NTRES 6100:

Collaborative and Reproducible Data Science in R

Why are we here?

What are we going to do?

Who are we?

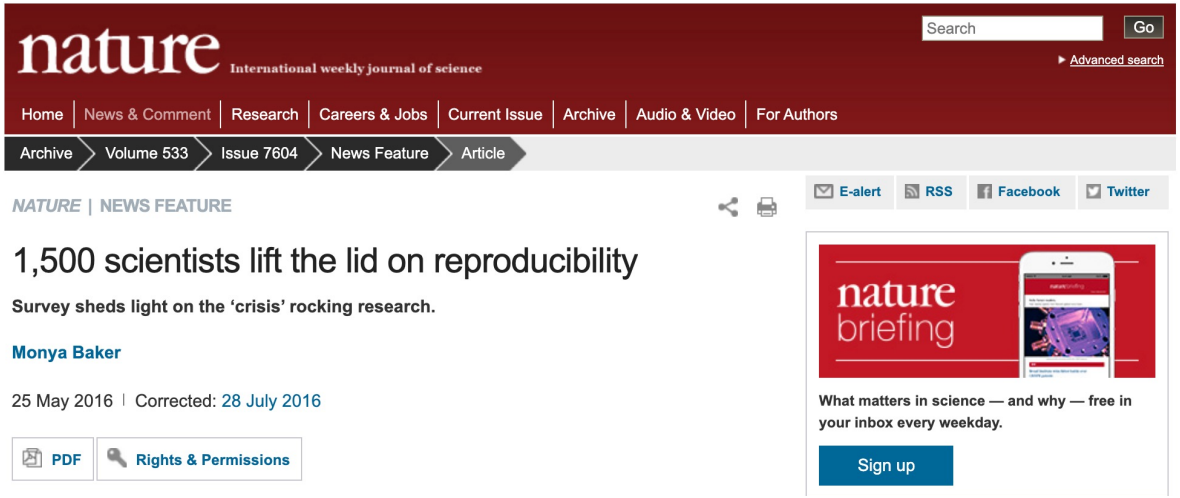
Why are we here?

What are we going to do?

Who are we?

And then we'll get situated with R/RStudio

Is science facing a reproducibility crisis?



"More than 70% of researchers have tried and failed to reproduce another scientist's experiments, and more than half have failed to reproduce their own experiments"

Nature's survey of 1,576 researchers

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

Discuss in breakout rooms

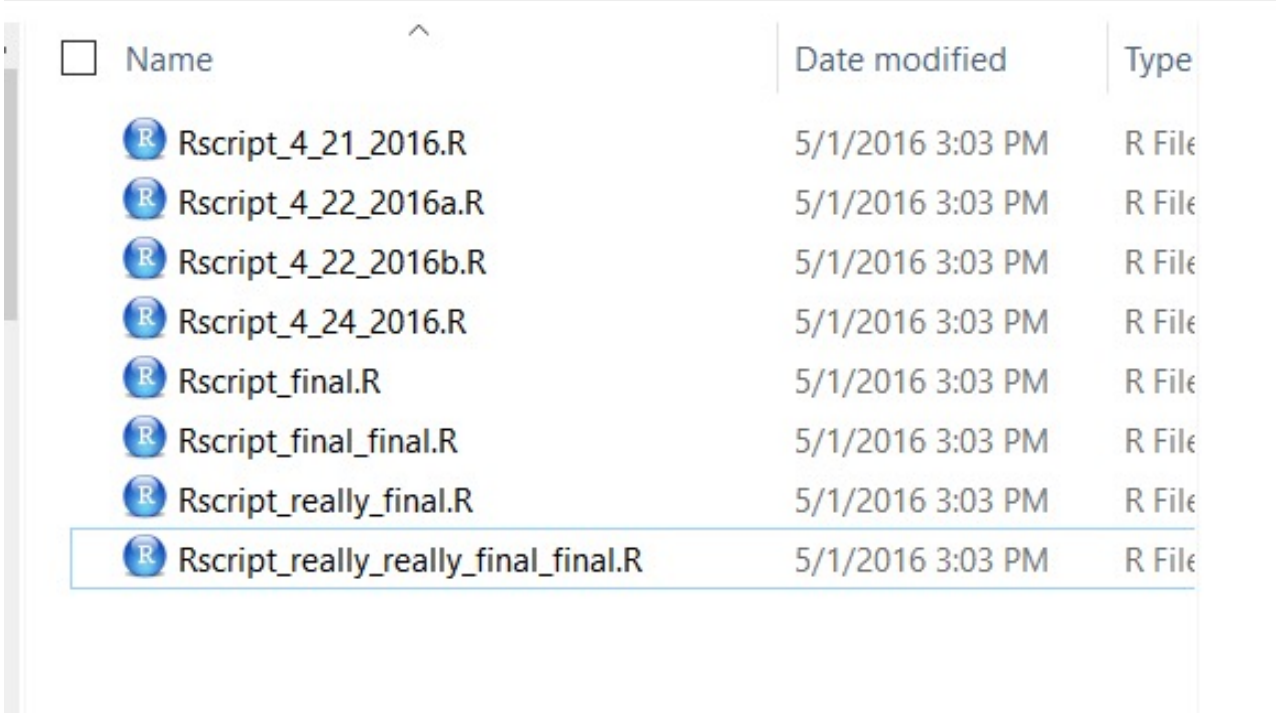
- Why may we sometimes fail to reproduce results in re-analysis of existing datasets?

- What can we do to ensure better reproducibility?

Why may we sometimes fail to reproduce results in re-analysis of existing datasets?

- Please add ideas here https://www.menti.com/f2x74wzfq5

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)

Nicolas Lou
(TA)

Let’s get to know each other

Meet your classmates in our shared Google Slides

What do we have in common?

