NTRES 6940:

Collaborative and Reproducible Data Science in R

Why are we here?

What are we going to do?

Who are we?

Why are we here?

What are we going to do?

Who are we?

And then we'll get situated with R/RStudio

Reproducible/open science

”The practice of distributing all data, software source code, and tools required to reproduce the results discussed in a research publication”

Definition from Colorado State University Libraries: shorturl.at/btN34

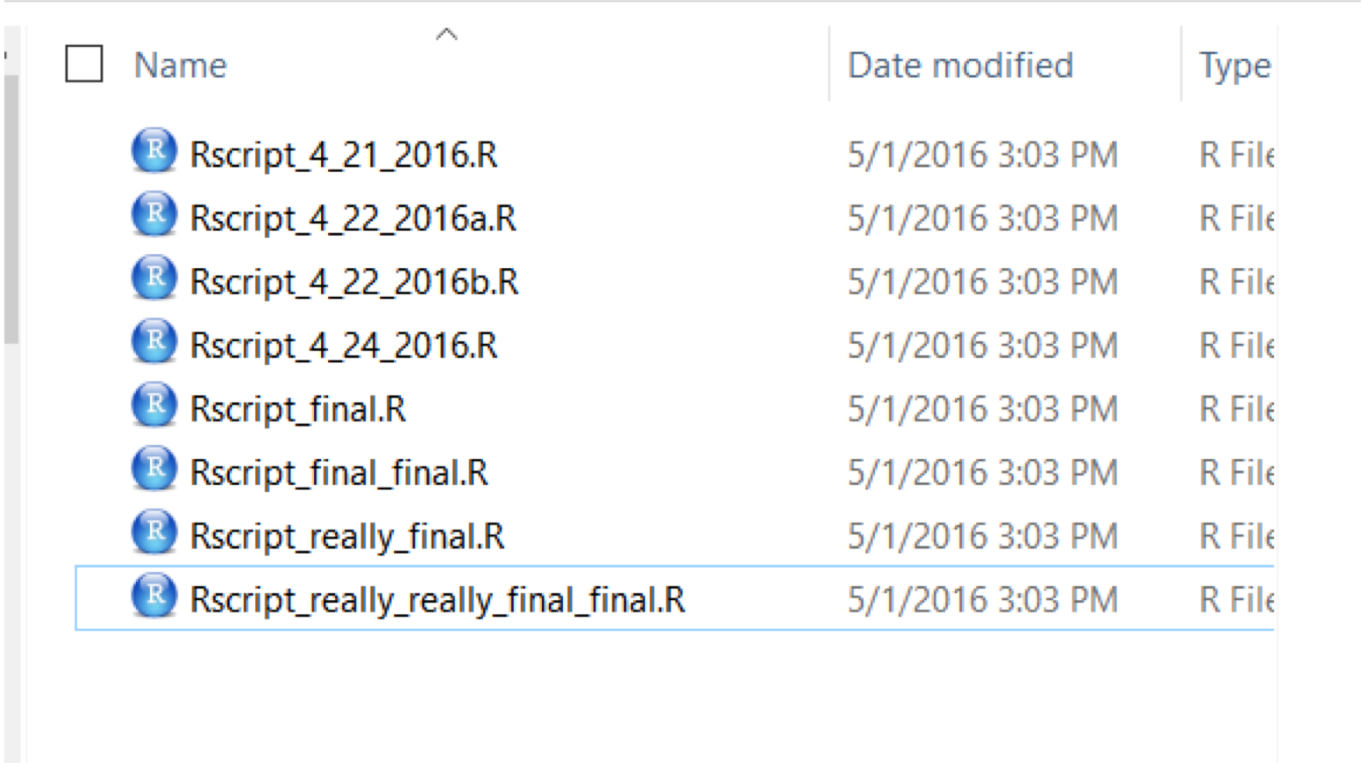
Discuss with your neighbor

- Is reproducibility the same as replicability?
- Why is there an increasing push towards open science now?
- Why should you adopt open science practices? Who will benefit and how?

Replication vs. Reproducibility

- **Replication:** the confirmation of results and conclusions from one study obtained independently in another (if a phenomenon is true, it should show up again and again)
- **But some studies can't be replicated:** too big, too costly, too time consuming, one time event, rare samples
- **Reproducibility:** minimum standard for assessing the value of scientific claims, particularly when full independent replication of a study is not feasible

Does this look familiar?



Who benefits from open science practices?

- YOU!
  - Future you will thank you
  - Increased research efficiency
- The scientific community
  - More transparency, easier to build off each other's work
- Society
  - More accurate science: errors are more likely to get detected

Tools for

better science in less time

Tools for

better science in less time

and with less pain

Who are we?

Nina Overgaard Therkildsen
(instructor)

Maria Akopyan
(TA)

Let’s get to know each other

- Here we are: https://docs.google.com/presentation/d/1UrV3oB-Zkl-CWfV2DDXUR9pXPK-dOG3FSwhD_zfbdzs/edit#slide=id.g93fda35b8d_0_62

