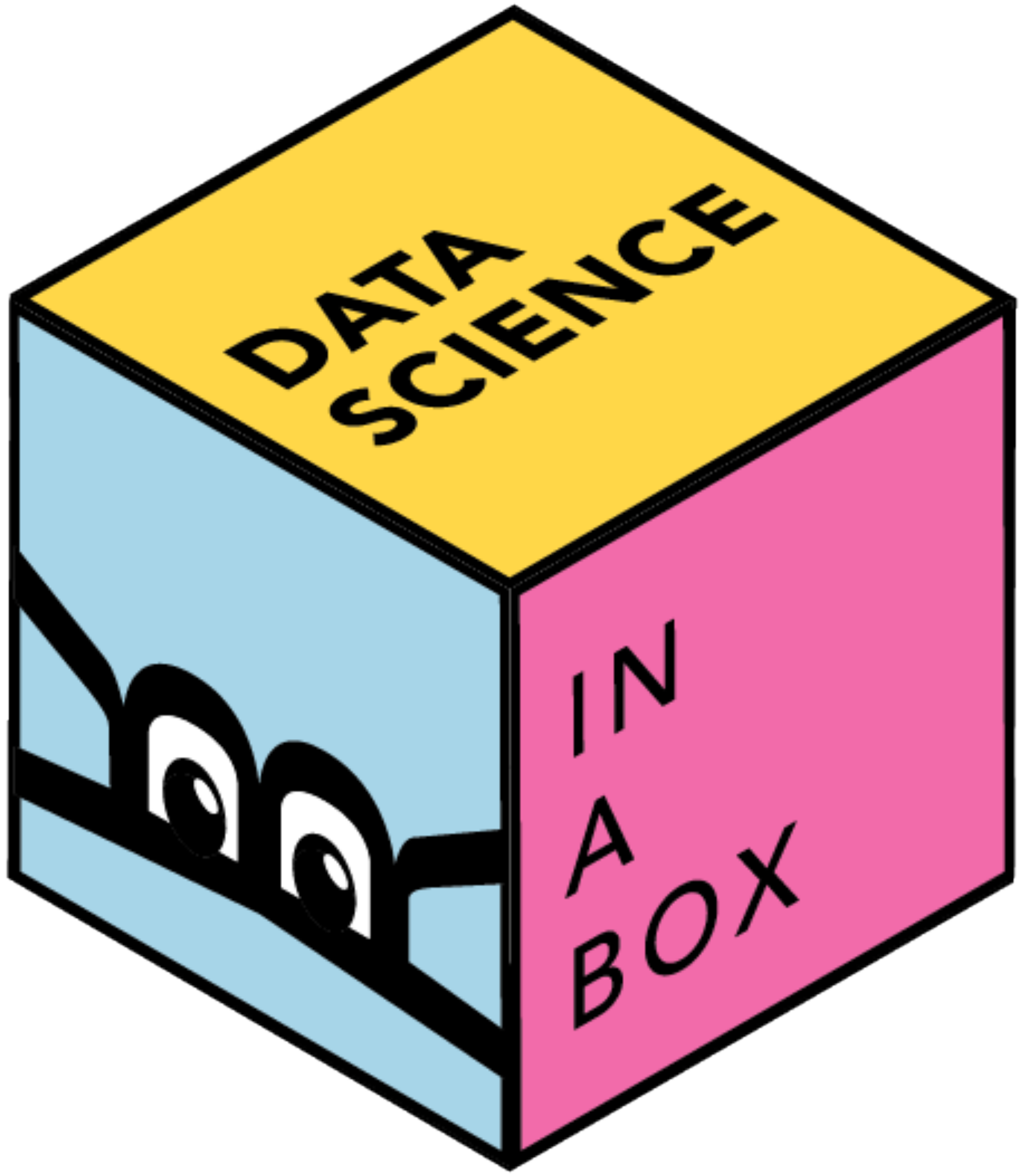


MINE ÇETINKAYA-RUNDEL

UNIVERSITY OF EDINBURGH + RSTUDIO

@minebocek

mine-cetinkaya-rundel

cetinkaya.mine@gmail.com

Three questions that keep me up at night…

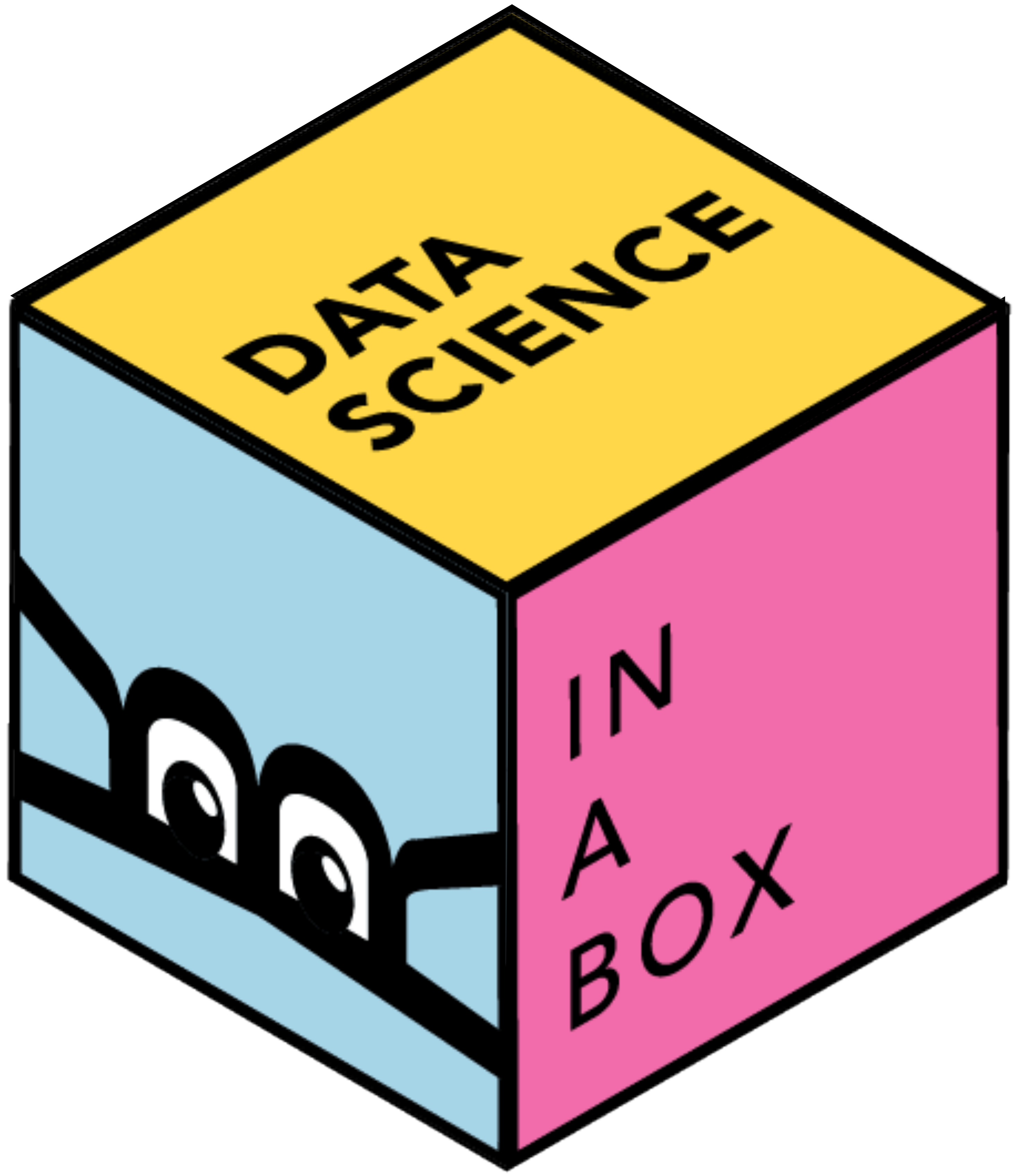
1 What should my students learn?
2 How will my students learn best?
3 What tools will enhance my students’ learning?