Let’s get to know each other

Meet your classmates in our shared Google Slides

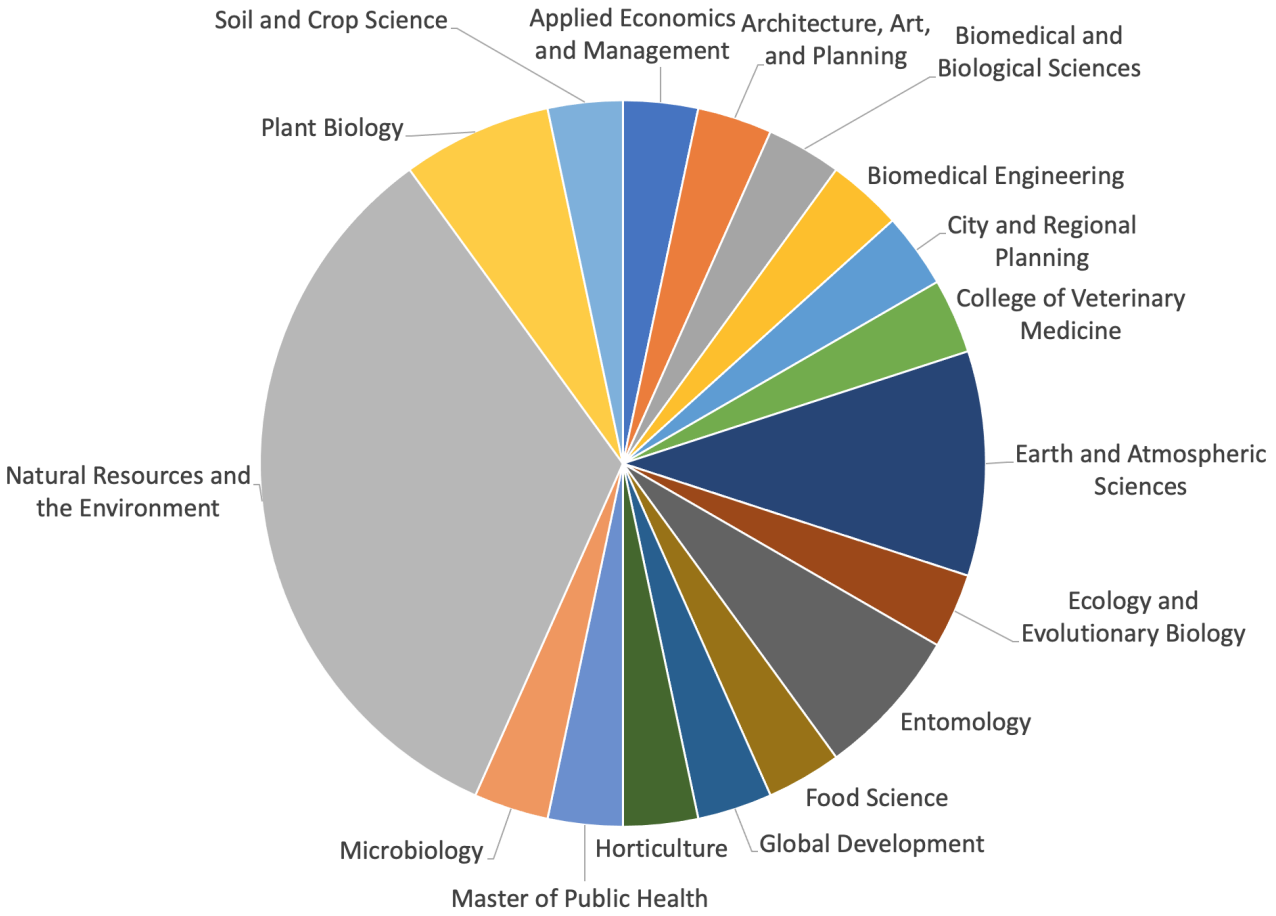
What do we have in common? https://www.menti.com/vkwnpxufs7

Thanks for completing the pre-course survey!

