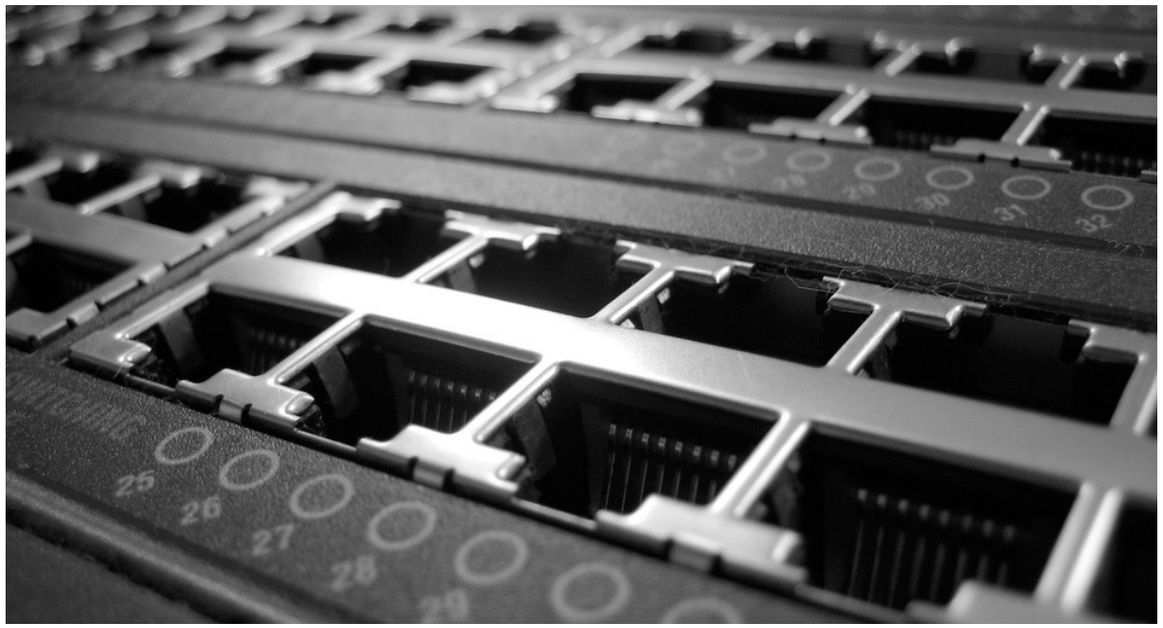


R Markdown

from RStudio

GitHub

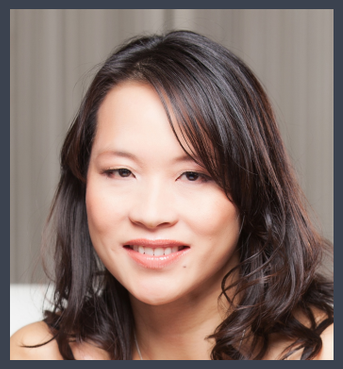
git

GitKraken

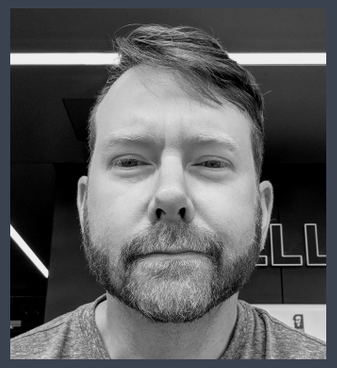
Tools for Reproducible Research

Harvard Chan Bioinformatics Core

https://tinyurl.com/hbc-trr



Shannan Ho Sui
Director



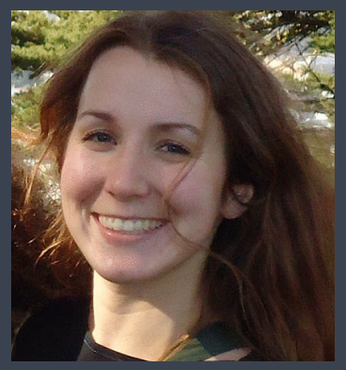
John Hutchinson
Associate Director

Victor Barrera

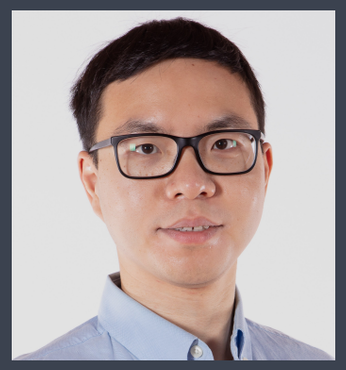
Zhu Zhuo

Preetida Bhetariya

Meeta Mistry



Mary Piper
Assoc. Training Director



Jihe Liu

Radhika Khetani
Training Director

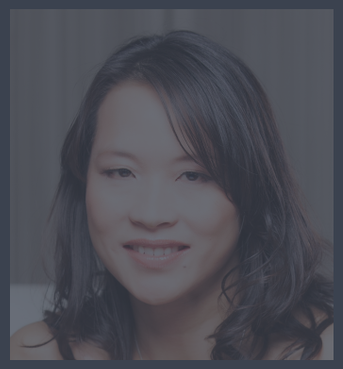
Maria Simoneau

James Billingsley

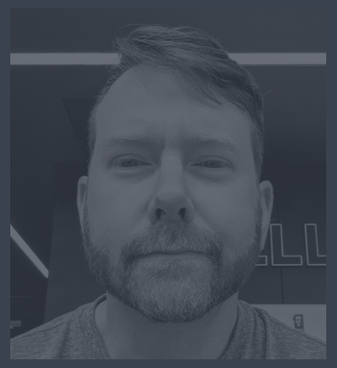
Sergey Naumenko

Joon Yoon

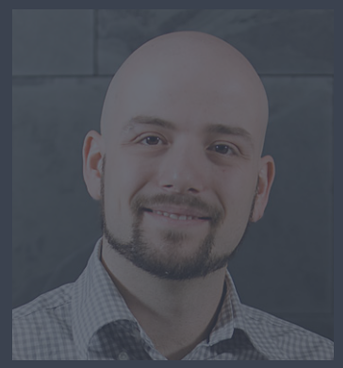
Peter Kraft
Faculty Advisor



Shannan Ho Sui
Director

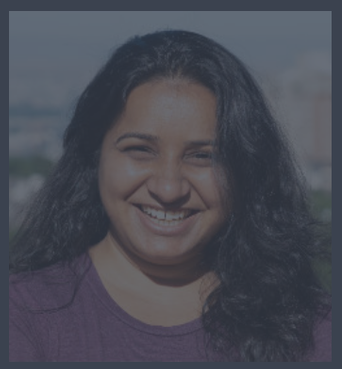


John Hutchinson
Associate Director



Victor Barrera

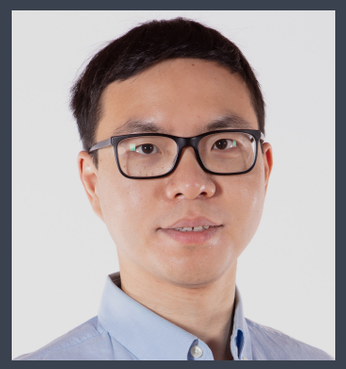
Zhu Zhuo



Preetida Bhetariya

Meeta Mistry

Mary Piper
Assoc. Training Director

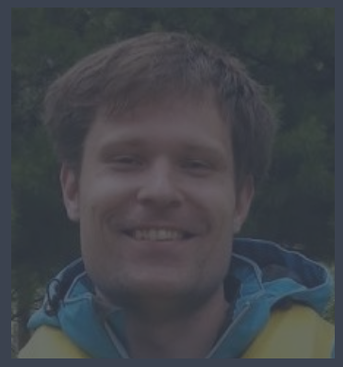


Jihe Liu

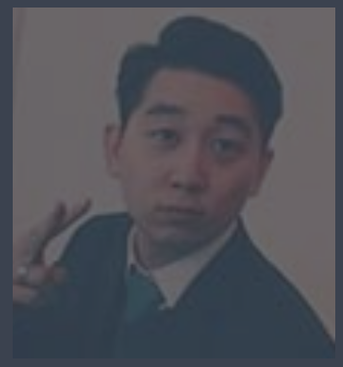
Radhika Khetani
Training Director

Maria Simoneau

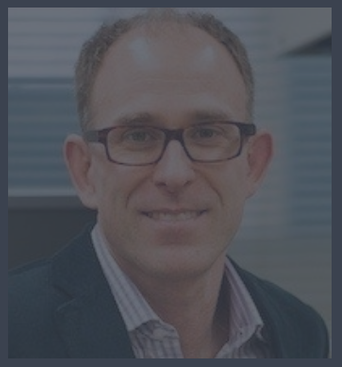
James Billingsley



Sergey Naumenko



Joon Yoon



Peter Kraft
Faculty Advisor

Consulting

- RNA-seq analysis: bulk, single cell, small RNA
- ChIP-seq and ATAC-seq analysis
- Genome-wide methylation
- WGS, resequencing, exome-seq and CNV studies
- QC & analysis of gene expression arrays
- Functional enrichment analysis
- Grant support

http://bioinformatics.sph.harvard.edu/

CFAR
HARVARD UNIVERSITY
CENTER FOR AIDS RESEARCH

HARVARD
T.H. CHAN
SCHOOL OF PUBLIC HEALTH

NIEHS

HSCI
HARVARD STEM CELL
INSTITUTE

HARVARD
CATALYST
THE HARVARD CLINICAL
AND TRANSLATIONAL