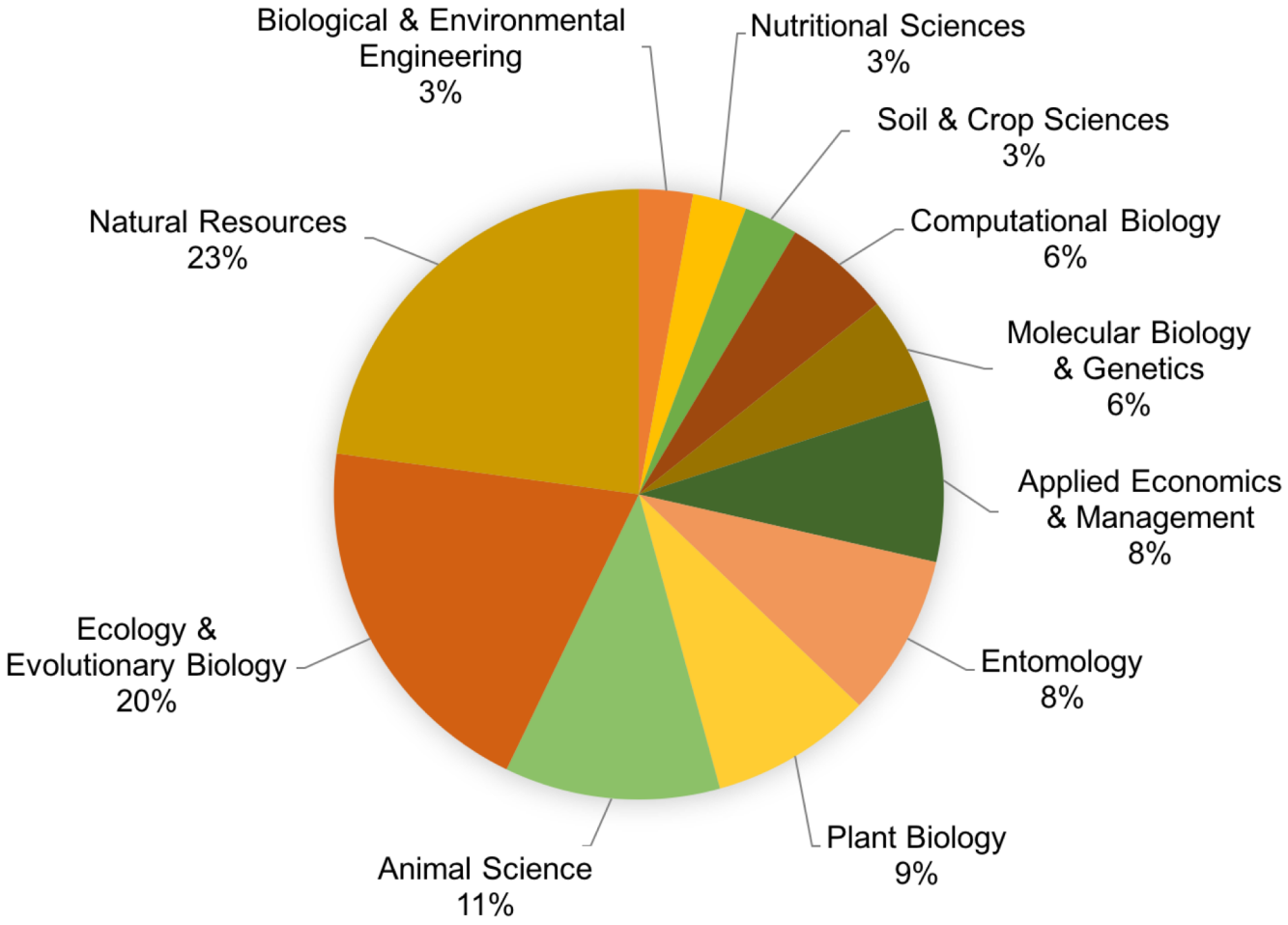
- What do we have in common?

Thanks for completing the pre-course survey!