35 respondents

Departments

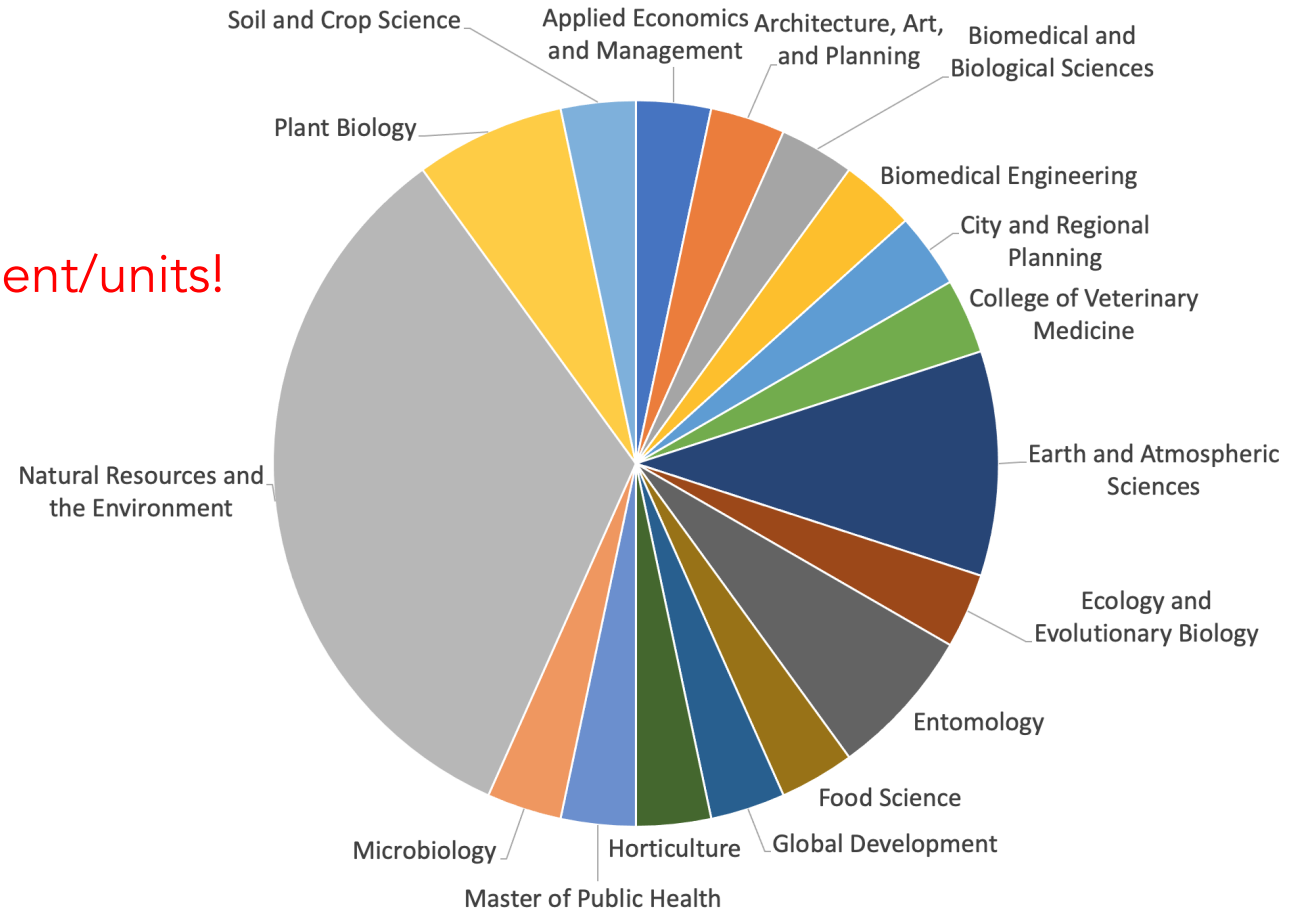
n = 31 respondents



Departments

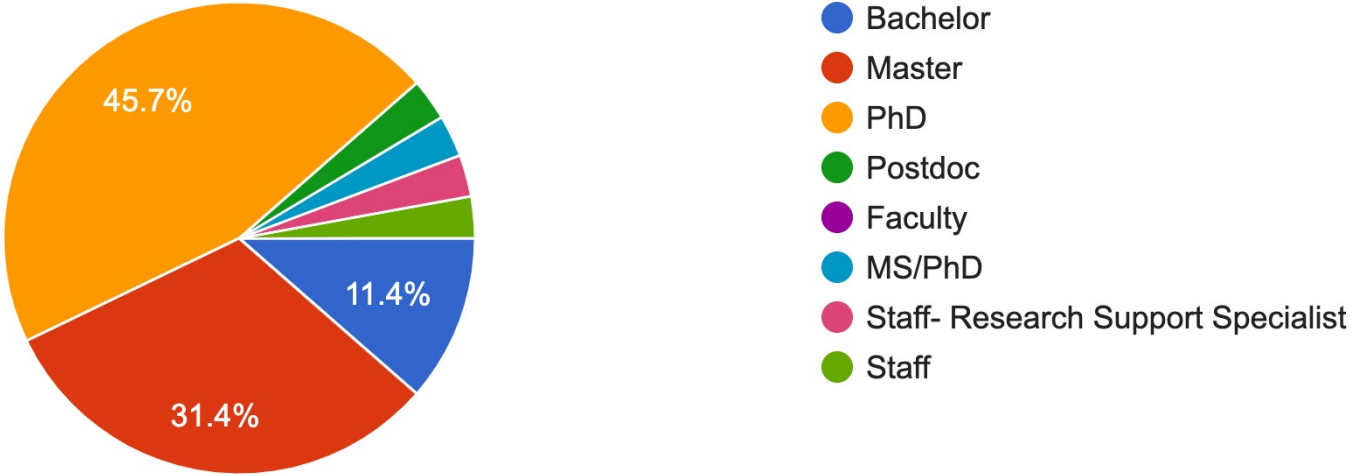
n = 31 respondents

17 different department/units!