SCIENCE CENTER

HARVARD
MEDICAL SCHOOL

Training

We have divided our short workshops into 2 categories:

1. Basic Data Skills - No prior programming knowledge needed (no prerequisites)

2. Advanced Topics: Analysis of high-throughput sequencing (NGS) data - Certain "Basic" workshops required as prerequisites.

Any participants wanting to take an advanced workshop will have to have taken the appropriate basic workshop(s) within the past 6 months.

http://bioinformatics.sph.harvard.edu/training/

https://hbctraining.github.io/main/

HARVARD T.H. CHAN SCHOOL OF PUBLIC HEALTH

DF/HCC DANA-FARBER / HARVARD CANCER CENTER

HSCI HARVARD STEM CELL INSTITUTE

HARVARD CATALYST THE HARVARD CLINICAL AND TRANSLATIONAL SCIENCE CENTER

HARVARD MEDICAL SCHOOL

Training

We have divided our short workshops into 2 categories:

1. Basic Data Skills - No prior programming knowledge needed (no prerequisites)

2. Advanced Topics: Analysis of high-throughput sequencing (NGS) data - Certain “Basic” workshops required as prerequisites.

Any participants wanting to take an advanced workshop will have to have taken the appropriate basic workshop(s) within the past 6 months.

http://bioinformatics.sph.harvard.edu/training/

https://hbctraining.github.io/main/

HARVARD T.H. CHAN SCHOOL OF PUBLIC HEALTH

DF/HCC DANA-FARBER / HARVARD CANCER CENTER

HSCI HARVARD STEM CELL INSTITUTE

HARVARD CATALYST THE HARVARD CLINICAL AND TRANSLATIONAL SCIENCE CENTER

HARVARD MEDICAL SCHOOL

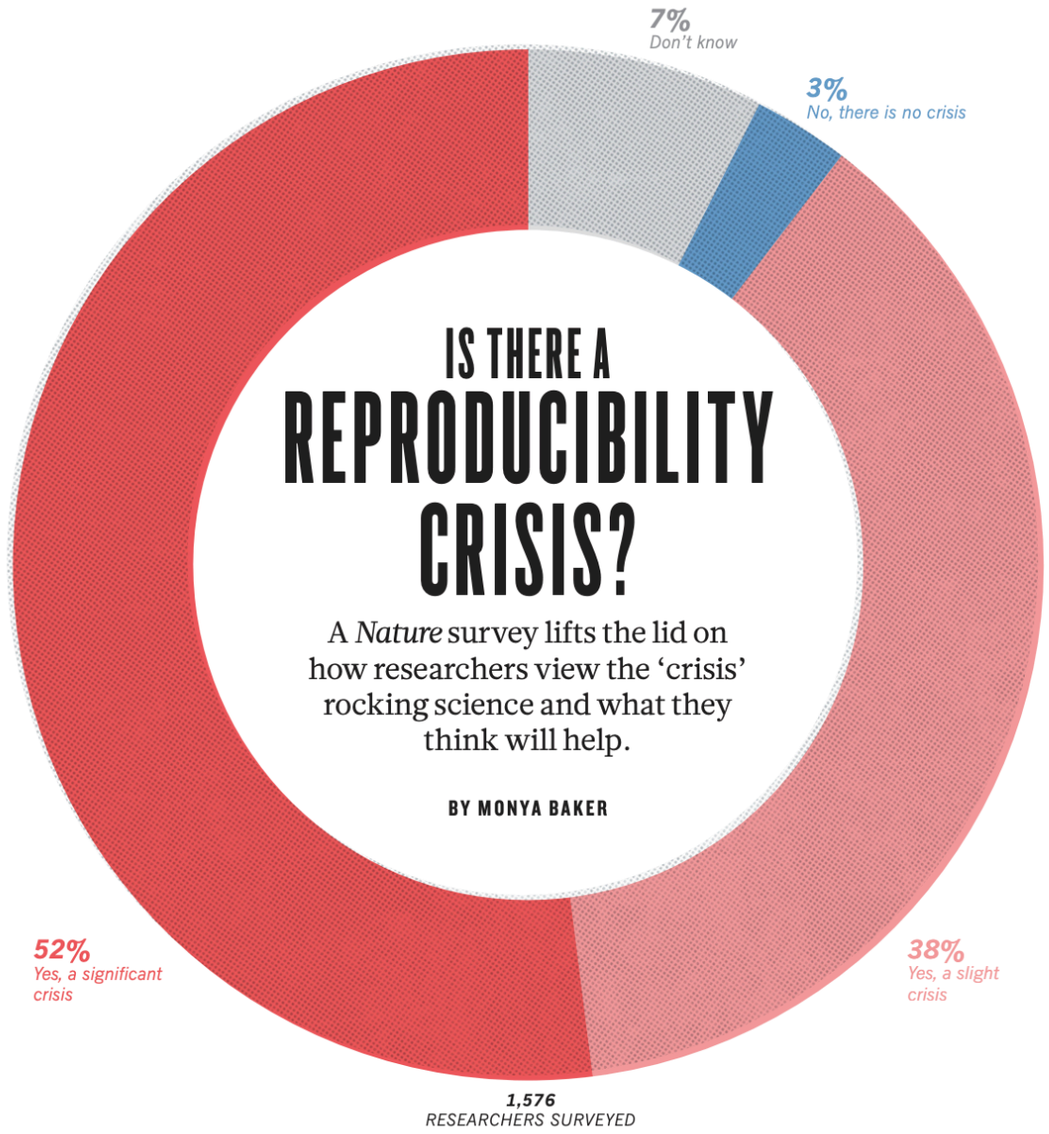
Workshop scope

Reproducibility means…

- that the methods of an experiment can be repeated?
- the results of a subsequent experiment based on those methods would generate identical results?
- that two groups analyzing the same data would reach the same conclusions?