42 respondents

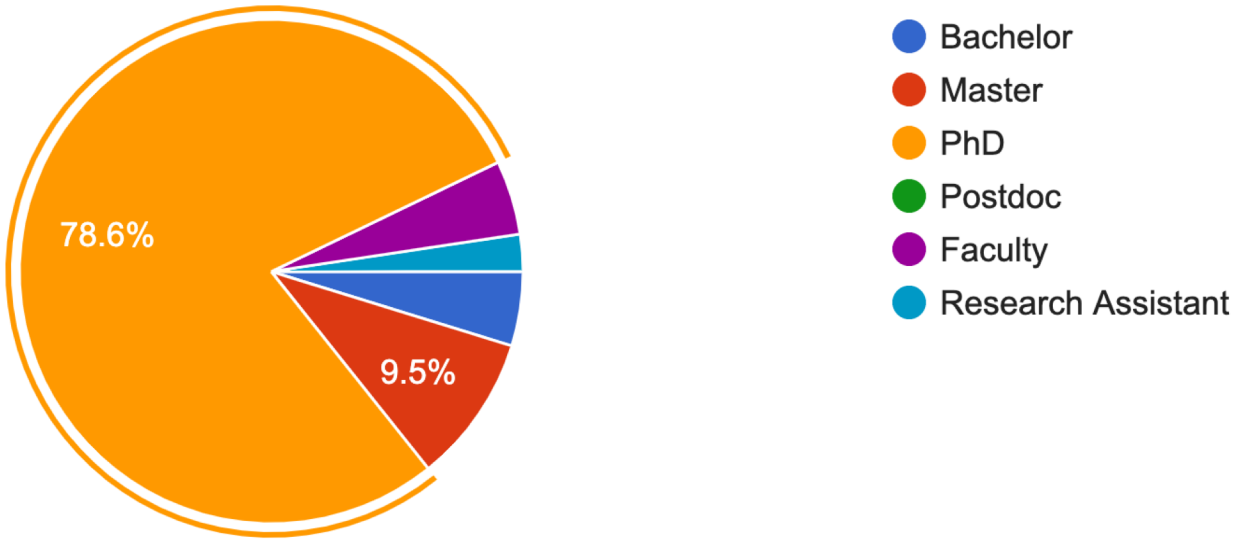
Departments

n = 35 respondents



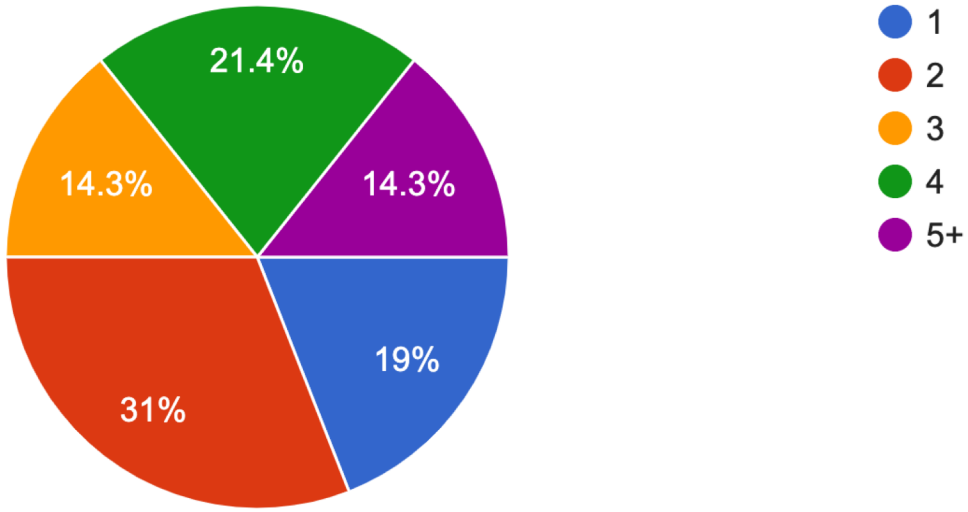
Degree program or position

42 responses



Year in your program/position

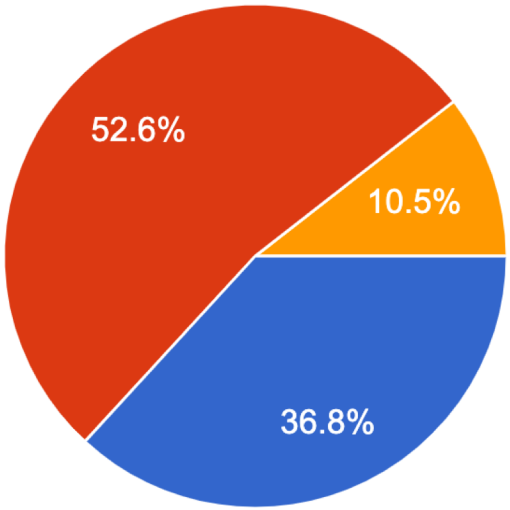
42 responses



Participation in live classes

What is your preference for breakout rooms?

38 responses

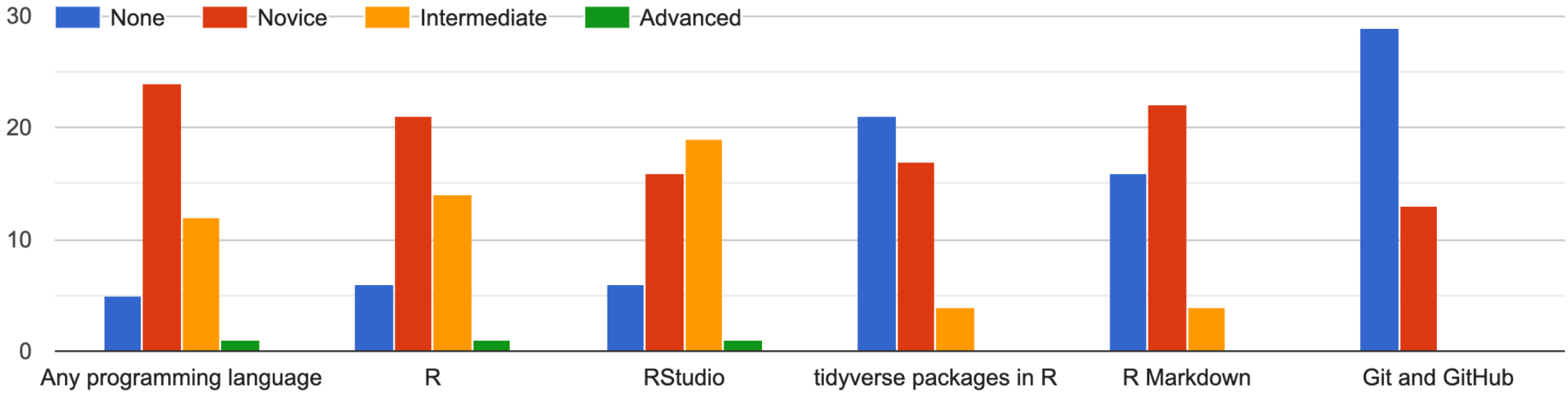


- I would prefer to be in a breakout room where people have their cameras on and actively work on exercises together
- I would prefer to be a breakout room where everyone works on the exercise alone, but can ask others if they need help
- I would prefer to just work on exercises on my own and will probably not participate in group discussions in breakout rooms

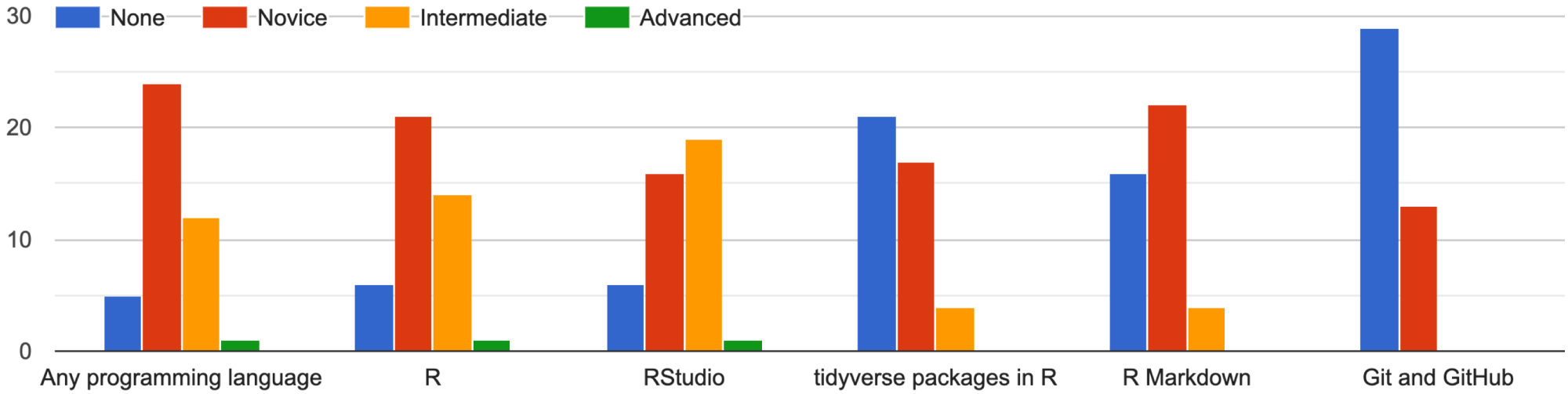
Dataset preference for demo'ing methods and practicing our skills