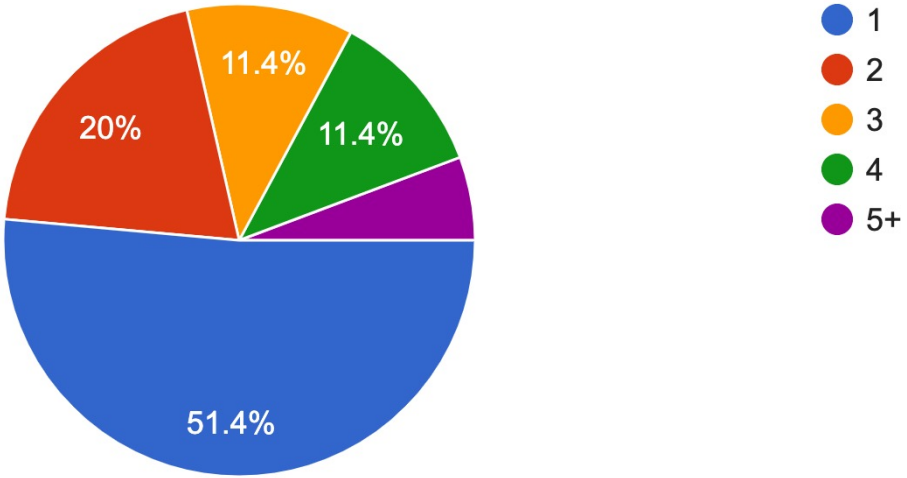
Degree program or position

35 responses

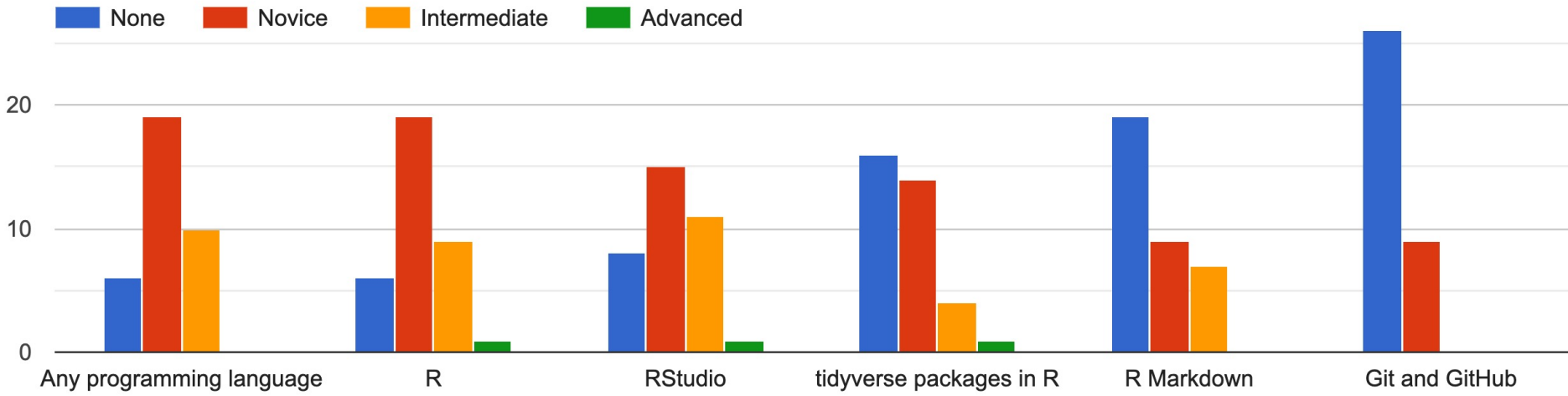


Year in your program/position

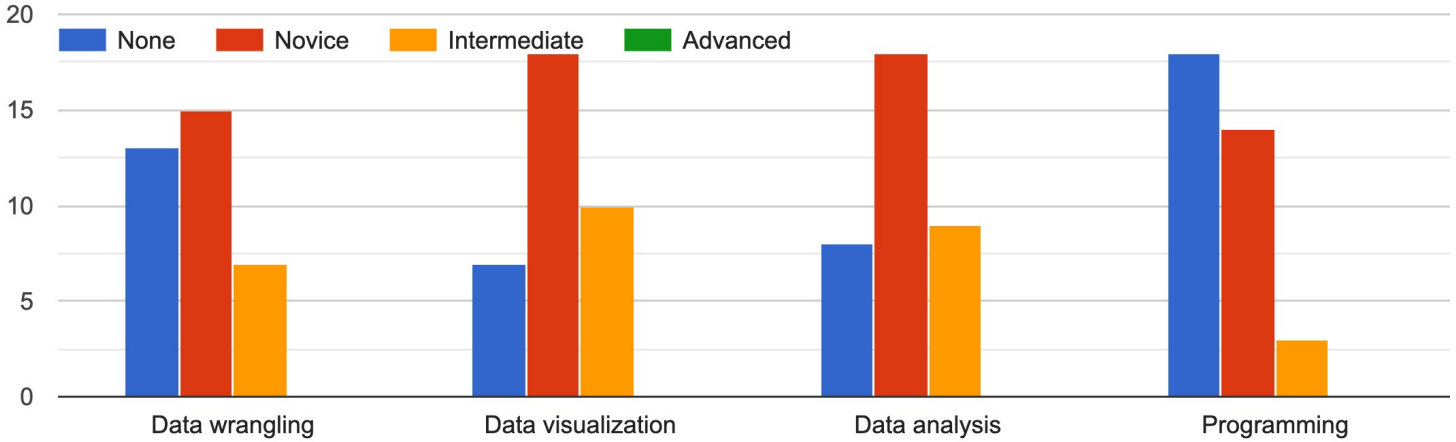
35 responses



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

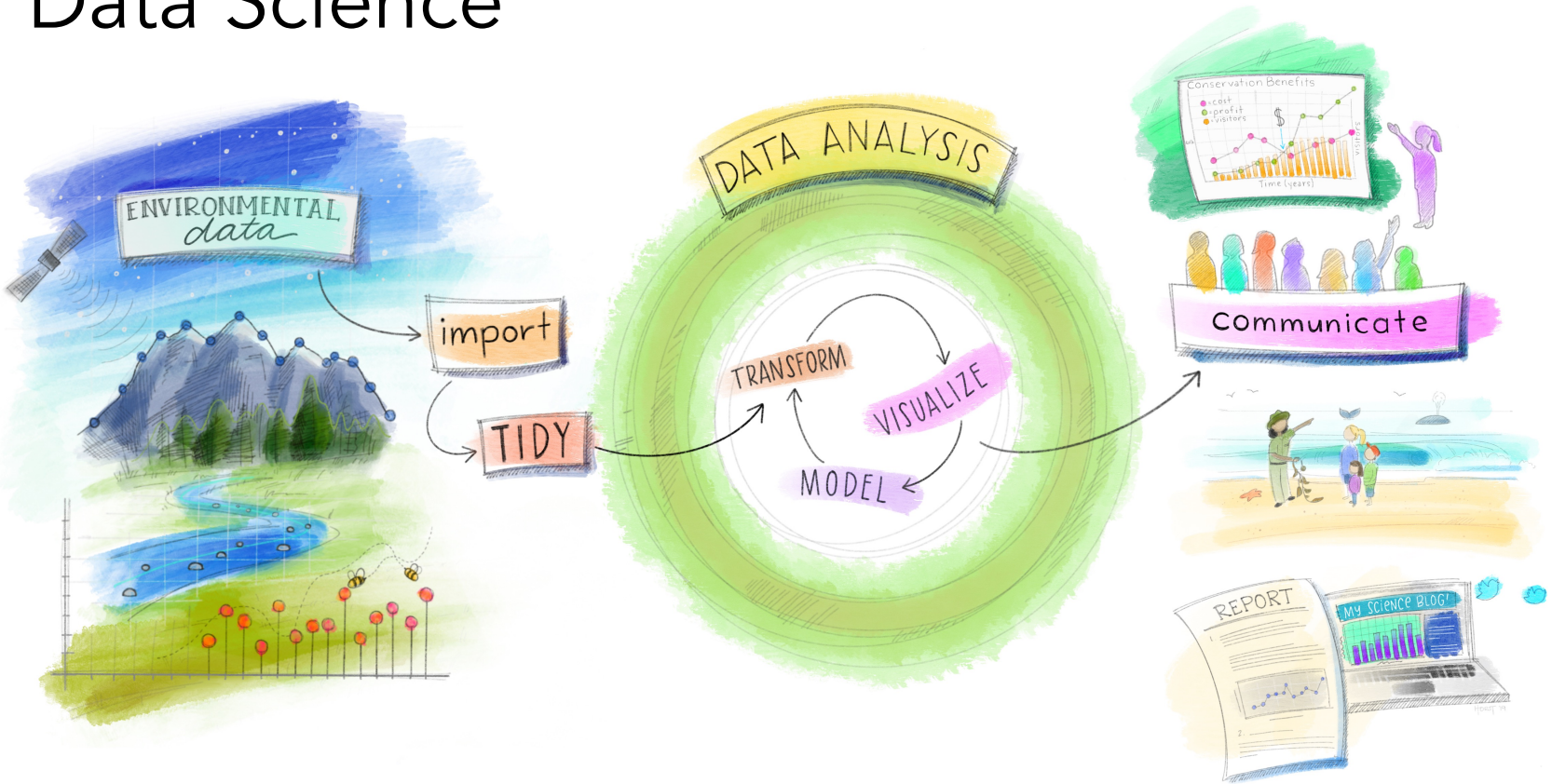