“Reproducibility is a minimum necessary condition for a finding to be believable and informative.”

- Goodman et al, 2016



a replication study. When work does not reproduce, researchers often assume there is a perfectly valid (and probably boring) reason. What's more, incentives to publish positive replications are low and journals can be reluctant to publish negative findings. In fact, several respondents who had published a failed replication said that editors and reviewers demanded that they play down comparisons with the original study.

Nevertheless, 24% said that they had been able to publish a successful replication and 13% had published a failed replication. Acceptance was more common than persistent rejection: only 12% reported being unable to publish successful attempts to reproduce others' work; 10% reported being unable to publish unsuccessful attempts.

Survey respondent Abraham Al-Ahmad at the Texas Tech University Health Sciences Center in Amarillo expected a "cold and dry rejection" when he submitted a manuscript explaining why a stem-cell technique had stopped working in his hands. He was pleasantly surprised when the paper was accepted[3]. The reason, he thinks, is because it offered a workaround for the problem.

Others place the ability to publish replication attempts down to a combination of luck, persistence and editors' inclinations. Survey respondent Michael Adams, a drug-development consultant, says that work showing severe flaws in an animal model of diabetes has been rejected six times, in part because it does not reveal a new drug target. By contrast, he says, work refuting the efficacy of a compound to treat Chagas disease was quickly accepted[4].

"REPRODUCIBILITY IS LIKE BRUSHING YOUR TEETH. ONCE YOU LEARN IT, IT BECOMES A HABIT."

THE CORRECTIVE MEASURES

One-third of respondents said that their labs had taken concrete steps to improve reproducibility within the past five years. Rates ranged from a high of 41% in medicine to a low of 24% in physics and engineering. Free-text responses suggested that redoing the work or asking someone else within a lab to repeat the work is the most common practice. Also common are efforts to beef up the documentation and standardization of experimental methods.

Any of these can be a major undertaking. A biochemistry graduate student in the United Kingdom, who asked not to be named, says that efforts to reproduce work for her lab's projects doubles the time and materials used — in addition to the time taken to troubleshoot when some things invariably don't work. Although replication does boost confidence in results, she says, the costs mean that she performs checks only for innovative projects or unexpected results.

Consolidating methods is a project unto itself, says Laura Shankman, a postdoc studying smooth muscle cells at the University of Virginia, Charlottesville. After several postdocs and graduate students left her lab within a short time, remaining members had trouble getting consistent results in their experiments. The lab decided to take some time off from new questions to repeat published work, and this revealed that lab protocols had gradually diverged. She thinks that the lab saved money overall by getting synchronized instead of troubleshooting failed experiments piecemeal, but that it was a long-term investment.

people mentioned this strategy. One who did was Hanne Watkins, a graduate student studying moral decision-making at the University of Melbourne in Australia. Going back to her original questions after collecting data, she says, kept her from going down a rabbit hole. And the process, although time consuming, was no more arduous than getting ethical approval or formatting survey questions. "If it's built in right from the start," she says, "it's just part of the routine of doing a study."

THE CAUSE

The survey asked scientists what led to problems in reproducibility. More than 60% of respondents said that each of two factors — pressure to publish and selective reporting — always or often contributed. More than half pointed to insufficient replication in the lab, poor oversight or low statistical power. A smaller proportion pointed to obstacles such as variability in reagents or the use of specialized techniques that are difficult to repeat.

But all these factors are exacerbated by common forces, says Judith Kimble, a developmental biologist at the University of Wisconsin–Madison: competition for grants and positions, and a growing burden of bureaucracy that takes away from time spent doing and designing research. "Everyone is stretched thinner these days," she says. And the cost extends beyond any particular research project. If graduate students train in labs where senior members have little time for their juniors, they may go on to establish their own labs without having a model of how training and mentoring should work. "They will go off and make it worse," Kimble says.

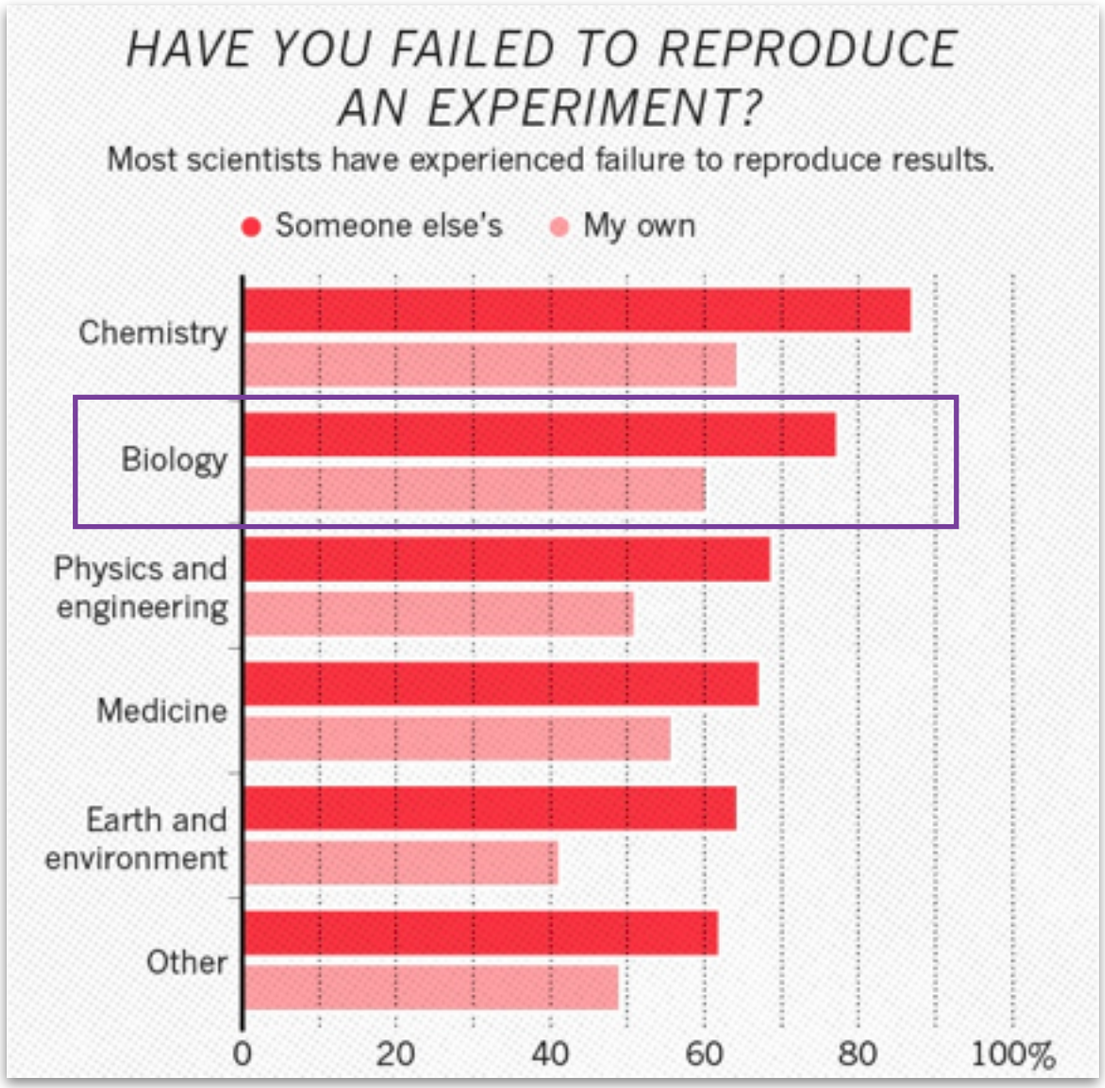
WHAT CAN BE DONE?