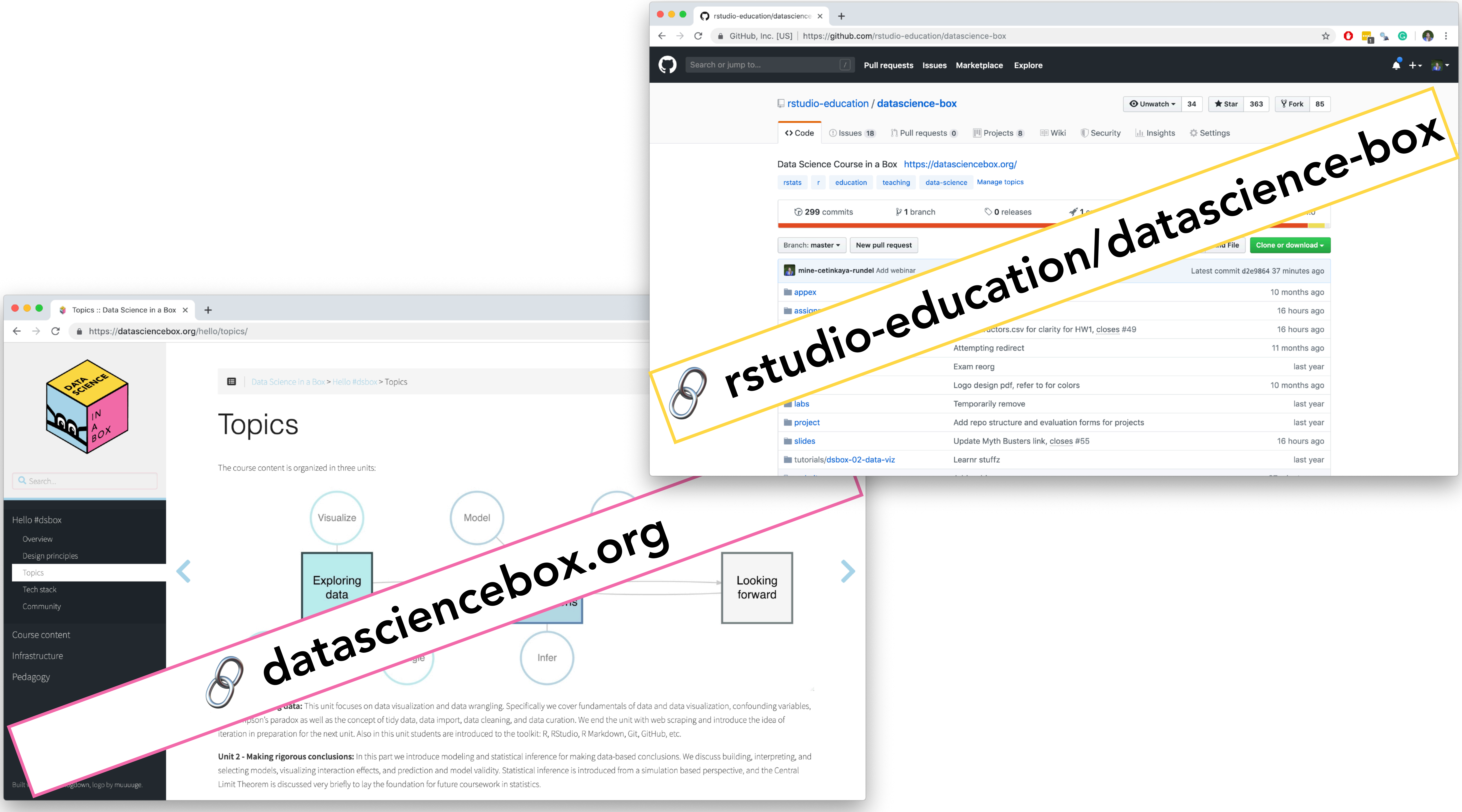
Three questions that keep me up at night…

Content 1 What should my students learn?

Pedagogy 2 How will my students learn best?

Infrastructure 3 What tools will enhance my students’ learning?


DATA SCIENCE
IN A BOX



rstudio-education/datascience-box

datasciencebox.org

rstudio-education/datascience

GitHub, Inc. [US] | https://github.com/rstudio-education/datascience-box

rstudio-education / datascience-box

Data Science Course in a Box https://datasciencebox.org/

mine-cetinkaya-rundel Add webinar Latest commit d2e9864 37 minutes ago

appex		10 months ago
assign...		16 hours ago
	...actors.csv for clarity for HW1, closes #49	16 hours ago
	Attempting redirect	11 months ago
	Exam reorg	last year
	Logo design pdf, refer to for colors	10 months ago
labs	Temporarily remove	last year
project	Add repo structure and evaluation forms for projects	last year
slides	Update Myth Busters link, closes #55	16 hours ago
tutorials/dsbox-02-data-viz	Learnr stuffz	last year