better science in less time

and with less pain

Data Science

Turning raw data into understanding, insight, and knowledge

Data Science



Artwork from @juliesquid for @openscapes (illustrated by @allison_horst)

Learning outcomes

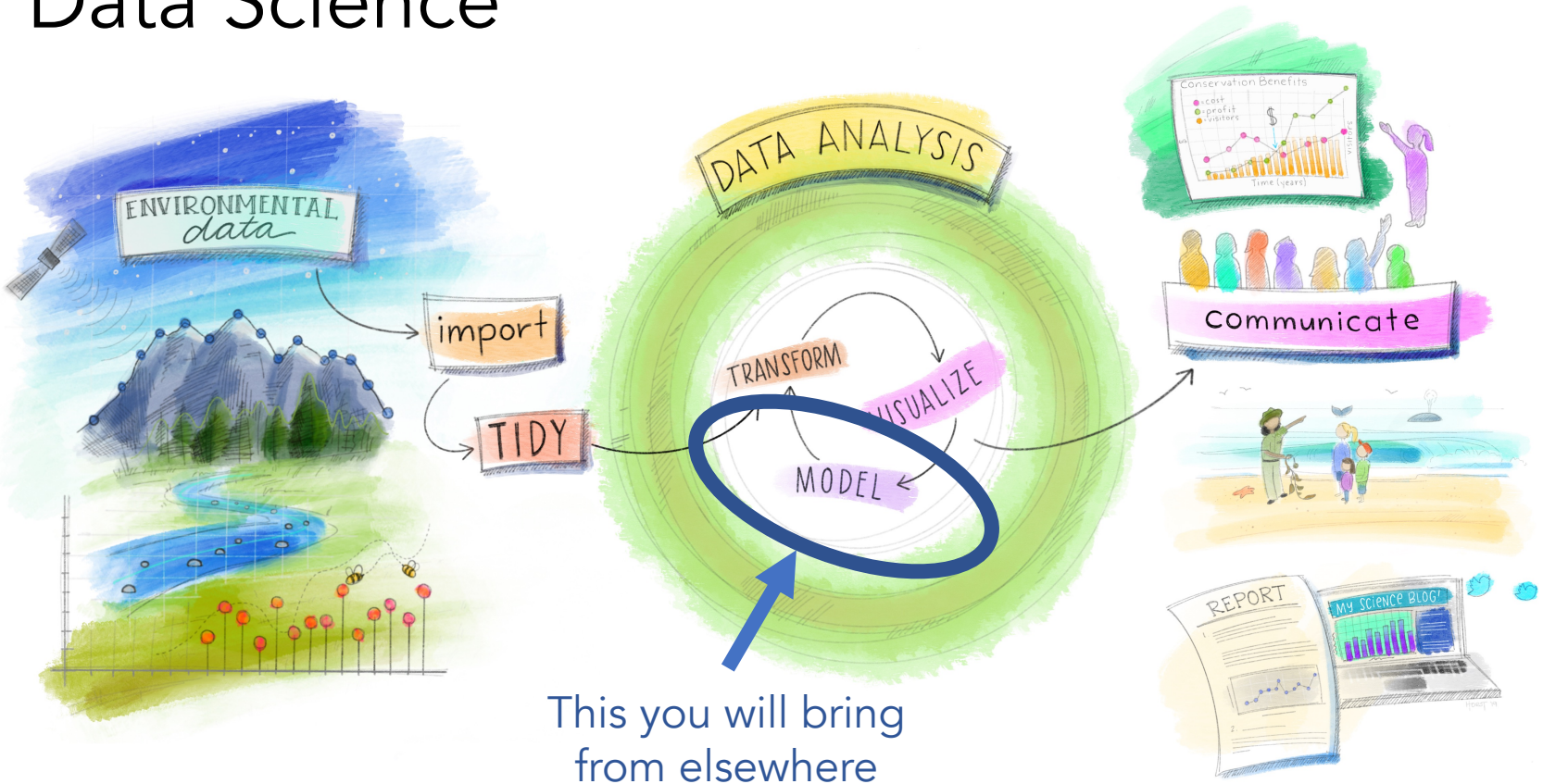
By the end of this course, students will be able to