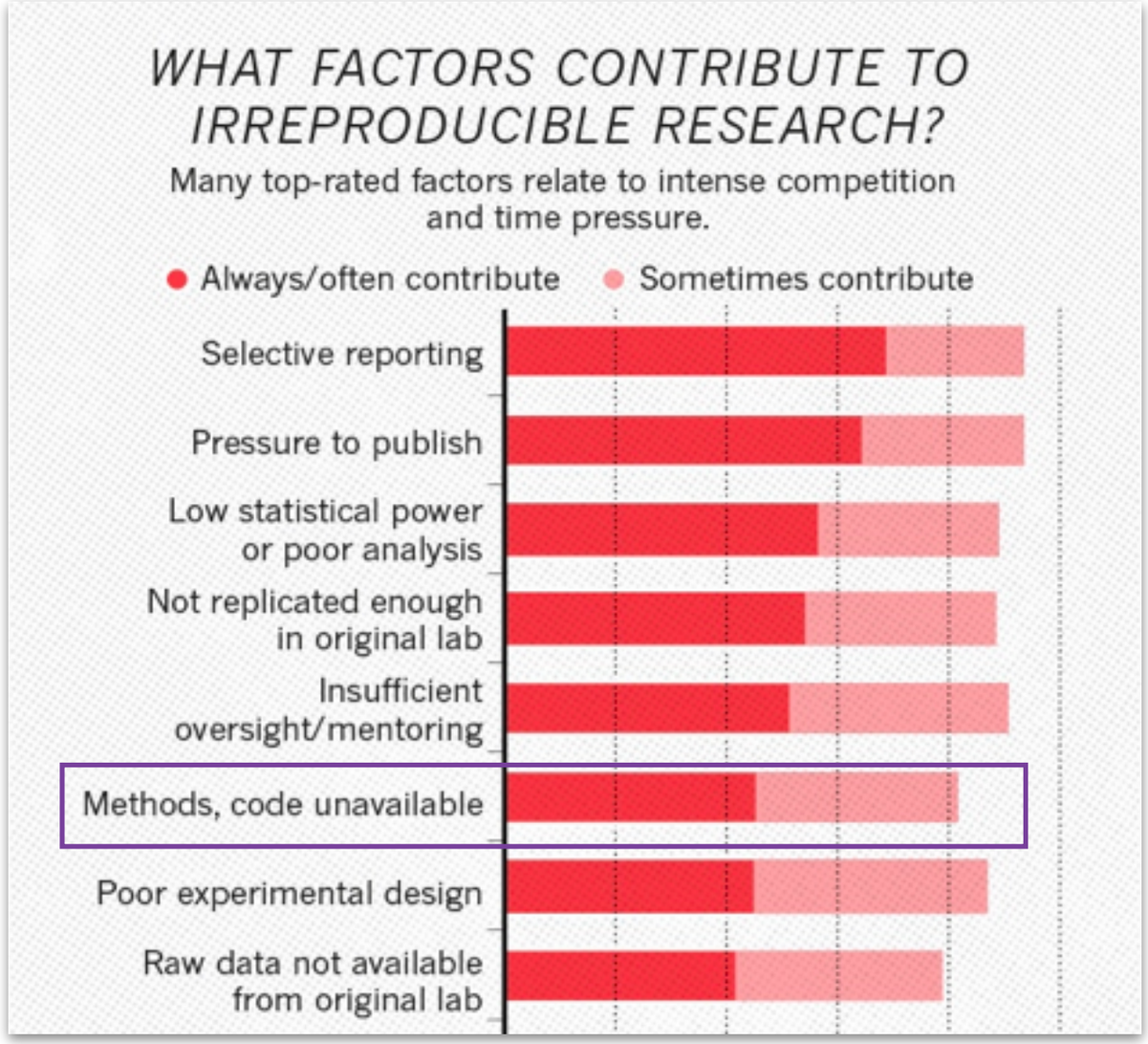
Respondents were asked to rate 11 different approaches to improving reproducibility in science, and all got ringing endorsements. Nearly 90% — more than 1,000 people — ticked "More robust experimental design" "better statistics" and "better mentorship". Those ranked higher than the option of providing incentives (such as funding or credit towards tenure) for reproducibility-enhancing practices. But even the lowest-ranked item — journal checklists — won a whopping 69% endorsement.

The survey — which was e-mailed to *Nature* readers and advertised on affiliated websites and social-media outlets as being 'about reproducibility' — probably selected for respondents who are more receptive to and aware of concerns about reproducibility. Nevertheless, the results suggest that journals, funders and research institutions that advance policies to address the issue would probably find cooperation, says John Ioannidis, who studies scientific robustness at Stanford University in California. "People would probably welcome such initiatives." About 80% of respondents thought that funders and publishers should do more to improve reproducibility.

"It's healthy that people are aware of the issues and open to a range of straightforward ways to improve them," says Munafo. And given that these ideas are being widely discussed, even in mainstream media, tack-

Baker, Monya. 2016. "1,500 scientists lift the lid on reproducibility." *Nature News* 533(7604): 452-454. https://dx.doi.org/10.1038/533452a





Baker, Monya. 2016. "1,500 scientists lift the lid on reproducibility." *Nature News* 533(7604): 452-454. https://dx.doi.org/10.1038/533452a

Adapted from Goldman and Obrycki, December 2020

What can I do?

- Organize your data!
  - Make plans for appropriate storage of your data, both raw and processed
  - Have project directories created before starting a project
  - Keep data organized during the analysis

You can't have any sort of reproducibility without good data management.

