and their two-character codes ("airlines").

1. Quickly decide individually then as a team which two-table function ("verb") AND which variable "key" (or multiple keys) would be most useful for the following:

- Identifying which airport destinations (from NYC) had the lowest average daily temperatures
- Determining the ages of the 20 airplanes having the highest median flight arrival delays
- Plotting the average hourly precipitation (over the whole year) for all the flights delayed by more than the 80 %tile departure delay
- Ordering the airlines according to the number of planes they flew in 2013 (from highest to lowest)
- Plotting the locations of airports that did not receive a flight from NYC in 2013
- Determining the 90 %tile wind speed (in NYC) at each hour for the scheduled departure time for flights with at least a 30-minute arrival delay
- Plotting a bar chart of the proportion of flights late by more than 15 minutes for the 5 airports with the highest percentage of late-arriving flights (with at least 100 flights landed), in order of lowest to highest

A. left_join()
B. inner_join()
C. full_join()
D. semi_join()
E. anti_join()

A. tailnum
B. month, day, hour
C. dest
D. carrier
E. origin

AE. None of these (what would it be instead?)

Team-Based Learning (TBL) combines the best aspects of flipped-classroom, problem-based learning with small-group learning

TBL keys for effective small-group learning:

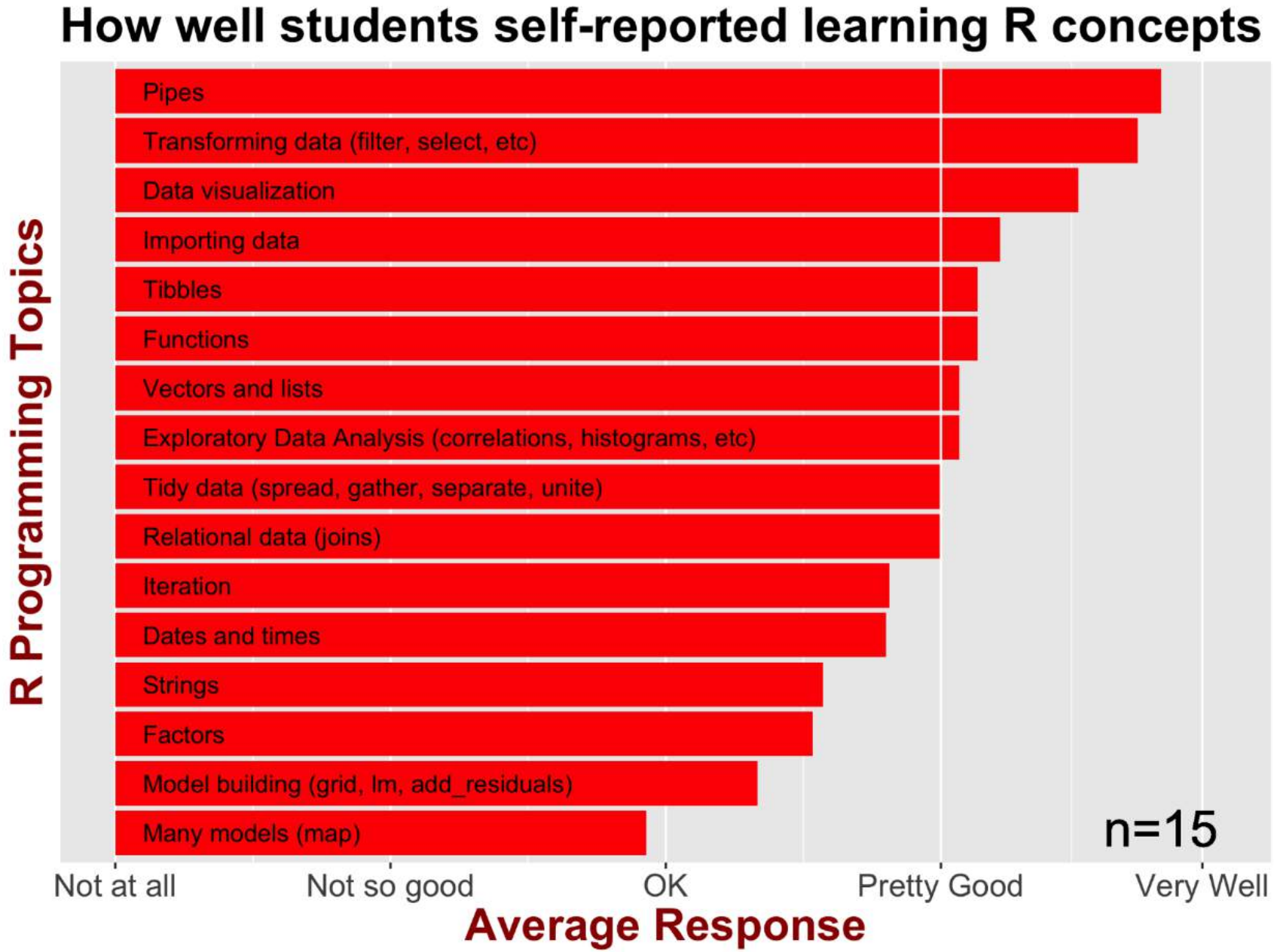
- Balanced, permanent teams of 5-7 students transparently chosen by the professor