- 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 Science



This you will bring from elsewhere

Artwork from @juliesquid for @openscapes (illustrated by @allison_horst)

For Big-Data Scientists, 'Janitor Work' Is Key Hurdle to Insights – NYTimes.com

TECHNOLOGY | For Big-Data Scientists, 'Janitor Work' Is Key Hurdle to Insights

SHARE

For Big-Data Scientists, 'Janitor Work' Is Key Hurdle to Insights

By STEVE LOHR AUG. 17, 2014

Monica Rogati, Jawbone's vice president for data science, with Brian Wilt, a senior data scientist.
Peter DaSilva for The New York Times

EMAIL

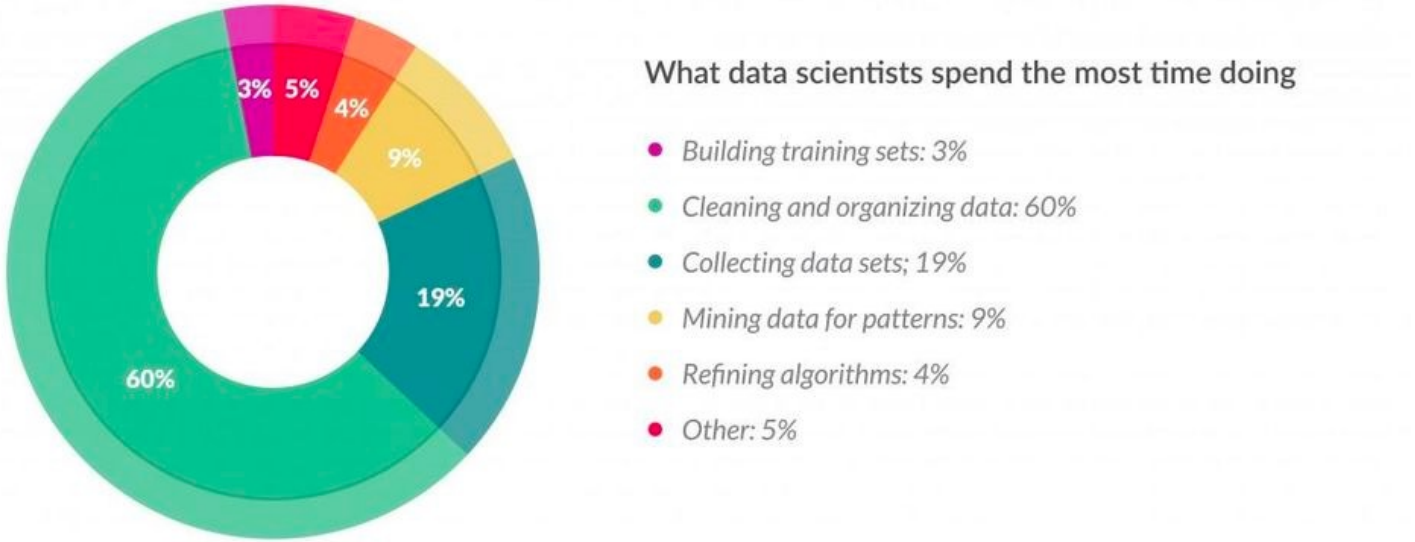
Technology revolutions come in measured, sometimes foot-dragging steps. The lab science and marketing enthusiasm tend to underestimate the

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
  - We will use tools can handle 100s Mb of data (up to ~1–2Gb)
  - Many big data problems are small data problems in disguise
    - Subset, subsample, summarize
    - 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