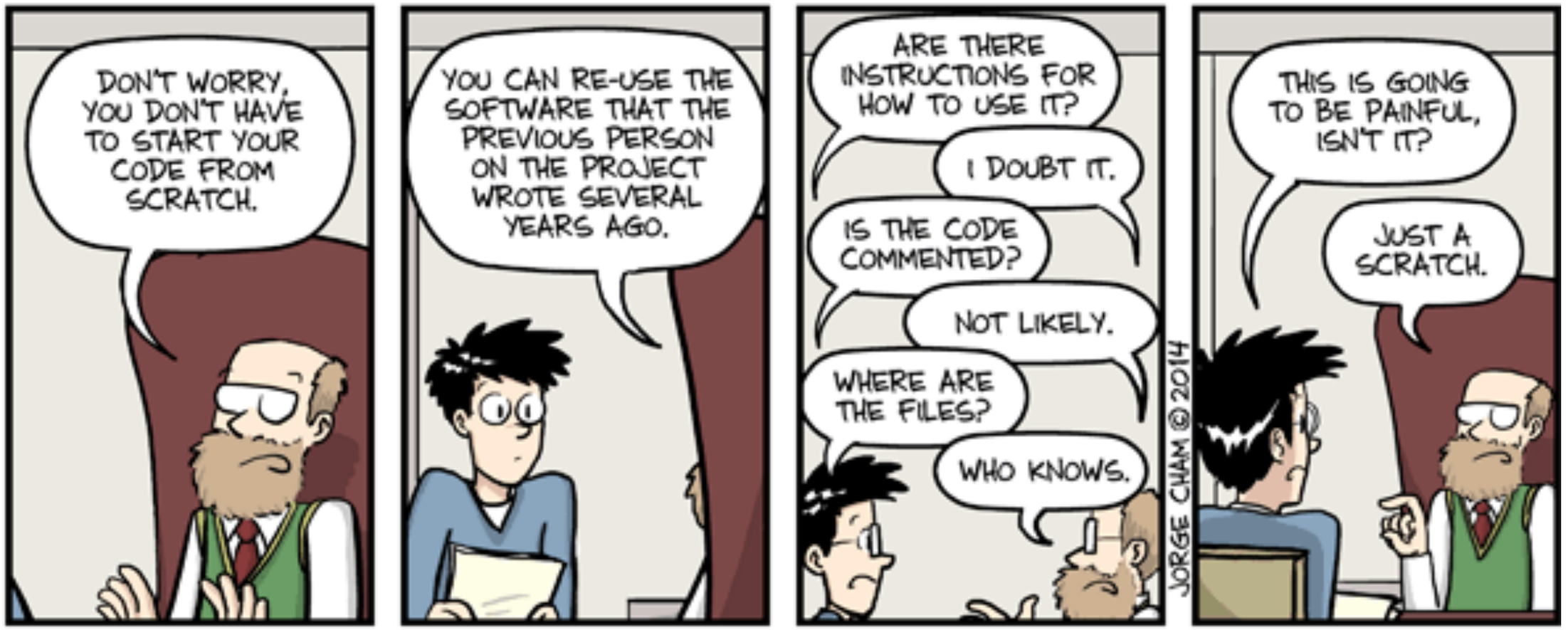
Adapted from Goldman, June 2020 and Goldman and Obrycki, December 2020

What can I do?

- Document everything!
  - Create READMEs that detail data organization, analysis methods, dates, naming conventions, etc.
  - Which tools and parameters have you tried, what were the version numbers? What were the results in each case?
  - What were the exact commands you ran throughout the workflow?
  - Annotate your code with comments
  - Use version control to track code updates and changes to other text-based documentation

Adapted from Goldman, June 2020

Piled Higher and Deeper *by Jorge Cham*



title: "Scratch" - originally published 3/12/2014 WWW.PHDCOMICS.COM

http://phdcomics.com/comics.php?f=1689

BIOMEDICAL RESEARCH DATA LIFECYCLE

https://datamanagement.hms.harvard.edu/

Tools for documentation and version control

- Documentation - Rmarkdown, Jupyter notebooks
- Version Control - Git, Subversion, Bitbucket
- Collaboration and Version Control - GitHub, Bitbucket
- Containerization to preserve workflows, tools and versions - Docker

Learning Objectives

- ✓ Describe methods for documenting computational analyses
- ✓ Generate reports for R-based analyses using RMarkdown
- ✓ Track changes using the Git version control system and the GitKraken tool
- ✓ Collaborate effectively, and disseminate code & other documents using Github

Logistics

Course webpage

https://tinyurl.com/hbc-trr

Course schedule online

Workshop Schedule

Day 1

Time	Topic	Instructor
09:30 - 10:10	Workshop Introduction	Radhika
10:10 - 10:55	RMarkdown Basics	TBD
10:55 - 11:00	Break	
11:00 - 11:45	RMarkdown Intermediate	TBD
11:45 - 12:00	Assignment review	Jihe

Assignment #1

- Practice with RMarkdown
- Upload the files requested in the above exercise to Dropbox **day before the next class**.

Course materials online

Learning Objectives

- Use code chunk options to customize the report
- Describe how to add figures and tables to an RMarkdown
- Describe how to specify the output format for RMarkdown