42 responses

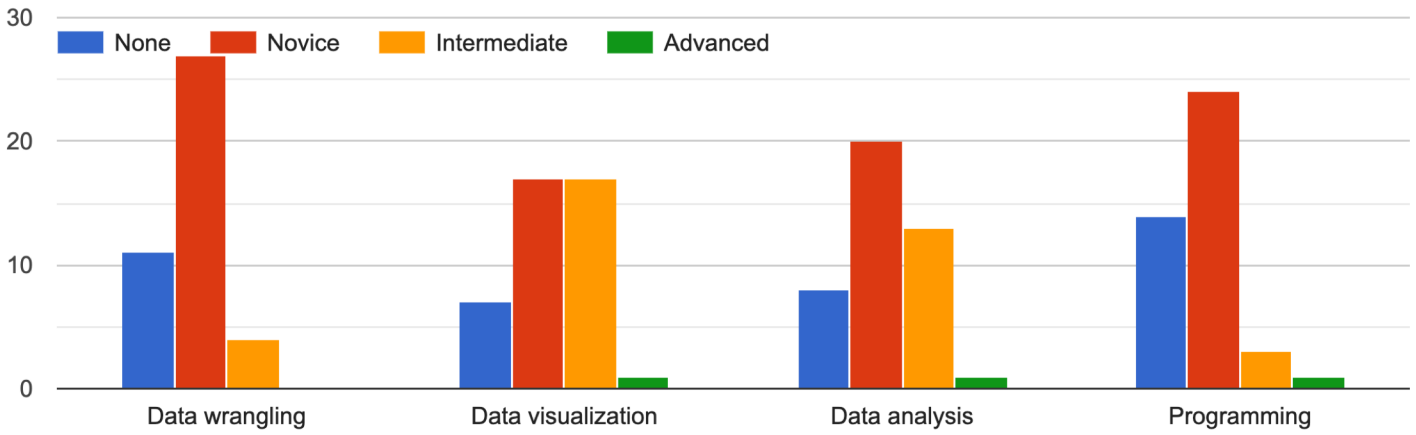
Please describe your experience level with



Please describe your experience level with



How would you describe your experience level with performing the following tasks in R:



You are all welcome here!

Code of conduct

We are dedicated to providing a welcoming and supportive environment for everyone, regardless of background, identity and prior experience level. Everyone in this course will be coming from a different place with different experiences and expectations.

We will not tolerate any form of language or behavior used to exclude, intimidate, or cause discomfort.