Topics :: Data Science in a Box

https://datasciencebox.org/hello/topics/

Data Science in a Box > Hello #dsbox > Topics

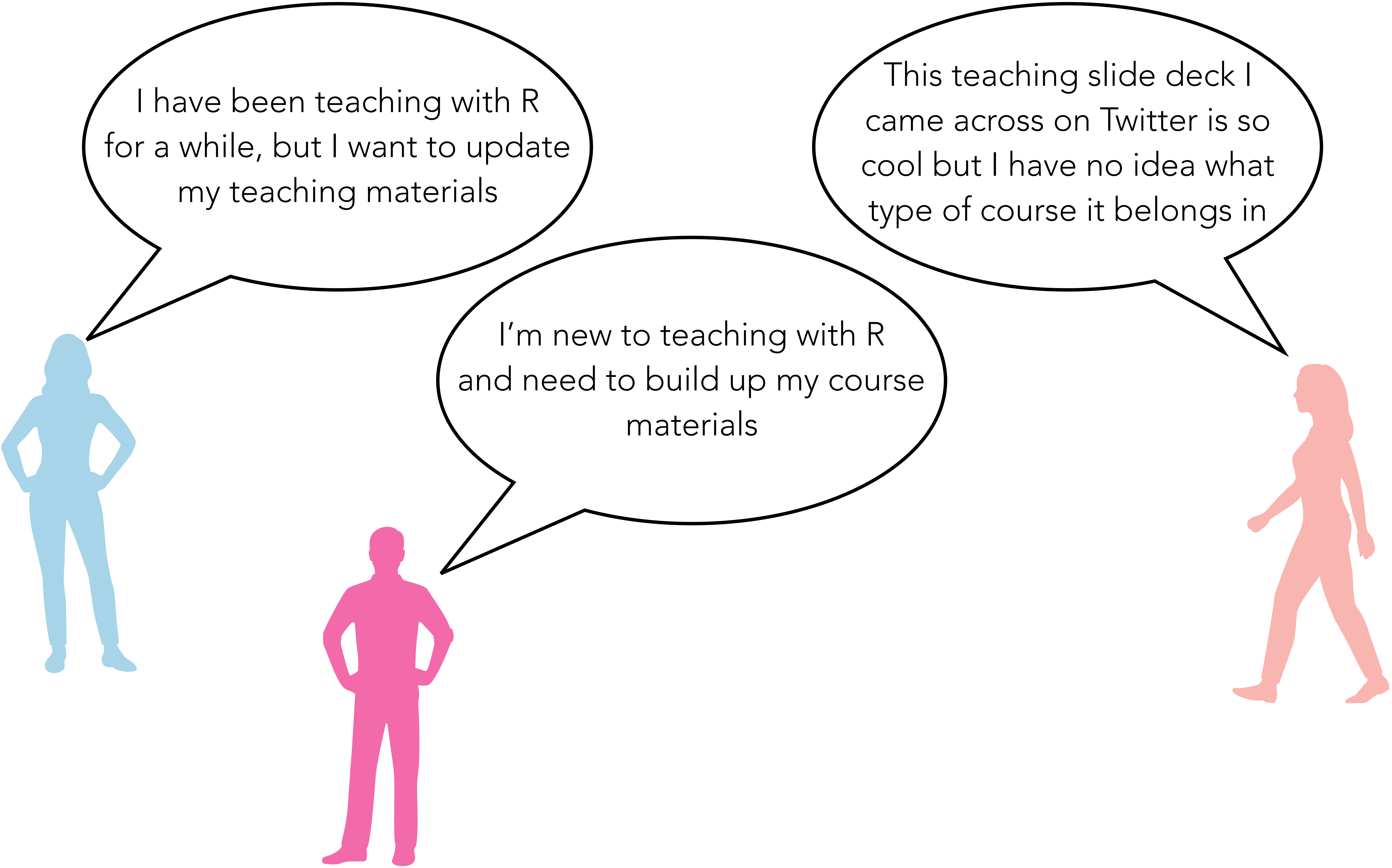
Topics

The course content is organized in three units:

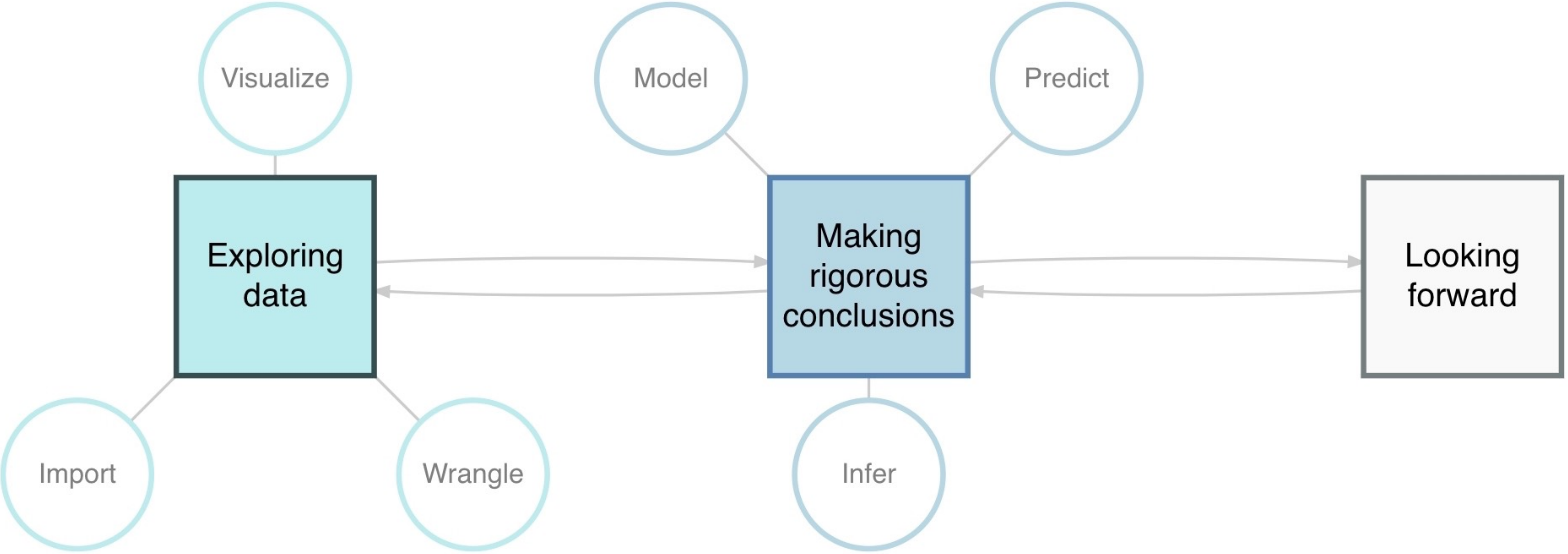
...g data: This unit focuses on data visualization and data wrangling. Specifically we cover fundamentals of data and data visualization, confounding variables, ...pson's paradox as well as the concept of tidy data, data import, data cleaning, and data curation. We end the unit with web scraping and introduce the idea of iteration in preparation for the next unit. Also in this unit students are introduced to the toolkit: R, RStudio, R Markdown, Git, GitHub, etc.

Unit 2 - Making rigorous conclusions: In this part we introduce modeling and statistical inference for making data-based conclusions. We discuss building, interpreting, and selecting models, visualizing interaction effects, and prediction and model validity. Statistical inference is introduced from a simulation based perspective, and the Central Limit Theorem is discussed very briefly to lay the foundation for future coursework in statistics.