Allows time to build cohesion and team identity

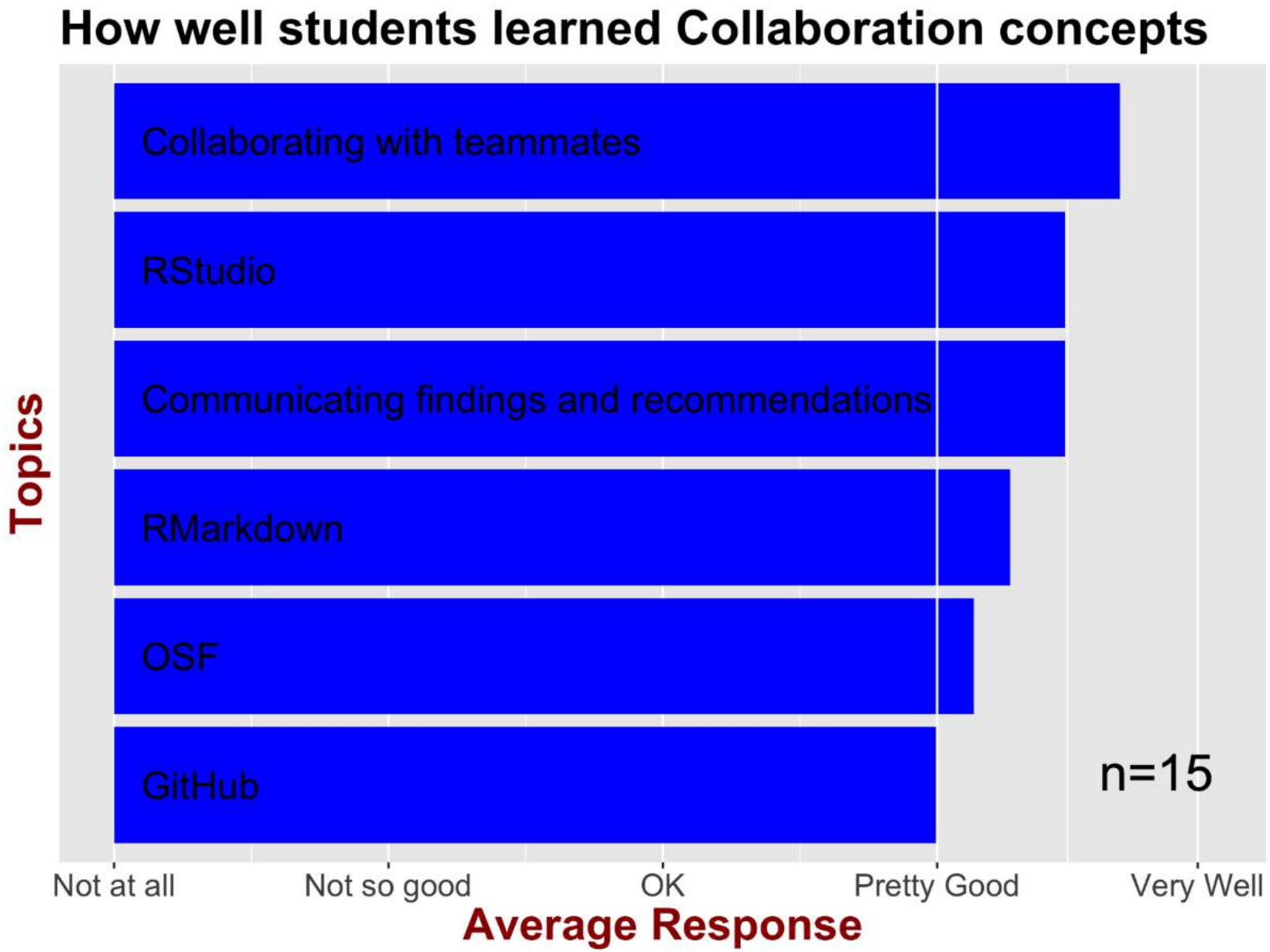
- Students are accountable to their peers

Peer Evaluation and Team Maintenance: qualitative and quantitative (20% of their grade) peer feedback 3 times per semester