Tools for

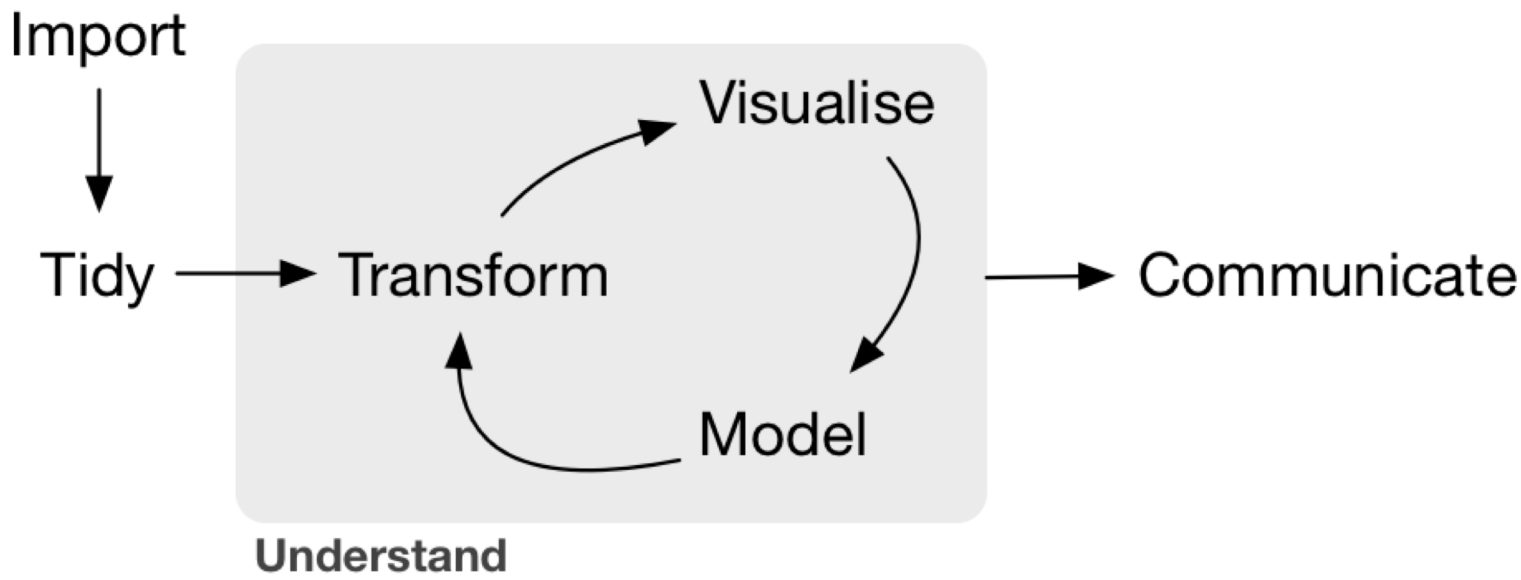
better science in less time

and with less pain

Data Science

Turning raw data into understanding, insight, and knowledge

Data Science



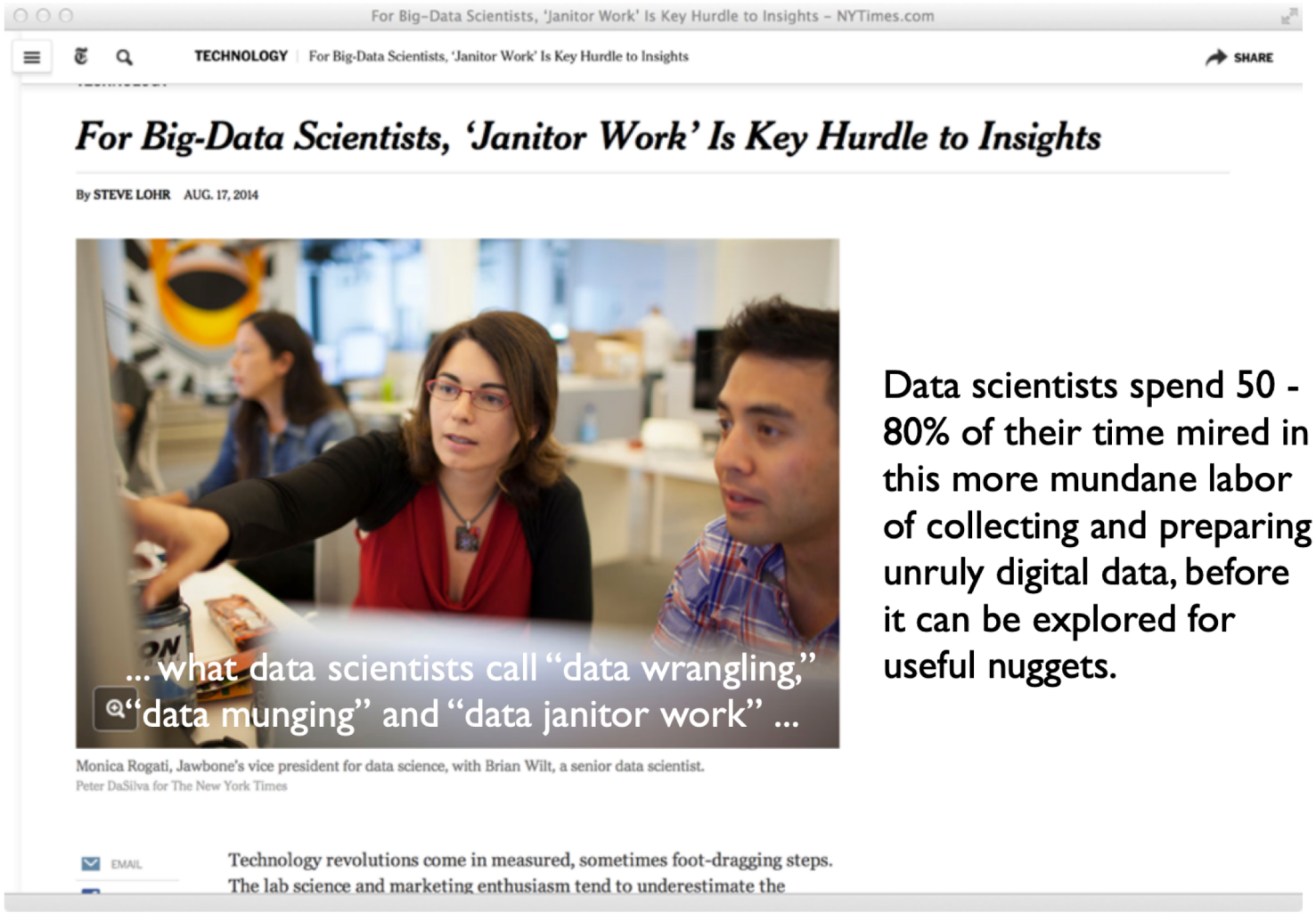
Learning outcomes

By the end of this course, students will be able to

- Describe strategies for ensuring that their data analysis is reproducible
- Demonstrate best practices for coding and project-oriented workflows in RStudio
- Import and clean messy data files using a variety of packages and functions in R
- Subset, reorganize, and merge diverse datasets in R
- Effectively explore and visualize patterns in complex datasets with ggplot in R
- Write simple functions/programs and data analysis pipelines in R
- Automate repeated analysis tasks in R
- Track the history of file changes (version control) and collaborate effectively on scripts with others with Git and GitHub
- Use R Markdown to combine text, equations, code, tables, and figures into reports, websites, and presentations