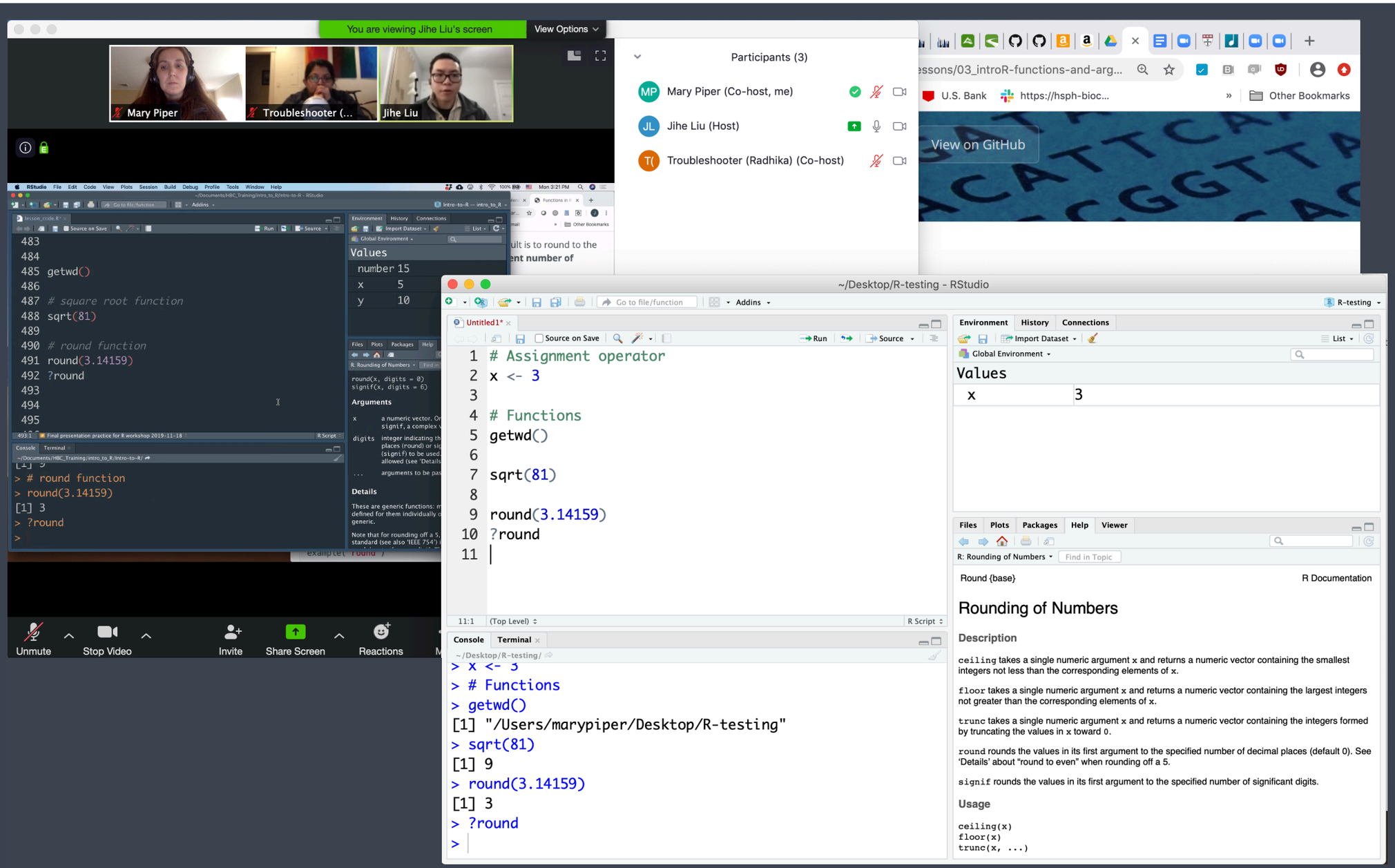
More about Code chunks

By this point, we have mentioned the word "knit" quite a few times, and you have installed and loaded the `knitr` package too. But, we have not yet fully defined what it is. knitr is an R package, developed by Yihui Xie, designed to convert RMarkdown and a couple of other file formats into a final report document in HTML or PDF or other formats.

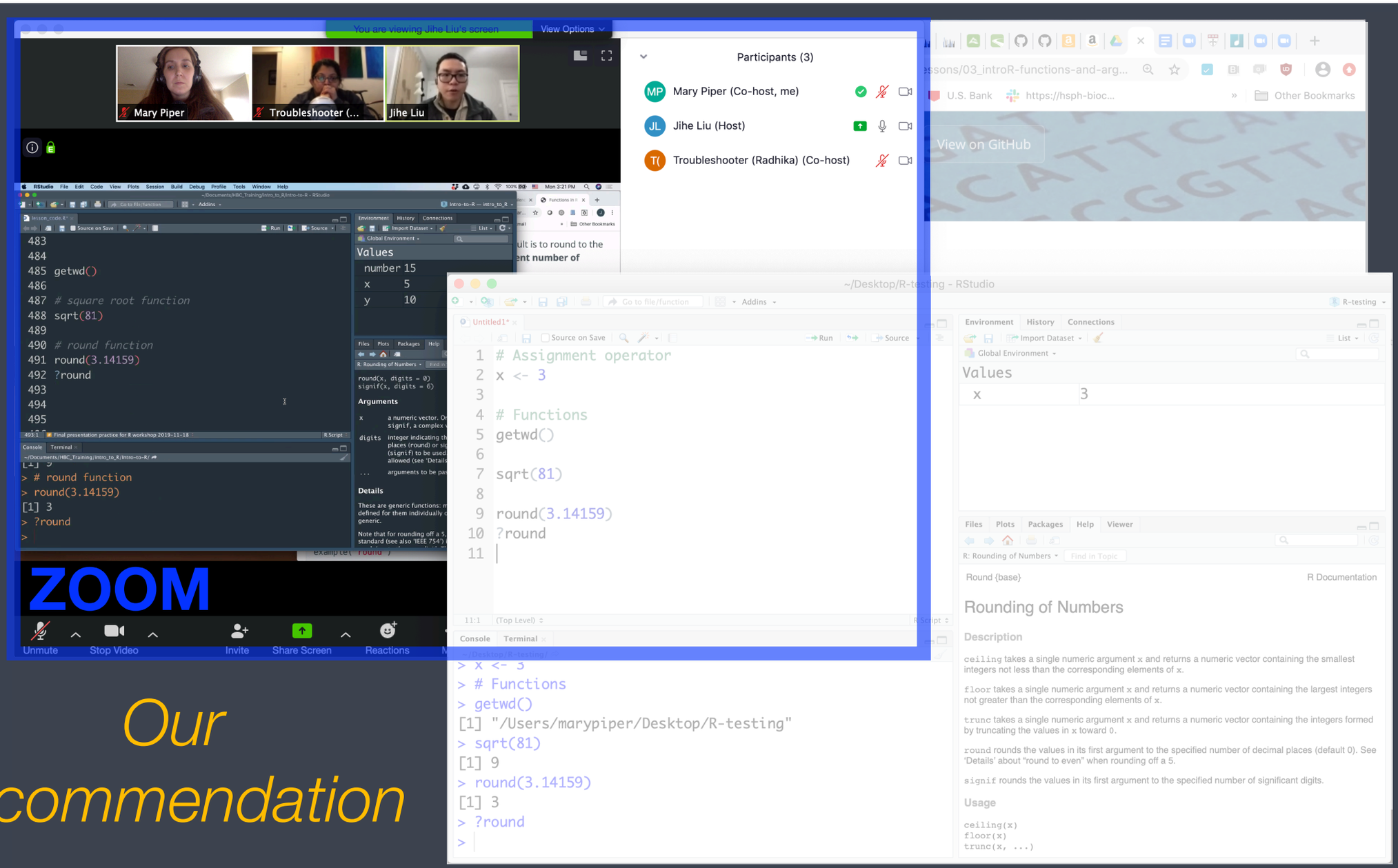
The `knitr` package provides a lot of customization options for code chunks embedded within the file. These options are written in the form of `tag=value`.

```
```{r chunk-name, echo=FALSE, warning=FALSE, message=FALSE}
x <- 4
y <- 2
x + y
```
```

Single screen & 3 windows?