AUDIENCE


I have been teaching with R for a while, but I want to update my teaching materials
I'm new to teaching with R and need to build up my course materials
This teaching slide deck I came across on Twitter is so cool but I have no idea what type of course it belongs in

TOPICS



Fundamentals of
data & data viz,
confounding variables,
Simpson's paradox
+
R / RStudio,
R Markdown, simple Git

Tidy data, data frames
vs. summary tables,
recoding & transforming,
web scraping & iteration
+
collaboration on GitHub

Building & selecting
models,
visualizing interactions,
prediction & validation,
inference via simulation

Data science ethics,
interactive viz &
reporting, text analysis,
Bayesian inference
+
communication &
dissemination

CONTENTS

26 lessons

10 application exercises

10 interactive tutorials

10 labs

10 homework assignments

2 take home exams

1 open-ended project

→

1 website

1 repo

1 package

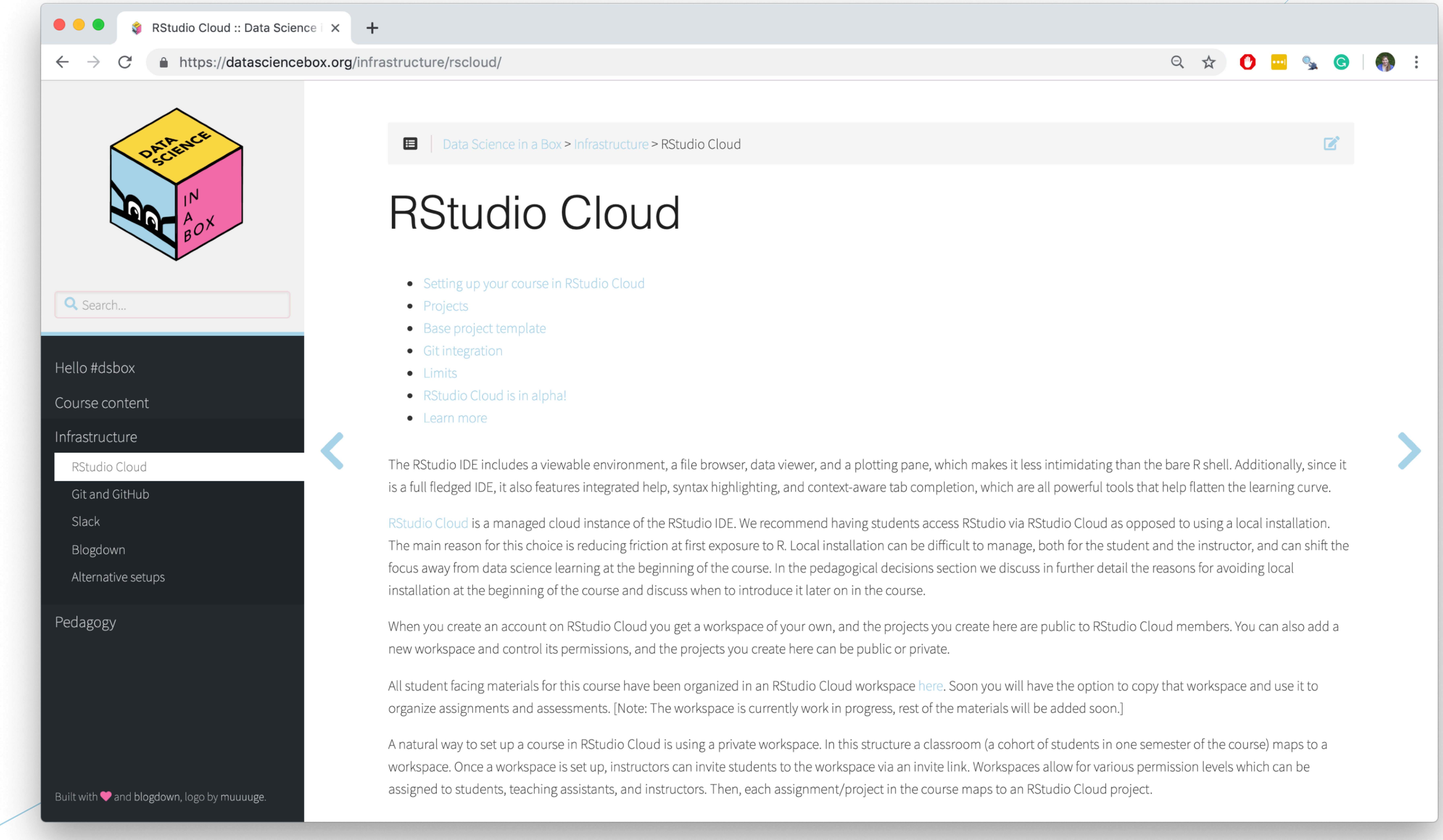
DESIGN PRINCIPLES

5 design principles

If you need to bake a cake, which kitchen would you rather get started in?

If you need to bake a cake, which kitchen would you rather get started in?

Cherish day one


RStudio Cloud :: Data Science
https://datasciencebox.org/infrastructure/rscloud/
DATA SCIENCE IN A BOX
Search...
Hello #dsbox
Course content
Infrastructure
RStudio Cloud
Git and GitHub
Slack
Blogdown
Alternative setups
Pedagogy
Built with ♥ and blogdown, logo by muuuuge.
Data Science in a Box > Infrastructure > RStudio Cloud

RStudio Cloud

- Setting up your course in RStudio Cloud
- Projects
- Base project template
- Git integration
- Limits
- RStudio Cloud is in alpha!
- Learn more

The RStudio IDE includes a viewable environment, a file browser, data viewer, and a plotting pane, which makes it less intimidating than the bare R shell. Additionally, since it is a full fledged IDE, it also features integrated help, syntax highlighting, and context-aware tab completion, which are all powerful tools that help flatten the learning curve.

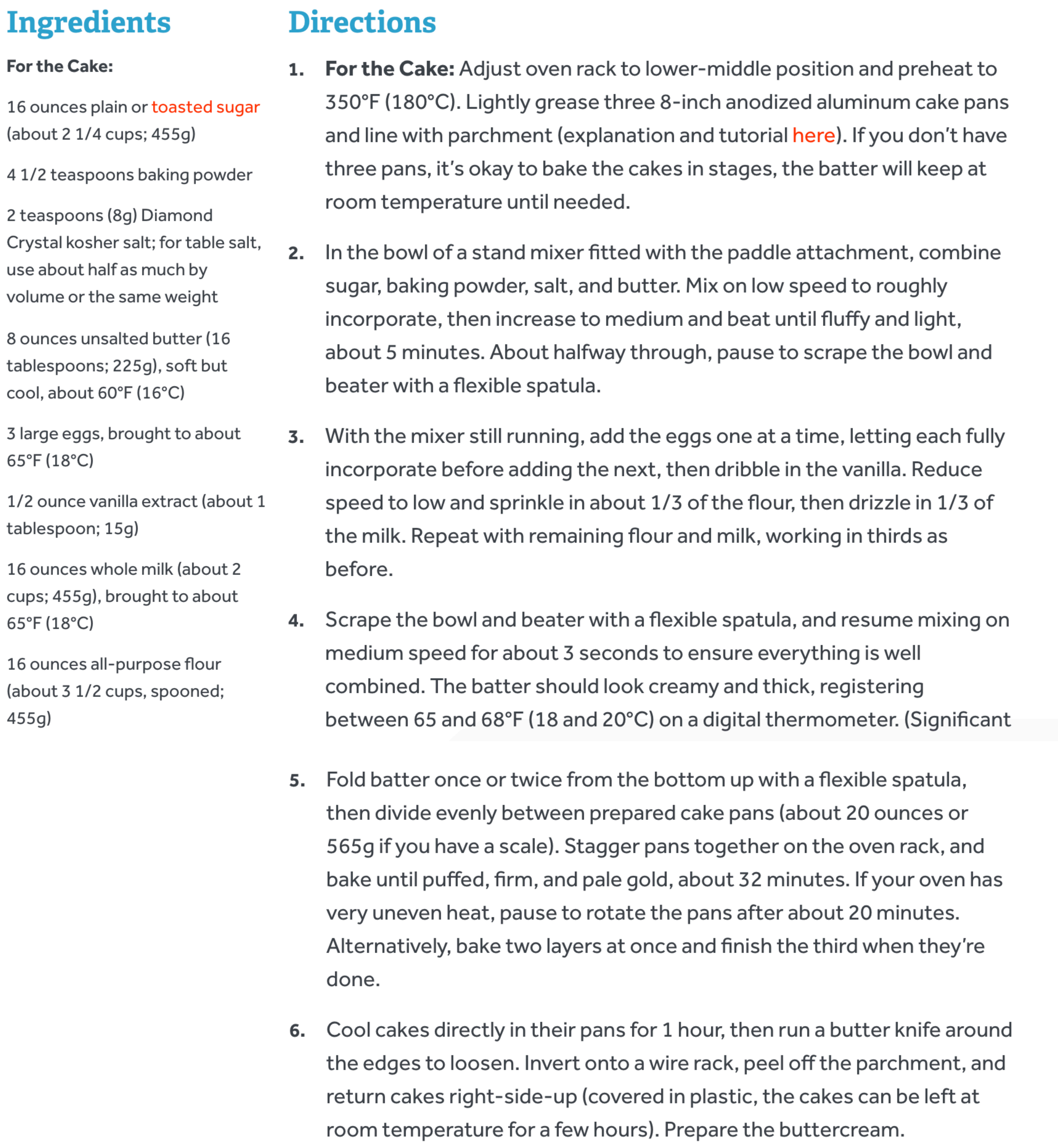
RStudio Cloud is a managed cloud instance of the RStudio IDE. We recommend having students access RStudio via RStudio Cloud as opposed to using a local installation. The main reason for this choice is reducing friction at first exposure to R. Local installation can be difficult to manage, both for the student and the instructor, and can shift the focus away from data science learning at the beginning of the course. In the pedagogical decisions section we discuss in further detail the reasons for avoiding local installation at the beginning of the course and discuss when to introduce it later on in the course.