What we will NOT cover

- Statistics and hypothesis confirmation

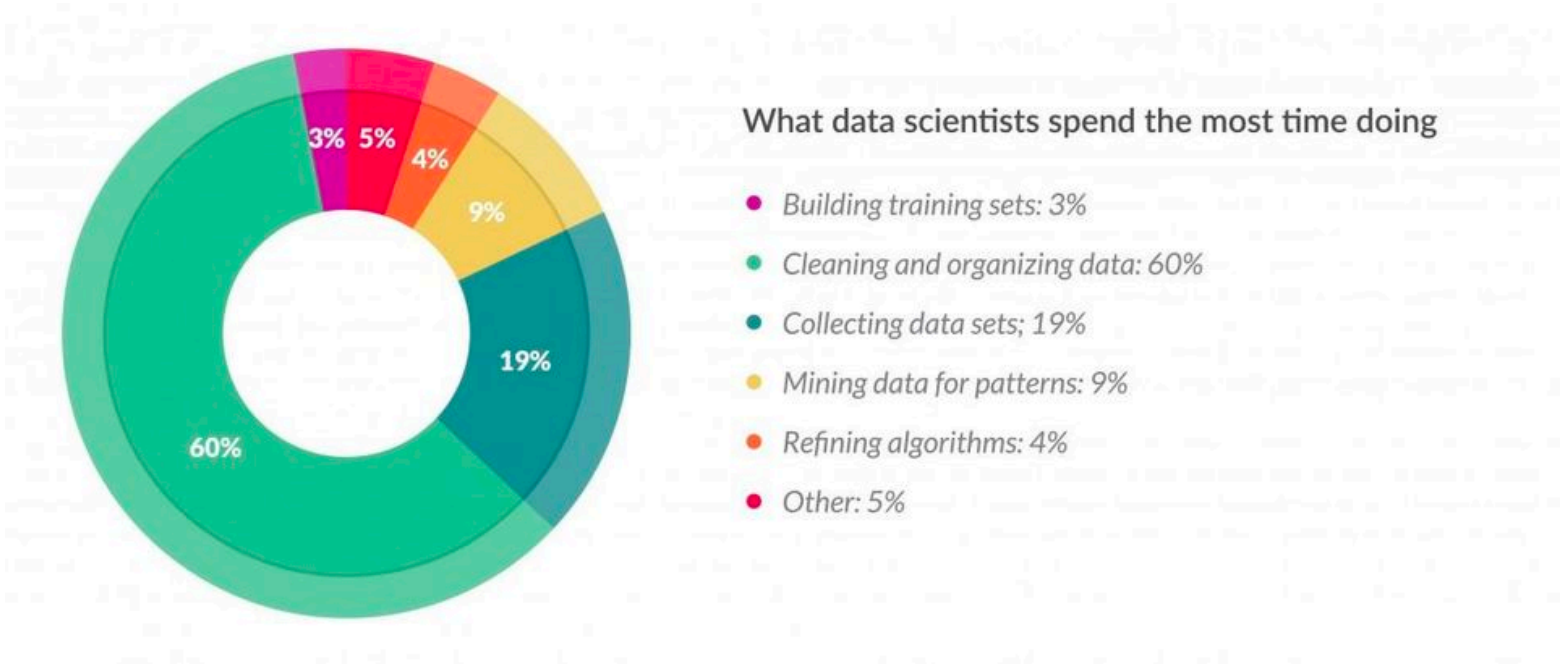


Data scientists spend 50 - 80% of their time mired in this more mundane labor of collecting and preparing unruly digital data, before it can be explored for useful nuggets.

http://www.nytimes.com/2014/08/18/technology/for-big-data-scientists-hurdle-to-insights-is-janitor-work.html?partner=rss&emc=rss&smid=tw-nytimesscience&_r=0

Survey of 80 data scientists

https://www.forbes.com/sites/gilpress/2016/03/23/data-preparation-most-time-consuming-least-enjoyable-data-science-task-survey-says/#25167ec06f63



Goals for data wrangling

- Understand how and why to tidy data and analyze tidy data, rather than making your analyses accommodate messy data
- Appreciate how there is a lot of decision-making involved with data analysis, and a lot of creativity
- Think ahead instead of only to get a single job done now
- Increase efficiency in your science and increase reproducibility
- Facilitate collaboration with others — especially Future You!

What we will NOT cover

- Statistics and hypothesis confirmation
- Modeling and simulation
- Big data

Big data

- Before we can handle big data, we need to handle small data
  - Tools can handle 100s Mb of data (up to 1–2Gb)
  - Many big data problems are small data problems in disguise
    - Subset, subsample, summary
    - Parallel analysis on multiple independent units?

What we will NOT cover

- Statistics and hypothesis confirmation
- Modeling and simulation
- Big data
- Any other programming languages than R

Why R?

- It’s free, open source, and available on every major platform
- A massive set of packages for statistical modelling, machine learning, visualization, and importing and manipulating data
- Cutting edge tools
- Supportive and welcoming community
- Powerful tools for communicating your results

What we will NOT cover

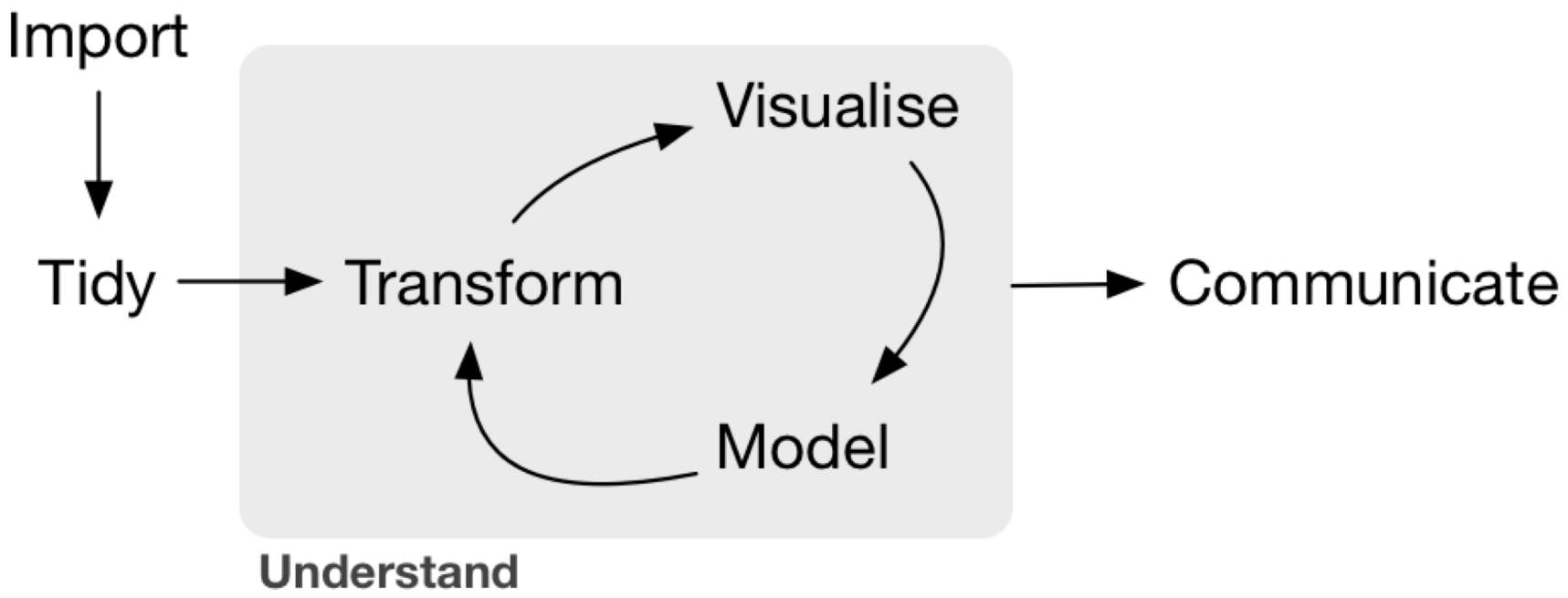
- Statistics and hypothesis confirmation
- Modeling and simulation
- Big data
- Any other programming languages than R
- Non-rectangular data (e.g. images, sounds, trees, text) or domain-specific applications