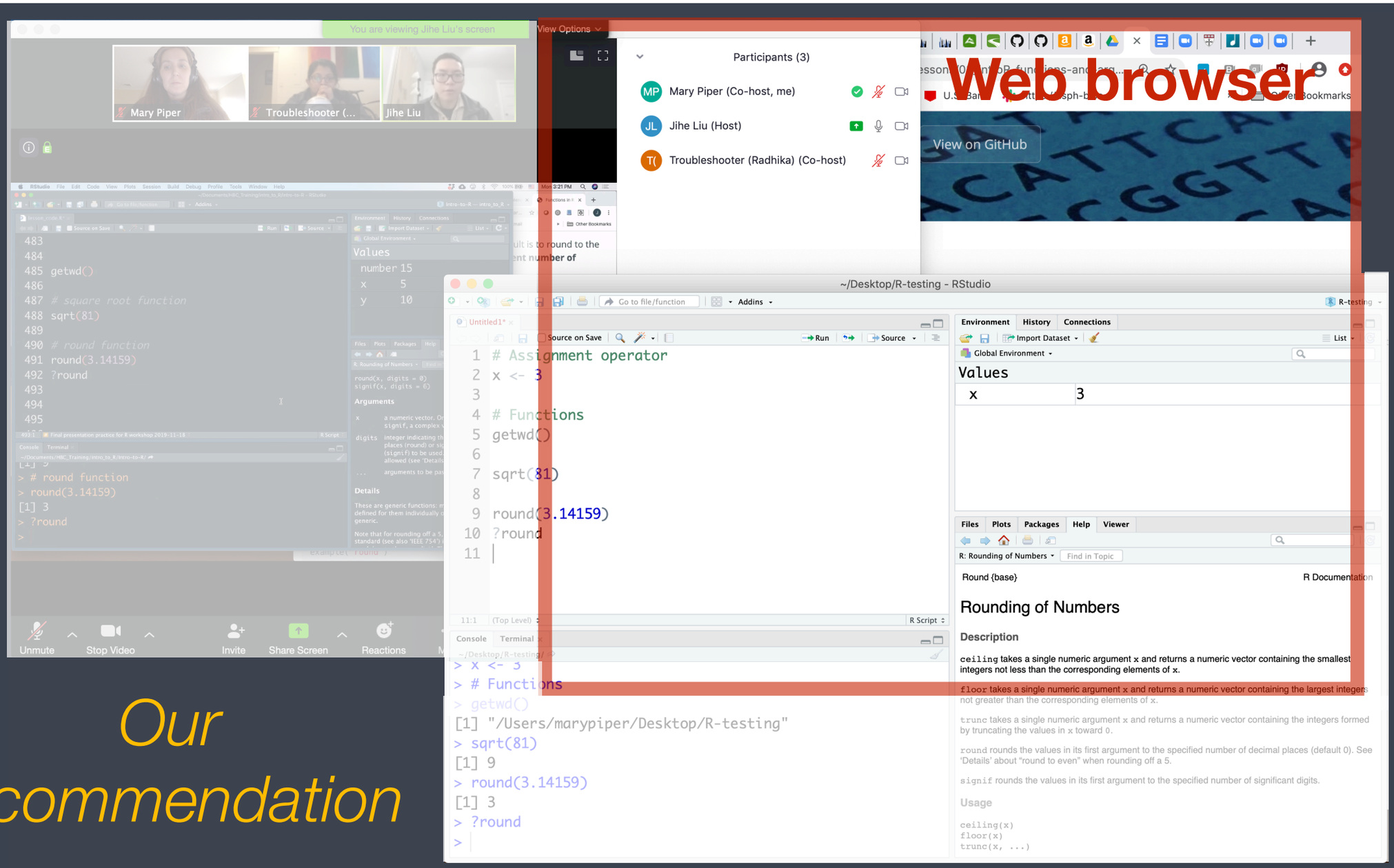


Single screen & 3 windows?



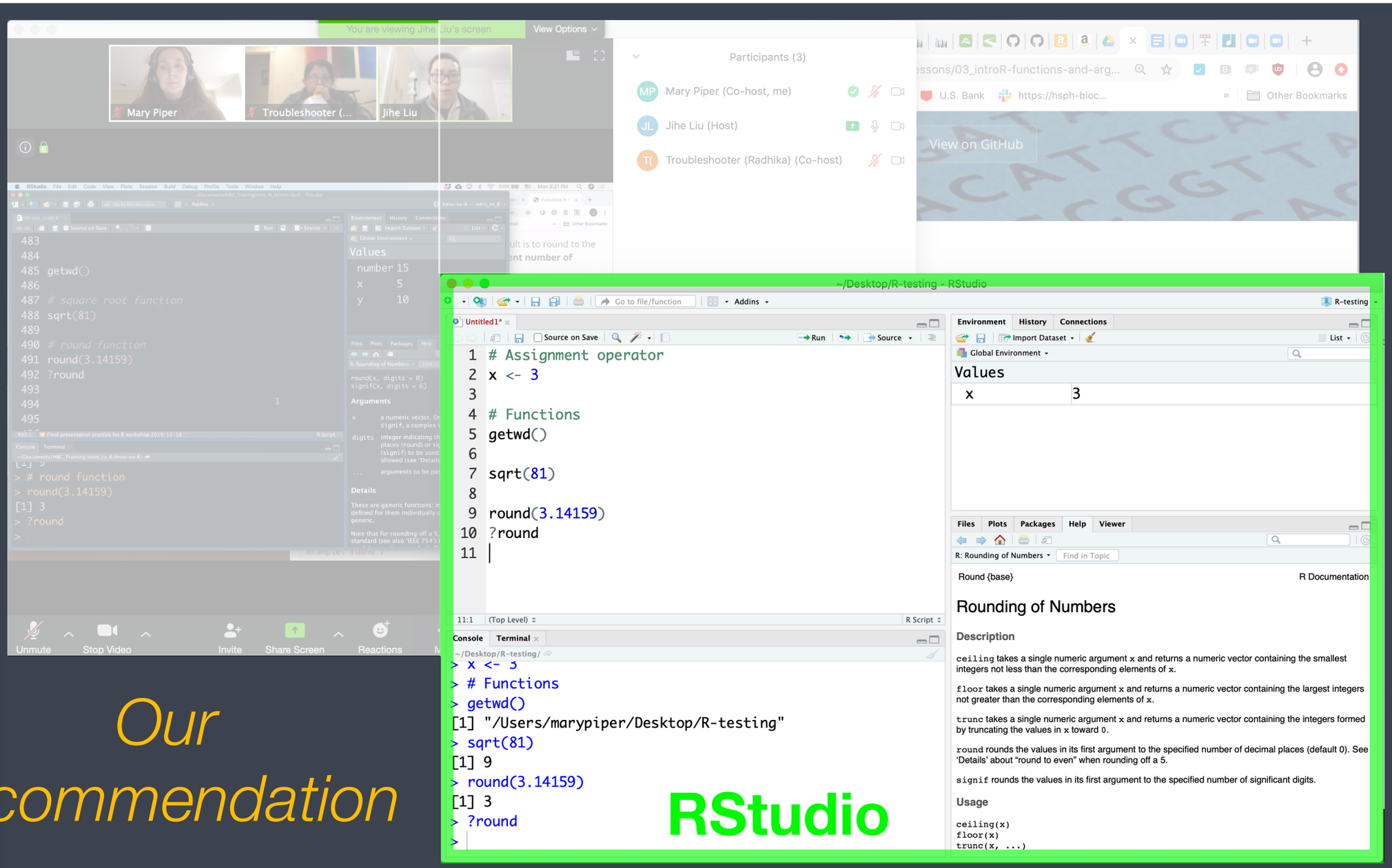
Our recommendation

Single screen & 3 windows?