When you create an account on RStudio Cloud you get a workspace of your own, and the projects you create here are public to RStudio Cloud members. You can also add a new workspace and control its permissions, and the projects you create here can be public or private.

All student facing materials for this course have been organized in an RStudio Cloud workspace here. Soon you will have the option to copy that workspace and use it to organize assignments and assessments. [Note: The workspace is currently work in progress, rest of the materials will be added soon.]

A natural way to set up a course in RStudio Cloud is using a private workspace. In this structure a classroom (a cohort of students in one semester of the course) maps to a workspace. Once a workspace is set up, instructors can invite students to the workspace via an invite link. Workspaces allow for various permission levels which can be assigned to students, teaching assistants, and instructors. Then, each assignment/project in the course maps to an RStudio Cloud project.

When baking a cake, which do you prefer: recipe only or recipe + pictures?



Ingredients

For the Cake:

16 ounces plain or toasted sugar (about 2 1/4 cups; 455g)

4 1/2 teaspoons baking powder

2 teaspoons (8g) Diamond Crystal kosher salt; for table salt, use about half as much by volume or the same weight

8 ounces unsalted butter (16 tablespoons; 225g), soft but cool, about 60°F (16°C)

3 large eggs, brought to about 65°F (18°C)

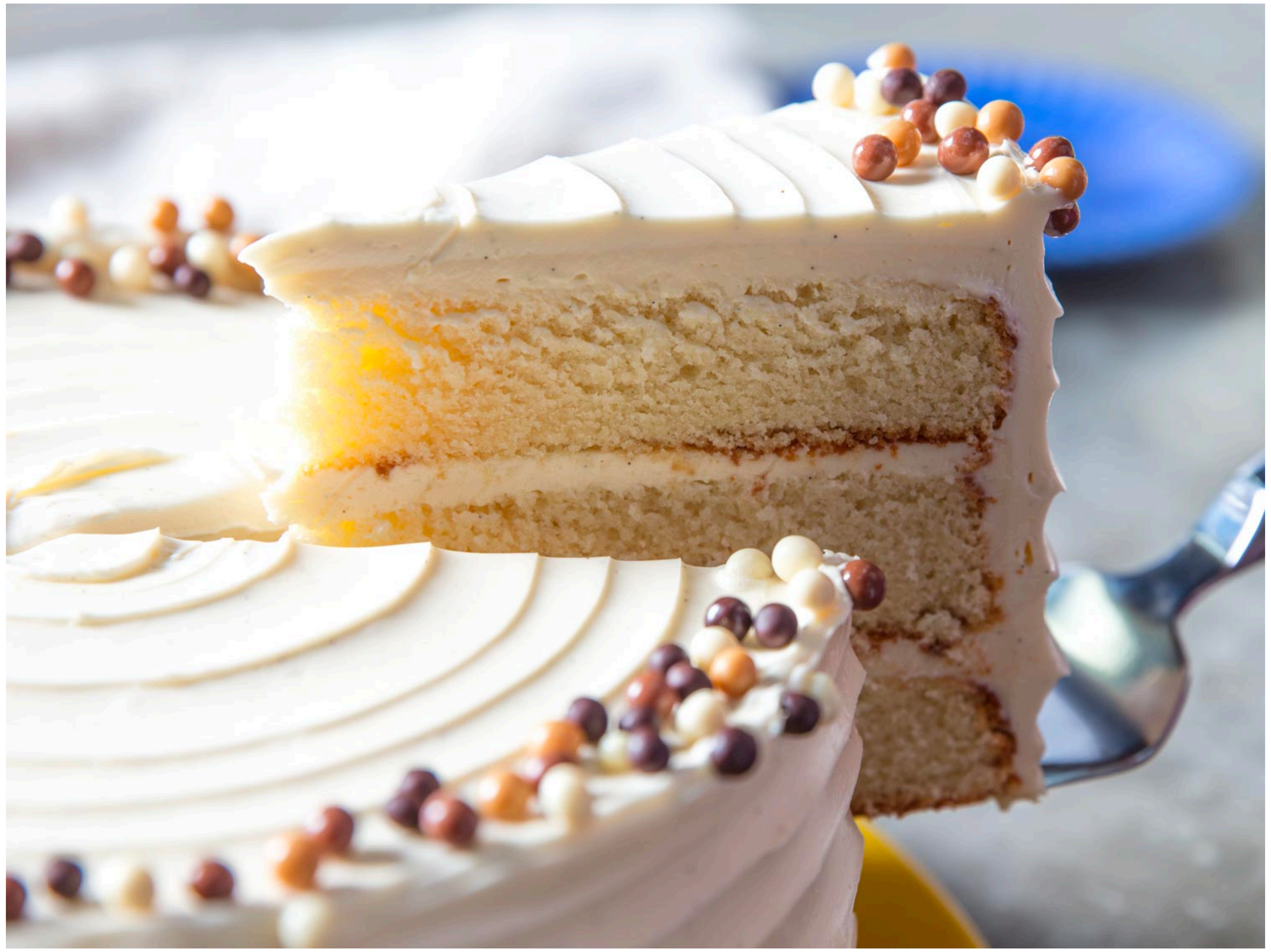
1/2 ounce vanilla extract (about 1 tablespoon; 15g)