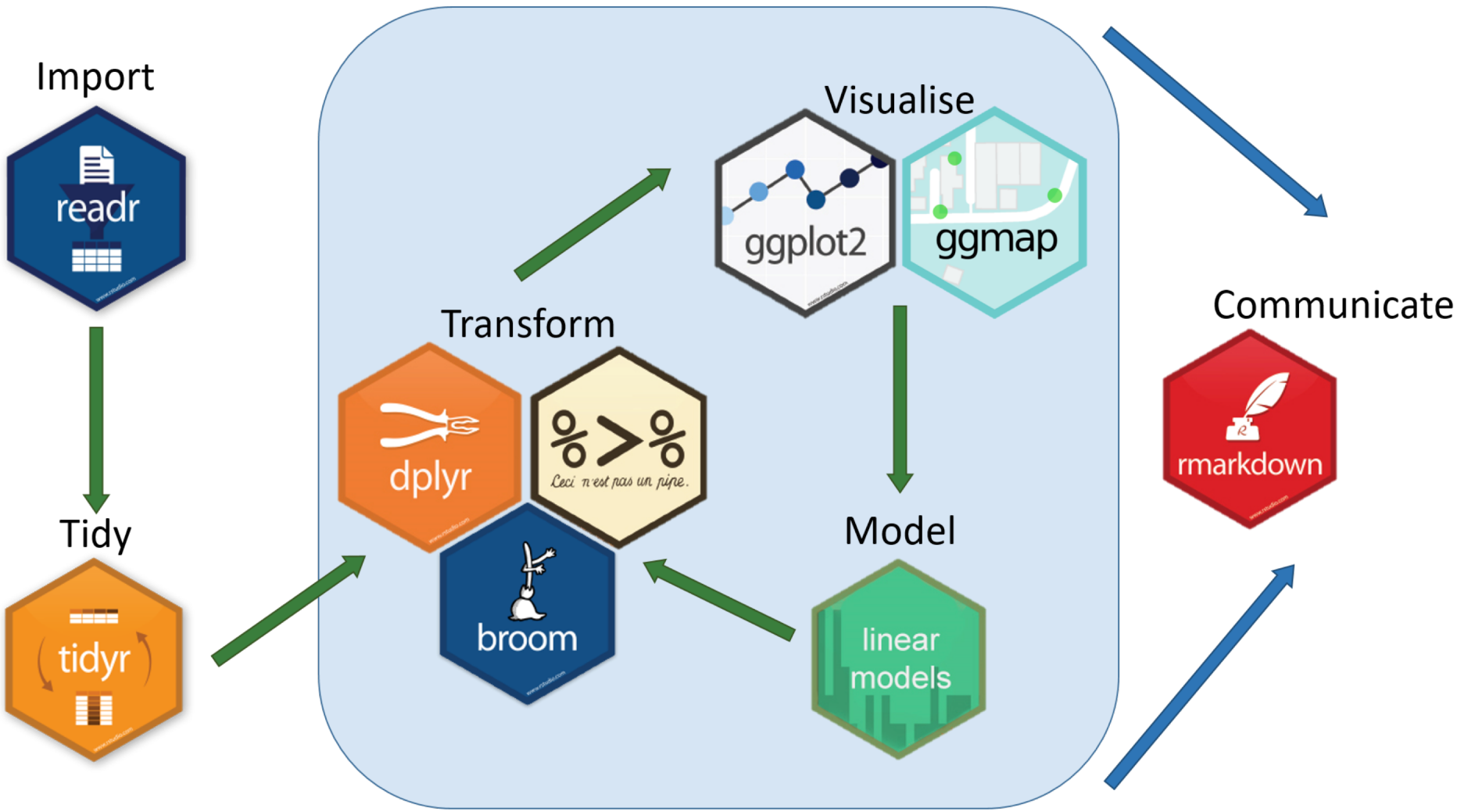
What we will NOT cover

- Statistics and hypothesis confirmation
- Modeling and simulation
- Big data
- Any other programming languages than R
- Non-rectangular data (e.g. images, sounds, trees, text) or domain-specific applications
- Base-R plotting and workflows (we will focus on the tidyverse)

The tidyverse

- An opinionated collection of R packages designed for data science
- All packages share an underlying design philosophy, grammar, and data structures


Import
Tidy
Transform
Visualise
Model
Communicate
Understand


Import
readr
Tidy
tidyr
Transform
dplyr
%>%
Leci n'est pas un pipe.
broom
Visualise
ggplot2
ggmap
Model
linear models
Communicate
rmarkdown

The tidyverse

- An opinionated collection of R packages designed for data science
- All packages share an underlying design philosophy, grammar, and data structures
- More streamlined and intuitive syntax and workflow than base R packages for most applications*

*Some would dispute that and swear by base R. We are not claiming that the tidyverse is superior to base R in all respects, only that it provides a set of very powerful tools

What we will NOT cover

- Statistics and hypothesis confirmation
- Modeling and simulation
- Big data
- Any other programming languages than R
- Non-rectangular data (e.g. images, sounds, trees, text) or domain-specific applications
- Base-R plotting and workflows (we will focus on the tidyverse)