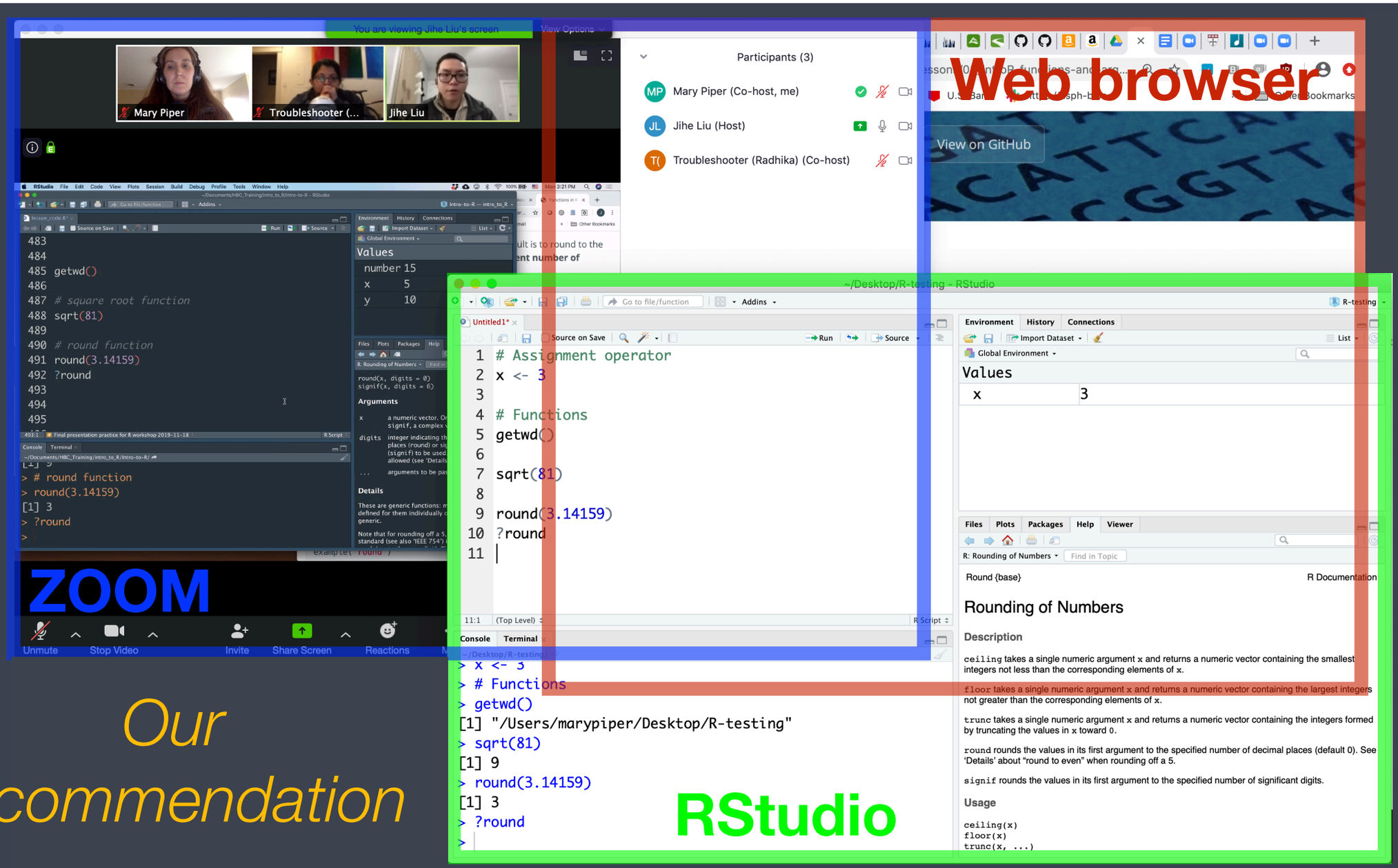
Our recommendation

Single screen & 3 windows?



Our recommendation

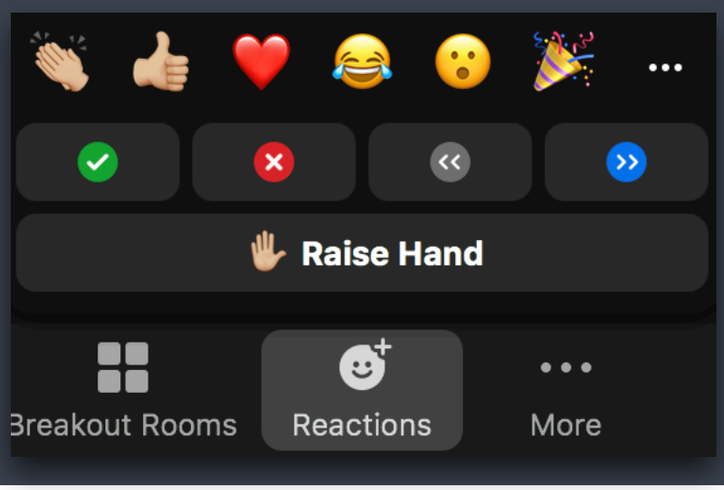
Single screen & 3 windows?



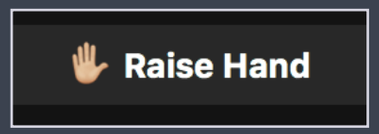
Our recommendation

Odds and Ends

- Quit/minimize all applications that are not required for class
- Are you all set?
  -  = "agree", "I'm all set" (equivalent to a **green post-it**)
  -  = "disagree", "I need help" (equivalent to a **red post-it**)

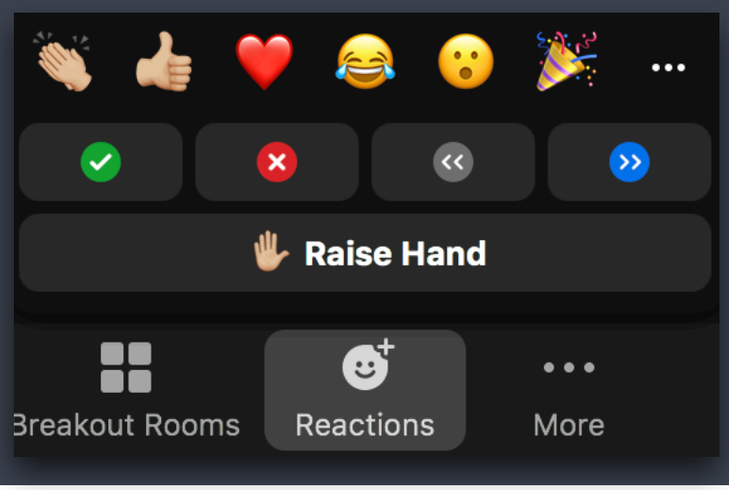


Odds and Ends (2/2)

- Questions for the presenter?
  - Post the question in the Chat window OR
  - 

 when the presenter asks for questions

- Technical difficulties with software?
  - Start a *private* chat with the *Troubleshooter* with a description of the problem.



Thanks!

- Julie Goldman, Countway Library
- John F. Obrycki, Boston Children's Hospital
- Data Carpentry

Contact us!

HBC training team: hbctraining@hsph.harvard.edu

O2 (HMS-RC): rchelp@hms.harvard.edu

HBC consulting: bioinformatics@hsph.harvard.edu

Twitter

HBC: @bioinfocore

HMS-RC: @hms_rc