16 ounces whole milk (about 2 cups; 455g), brought to about 65°F (18°C)

16 ounces all-purpose flour (about 3 1/2 cups, spooned; 455g)

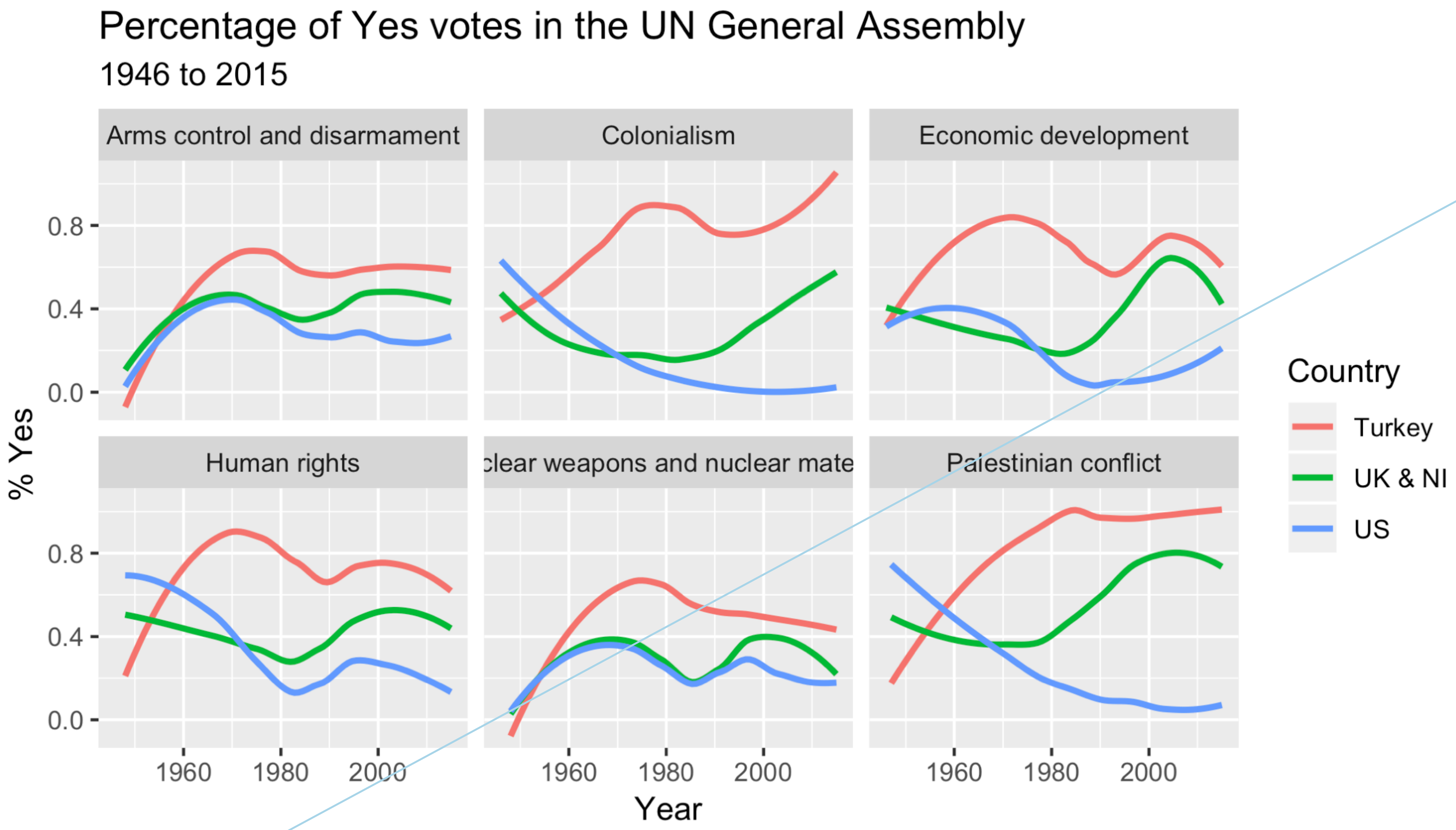
Directions

1. **For the Cake:** Adjust oven rack to lower-middle position and preheat to 350°F (180°C). Lightly grease three 8-inch anodized aluminum cake pans and line with parchment (explanation and tutorial here). If you don't have three pans, it's okay to bake the cakes in stages, the batter will keep at room temperature until needed.
2. In the bowl of a stand mixer fitted with the paddle attachment, combine sugar, baking powder, salt, and butter. Mix on low speed to roughly incorporate, then increase to medium and beat until fluffy and light, about 5 minutes. About halfway through, pause to scrape the bowl and beater with a flexible spatula.
3. With the mixer still running, add the eggs one at a time, letting each fully incorporate before adding the next, then dribble in the vanilla. Reduce speed to low and sprinkle in about 1/3 of the flour, then drizzle in 1/3 of the milk. Repeat with remaining flour and milk, working in thirds as before.
4. Scrape the bowl and beater with a flexible spatula, and resume mixing on medium speed for about 3 seconds to ensure everything is well combined. The batter should look creamy and thick, registering between 65 and 68°F (18 and 20°C) on a digital thermometer. (Significant
5. Fold batter once or twice from the bottom up with a flexible spatula, then divide evenly between prepared cake pans (about 20 ounces or 565g if you have a scale). Stagger pans together on the oven rack, and bake until puffed, firm, and pale gold, about 32 minutes. If your oven has very uneven heat, pause to rotate the pans after about 20 minutes. Alternatively, bake two layers at once and finish the third when they're done.
6. Cool cakes directly in their pans for 1 hour, then run a butter knife around the edges to loosen. Invert onto a wire rack, peel off the parchment, and return cakes right-side-up (covered in plastic, the cakes can be left at room temperature for a few hours). Prepare the buttercream.



Start with cake

- Open today's demo project
- Knit the document and discuss the results with your neighbor



- Then, change Turkey to a different country, and plot again

Start with cake

With great examples, comes a great amount of code…
but let's focus on the task at hand…

- Open today's demo project
- Knit the document and discuss the results with your neighbor
- Then, change Turkey to a different country, and plot again

Start with cake

```r
un_votes %>%
  filter(country %in% c("UK & NI", “US”, "Turkey")) %>%
  inner_join(un_roll_calls, by = "rcid") %>%
  inner_join(un_roll_call_issues, by = "rcid") %>%
  group_by(country, year = year(date), issue) %>%
  summarize(
    votes = n(),
    percent_yes = mean(vote == "yes")
  ) %>%
  filter(votes > 5) %>%  # only use records where there are more than 5 votes
  ggplot(mapping = aes(x = year, y = percent_yes, color = country)) +
  geom_smooth(method = "loess", se = FALSE) +
  facet_wrap(~ issue) +
  labs(
    title = "Percentage of Yes votes in the UN General Assembly",
    subtitle = "1946 to 2015",
    y = "% Yes",
    x = "Year",
    color = "Country"
  )
```