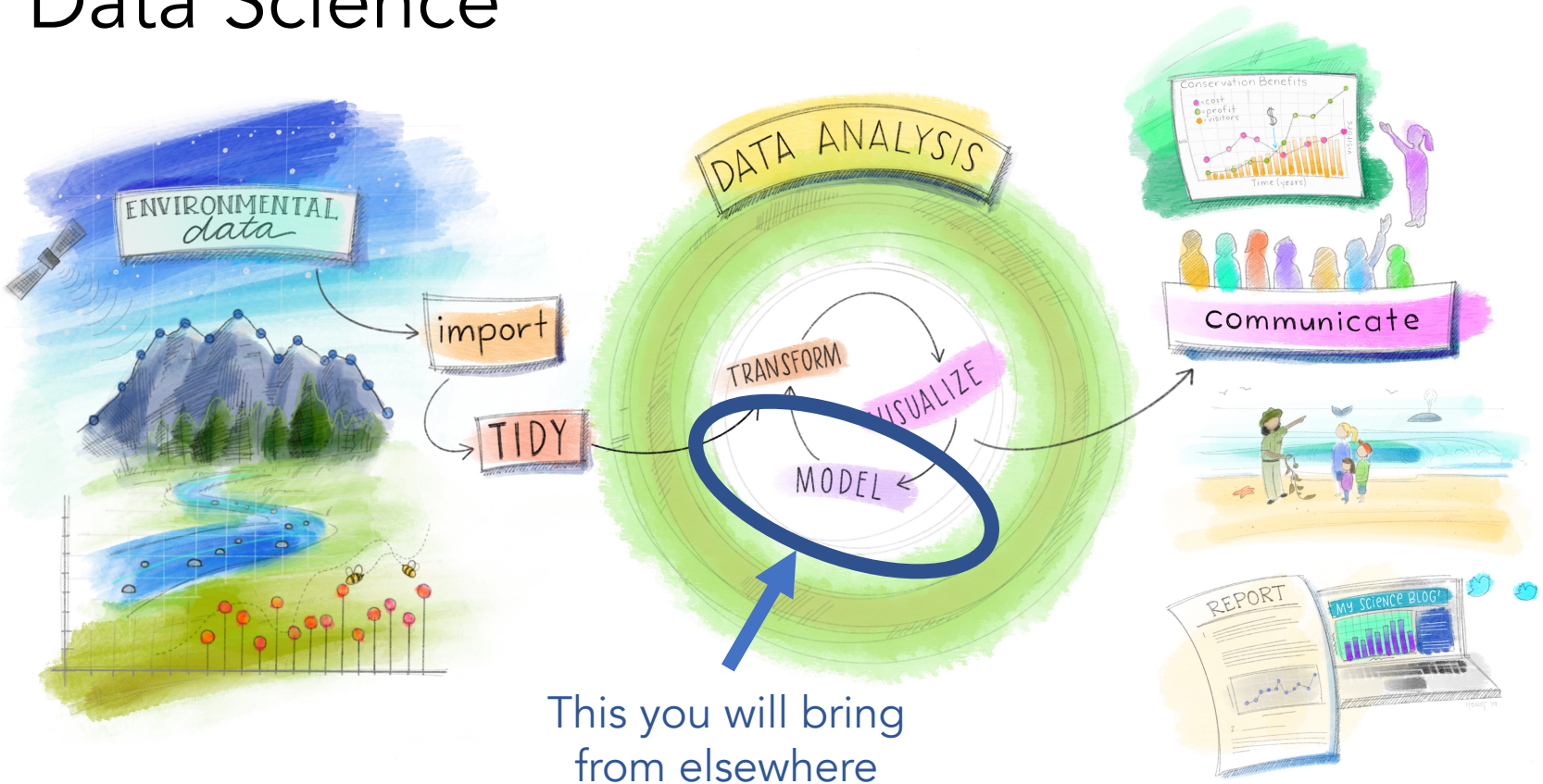
What we will NOT cover

- 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

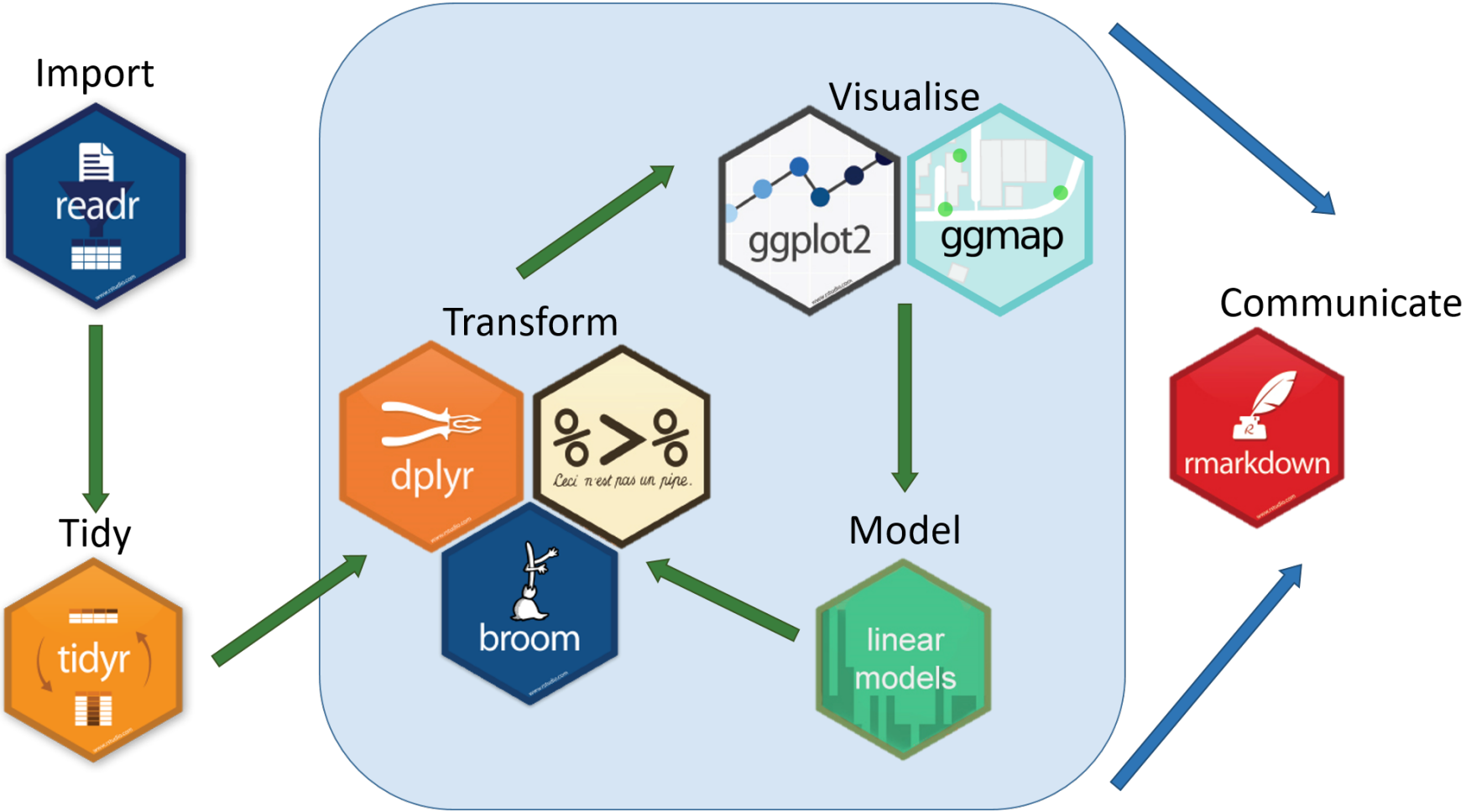
- An opinionated collection of R packages designed for data science
- All packages share an underlying design philosophy, grammar, and data structures