What we WILL cover

Coding with best practices (RStudio/tidyverse)

Collaborative book-keeping (Git/GitHub)

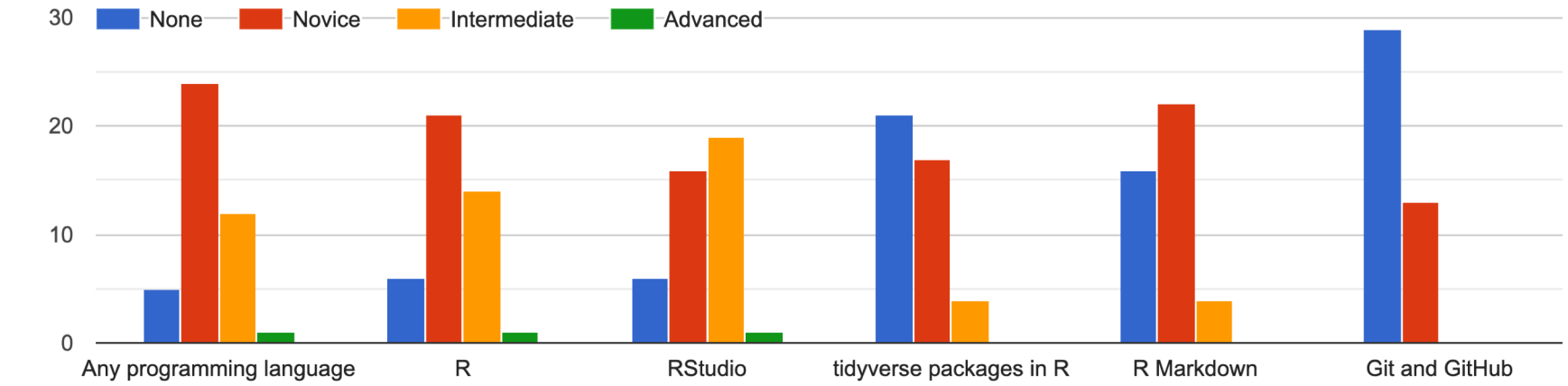
Reporting and communicating (RMarkdown/GitHub)

What we WILL cover

Coding with best practices (RStudio/tidyverse)

Collaborative book-keeping (Git/GitHub)

Reporting and communicating (RMarkdown/GitHub)



Example RMarkdowns

Course schedule

- https://github.com/nt246/NTRES6940-data-science#tentative-schedule-subject-to-adjustment

Course format

- Two weekly lectures – Mondays and Wednesday 2.55-4.10pm

- Optional “hacky hour” – Fridays 12:20-2:15pm
  - Come work on assignments or bring your own dataset for some end-of-week social coding

- Practice, practice, practice!

Lecture notes and assigned readings

- See course website https://nt246.github.io/NTRES6940-data-science/index.html
- Please complete the required readings before each class
- Optional readings are listed to help you dive further into the material

Live lectures

- During lecture please keep yourself muted unless you would like to ask a question
- Please do ask questions! You may ask questions by using the raised hand function in Zoom or through the Zoom chat or course Slack workspace
- This is a safe space and no question is dumb or pointless!
- If you're comfortable, please keep your video on (or post a photo if you will not have video on)
- Lectures will mostly be live coding, so type along with me!

Breakout Rooms

Starting Wednesday, we will send a survey via Slack prior to each class so you can indicate your breakout room preference for that day (due by noon). This will allow us to more efficiently form breakout groups and give you the opportunity to let us know if your preference changes or if you will be absent for a given live lecture.

blue	🔵 I would prefer to be in a breakout room where people have their **cameras on and actively work on exercises together**
red	🔴 I would prefer to be a breakout room where everyone works on the exercise **alone, but can ask others if they need help**
black	I would prefer to just work on exercises on my own and will probably **not participate in group discussions in breakout rooms**

Check the #logistics channel on Slack and respond to the message with an emoji circle
Blue will be your default if you don't respond for any given week

Lecture recordings

- Will be available on Zoom

Assignments

- Weekly problem sets
  - Assigned each Wednesday, due the following Wednesday at 10pm
  - Will submit on GitHub – more instructions to follow

Evaluation

- To pass this course you must:
  - Attend all lectures unless otherwise arranged (direct message Maria on Slack beforehand if you need to miss class)
  - Participate actively in class
  - Submit at least 6 of the 7 problem sets with demonstrated effort to complete all questions
  - Give a speed presentation (~2 mins) at the end of the course on how you are implementing something we have learned in your own work

We know the world is crazy right now

- Talk to us if you need species arrangements

Course communication

- All course communication will be via Slack and GitHub
  (occasionally we may use Canvas announcement, but check Slack to stay up-to-date)

- Help answer each other's questions and post cool tips you come across
  - The more we engage, the more we learn. You learn by helping others

- No questions are stupid, so no one should feel bad about asking. But if you prefer to be anonymous, we have installed the Anonymity Bot in Slack. Just type /anon

Where to find things

- Course website: https://nt246.github.io/NTRES6940-data-science/index.html
- Canvas page: https://canvas.cornell.edu/courses/21578
- Slack: https://app.slack.com/client/T01A5DRGMFV

Ongoing feedback

- Still getting used to the online format… So there is room for adjustment along the way
  - Ongoing input through the Slack ‘feedback’ channel
  - Regular quick check-ins

- If anyone has special accessibility concerns, please reach out

Check list

Have you:

- ❑ Gotten the current versions of R and RStudio working?
- ❑ Followed the instructions for installing Git and making a GitHub account?
- ❑ Joined the workspace on Slack (check that you have all the channels)
- ❑ Added a photo to you GitHub and Slack accounts? (optional)

Introduction to R/RStudio

- RStudio IDE orientation
- Shortcuts and autocomplete
- Install and load packages
- Scripts
- Home directory and RStudio projects
- Change settings to not save workspace