Start with cake

```
un_votes %>%
  filter(country %in% c("UK & NI", “US”, "Turkey")) %>%
  inner_join(un_roll_calls, by = "rcid") %>%
  inner_join(un_roll_call_issues, by = "rcid") %>%
  group_by(country, year = year(date), issue) %>%
  summarize(
    votes = n(),
    percent_yes = mean(vote == "yes")
  ) %>%
  filter(votes > 5) %>%  # only use records where there are more than 5 votes
  ggplot(mapping = aes(x = year, y = percent_yes, color = country)) +
  geom_smooth(method = "loess", se = FALSE) +
  facet_wrap(~ issue) +
  labs(
    title = "Percentage of Yes votes in the UN General Assembly",
    subtitle = "1946 to 2015",
    y = "% Yes",
    x = "Year",
    color = "Country"
  )
```

🍰 Start with cake

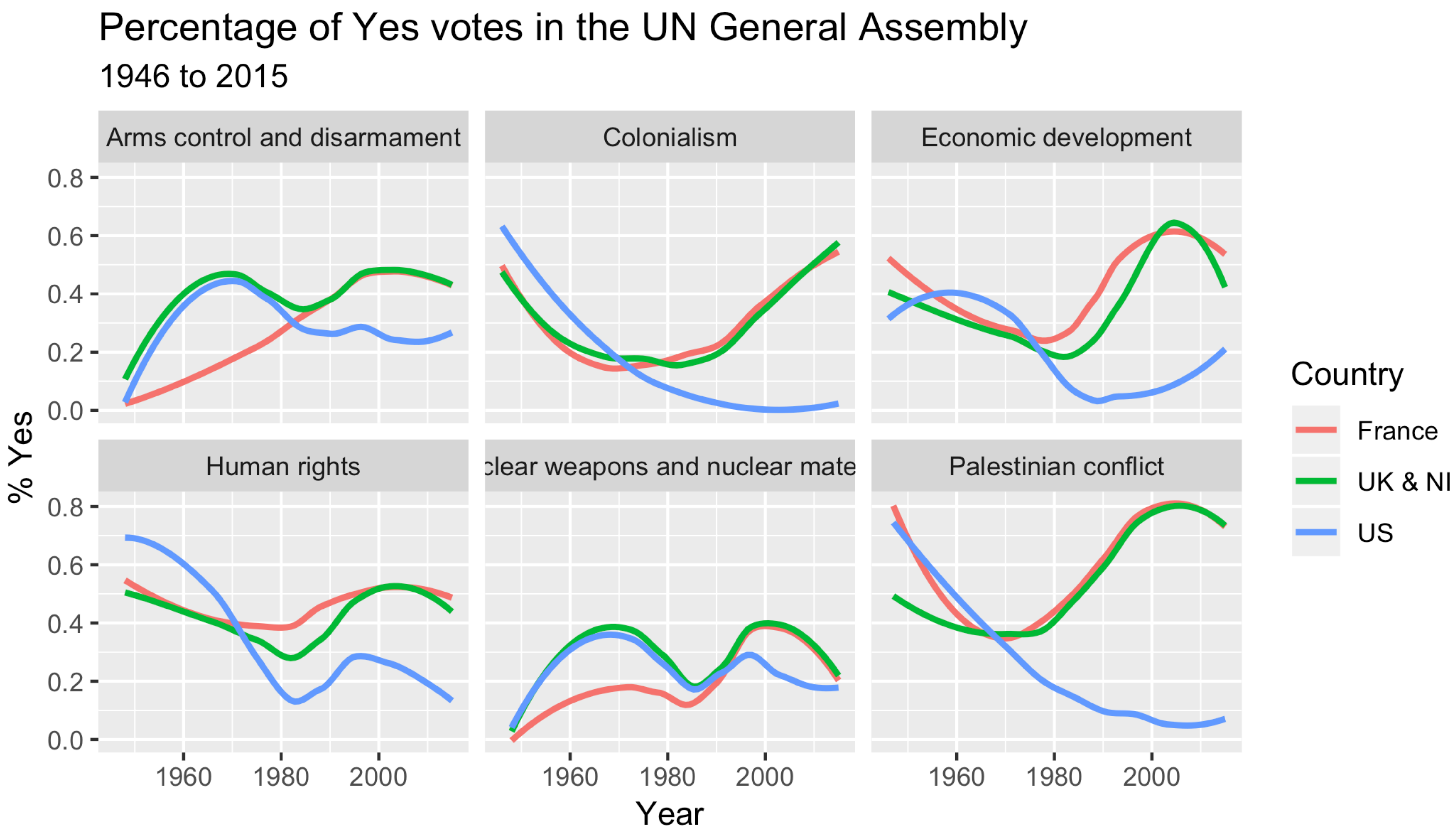
```
un_votes %>%
  filter(country %in% c("UK & NI", “US”, "Turkey")) %>%
  inner_join(un_roll_calls, by = "rcid") %>%
  inner_join(un_roll_call_issues, by = "rcid") %>%
  group_by(country, year = year(date), issue) %>%
  summarize(
    votes = n(),
    percent_yes = mean(vote == "yes")
  ) %>%
  filter(votes > 5) %>%  # only use records where there are more than 5 votes
  ggplot(mapping = aes(x = year, y = percent_yes, color = country)) +
  geom_smooth(method = "loess", se = FALSE) +
  facet_wrap(~ issue) +
  labs(
    title = "Percentage of Yes votes in the UN General Assembly",
    subtitle = "1946 to 2015",
    y = "% Yes",
    x = "Year",
    color = "Country"
  )
```

DESIGN PRINCIPLES

🍰 Start with cake

```
un_votes %>%
  filter(country %in% c("UK & NI", “US”, “France")) %>%
  inner_join(un_roll_calls, by = "rcid") %>%
  inner_join(un_roll_call_issues, by = "rcid") %>%
  group_by(country, year = year(date), issue) %>%
  summarize(
    votes = n(),
    percent_yes = mean(vote == "yes")
  ) %>%
  filter(votes > 5) %>%  # only use records where there are more than 5 votes
  ggplot(mapping = aes(x = year, y = percent_yes, color = country)) +
  geom_smooth(method = "loess", se = FALSE) +
  facet_wrap(~ issue) +
  labs(
    title = "Percentage of Yes votes in the UN General Assembly",
    subtitle = "1946 to 2015",
    y = "% Yes",
    x = "Year",
    color = "Country"
  )
```