Data Science



This you will bring from elsewhere

Artwork from @juliesquid for @openscapes (illustrated by @allison_horst)


Import
readr
Tidy
tidyr
Transform
dplyr
%>%
Ceci 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 structure
- 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)

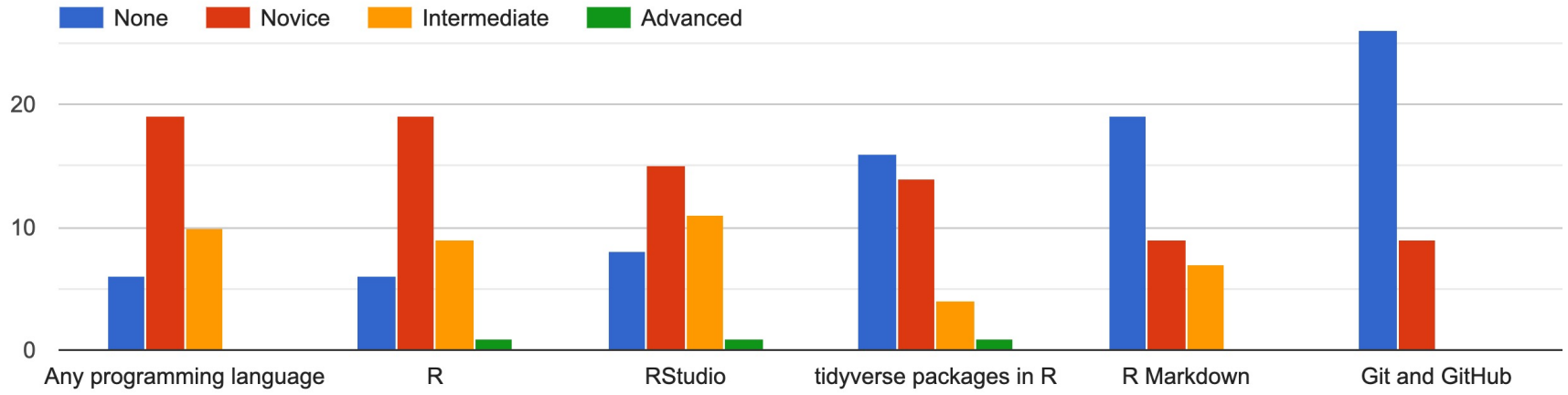
Reporting and communicating (RMarkdown/GitHub)

What we WILL cover

Coding with best practices (RStudio/tidyverse)

Collaborative book-keeping (Git/GitHub)

Reporting and communicating (RMarkdown/GitHub)



Example RMarkdowns

- https://github.com/therkildsen-lab/data-processing
- https://github.com/therkildsen-lab/greenland-cod
- https://github.com/therkildsen-lab/batch-effect

Course schedule

- https://nt246.github.io/NTRES-6100-data-science/syllabus.html

Course format

- Two weekly lectures – Tuesdays and Thursdays 9.40-10.55pm

- Optional labs – Thursdays OR Fridays 12:25-2:20pm (1 extra credit)
  - Exercises, reinforcement and expansion of lecture material, open-ended problem-solving

- Practice, practice, practice!

Lecture notes and assigned readings

- See course website https://nt246.github.io/NTRES-6100-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 Zoom lectures, 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 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!

Lecture recordings

- Will be available on Canvas

Assignments

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

Evaluation

- To pass this course you must:
  - Attend all lectures unless otherwise arranged (direct message Nicolas on Slack beforehand if you need to miss class)
  - Participate actively in class
  - Submit at least 8 of the 9 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 (and has been so for a while…)

- Talk to us if you need special 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

Tips for using Slack effectively

- Add your preferred name and a photo to your profile to make our workspace more personal
- Try to write your thoughts in a single direct message to minimize notifications
- Manage your notifications by turning on or scheduling Do Not Disturb
- Use channels and threads to keep the workspace organized
- Replace short follow-up messages with emoji reactions

More details here: https://slack.com/blog/collaboration/etiquette-tips-in-slack

Where to find things

- Course website: https://nt246.github.io/NTRES-6100-data-science/index.html
- Canvas page: https://canvas.cornell.edu/courses/39033
- Slack: https://join.slack.com/t/ntres6100-sp22/shared_invite/zt-11ztkw0t3-34p8BoXwoQiwOLeXqE_u2Q

Ongoing feedback

- We want you to get as much as possible out of this course and there is room for adjustment along the way
  - We always welcome input through the Slack ‘feedback’ channel
  - We will conduct regular quick check-ins
  - Feel free to use the “anonymity bot” on Slack to share your thoughts

- 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, but encouraged)

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

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