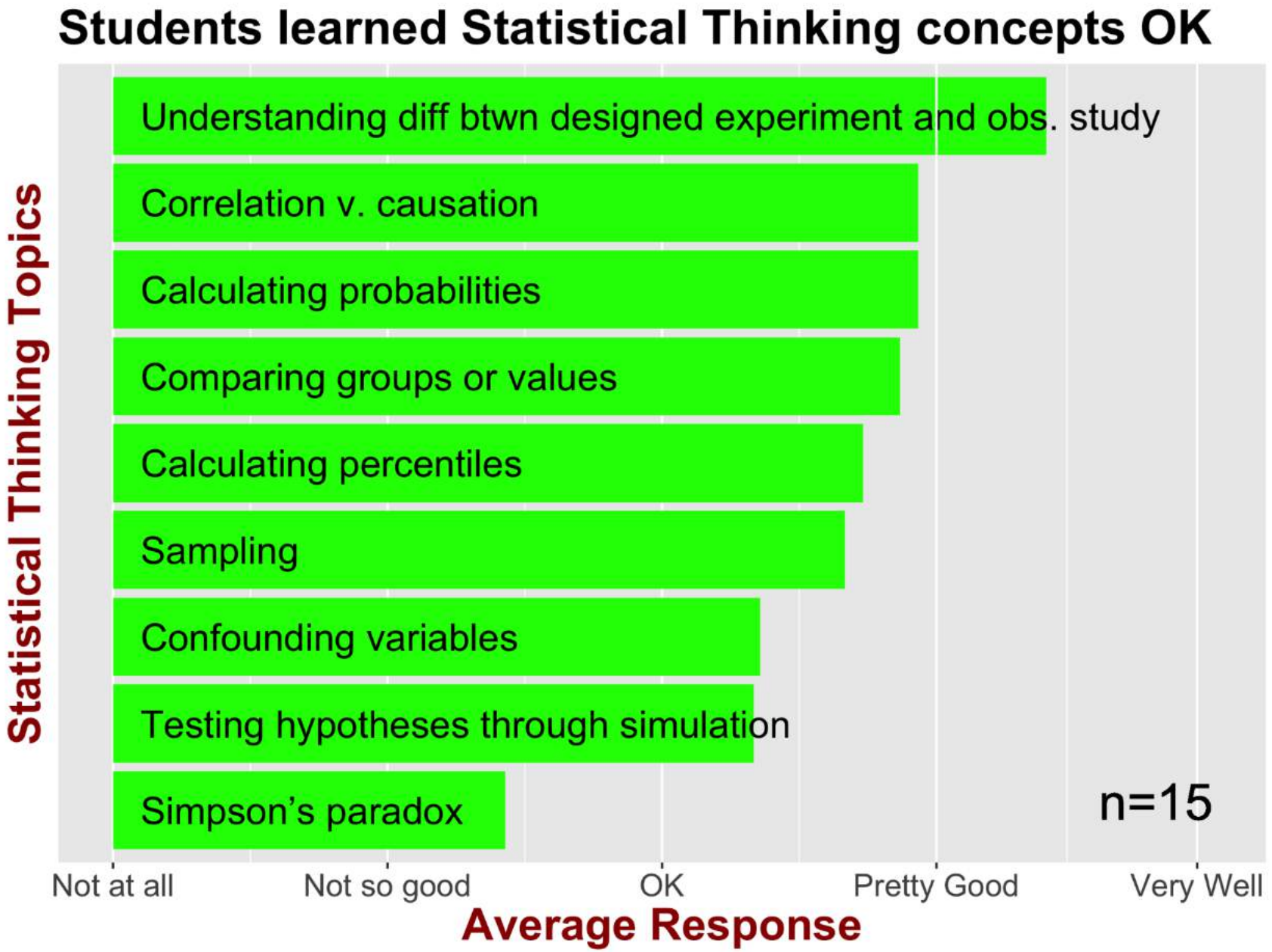
Students reported that they learned ~ half of the **R programming** concepts Very Well



Students reported that they learned **Collaboration** and **Workflow** very well



Students learned **Statistical Thinking** concepts just OK (room for improvement)



Students learned **Statistical Thinking** concepts just OK (room for improvement)

Students took a pre- and post-course Statistical Reasoning and Literacy Skills assessment (REALI)

Anelise Sabbag (2016) "Examining the Relationship Between Statistical Literacy and Statistical Reasoning"

- Scores increased <5% on average
- 4 students increased >10%
- 1 student decreased > 10%

No practical or statistically sig. improvement in REALI scores

My conclusions: **TBL in Data Science is worth doing and worth improving**

Next year I'm adding a 1-2 hour lab section
More class time will mean more time for improved exercises and lab assignments

I will prepare 15-minute mini lectures to help students approach the application exercises

Poll #3

Of the choices below, which best describes how you think about teaching Data Science, Statistical Thinking, and Collaboration with Team-Based Learning?

A. Great idea to implement right now!

B. Pretty good idea worth investigating further

C. Good idea, but I would implement it differently

D. Seems too complicated to implement right now

E. The standard class is better and easier to teach

Summary

- Collaboration skills are important for data scientists
- My students gain experience collaborating while learning R programming and (some) Statistical Thinking skills
- *Team-Based Learning* combines the best parts of flipped-classroom, project based learning with small-group learning and makes them work!
- Your results may vary

Summary

- Collaboration skills are important for data scientists
- My students gain experience collaborating while learning R programming and (some) Statistical Thinking skills
- *Team-Based Learning* combines the best parts of flipped-classroom, project based learning with small-group learning and makes them work!
- Your results may vary

Materials:
www.osf.io/xmtce

Questions?

Eric.Vance@Colorado.EDU