Start with cake



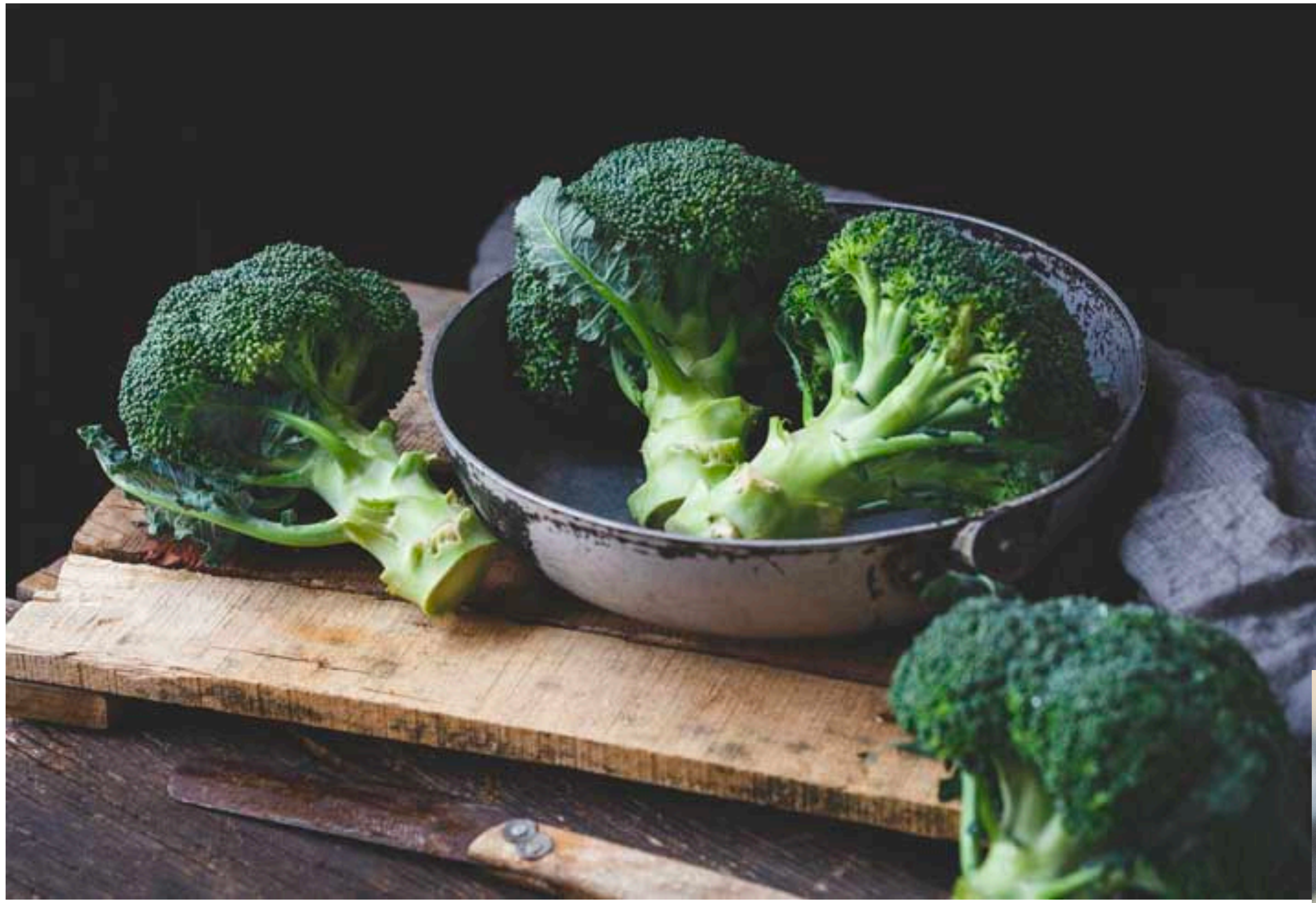
Which of the two motivates you more to learn how to cook eggs?

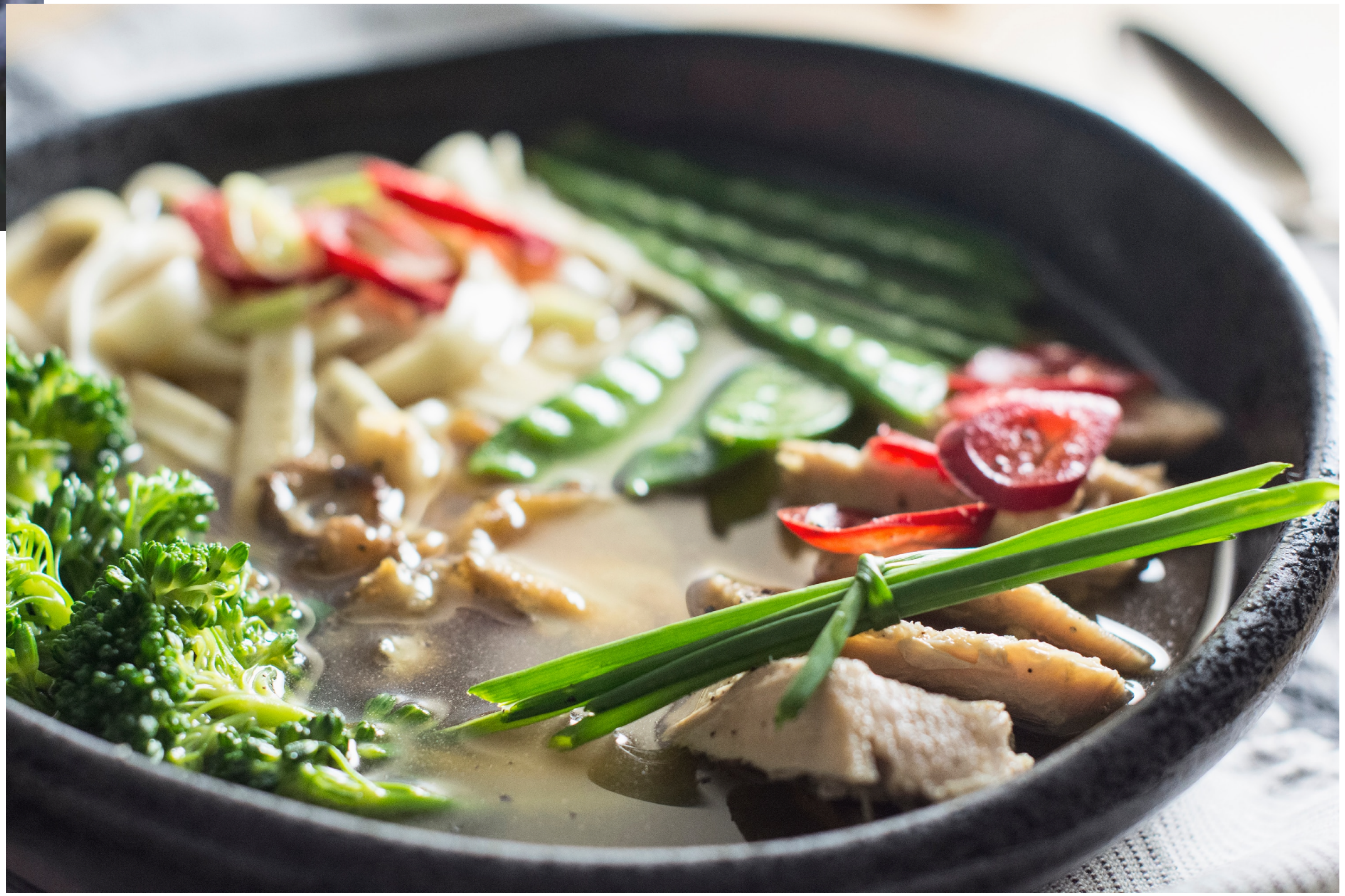
Which of the two motivates you more to learn how to cook eggs?

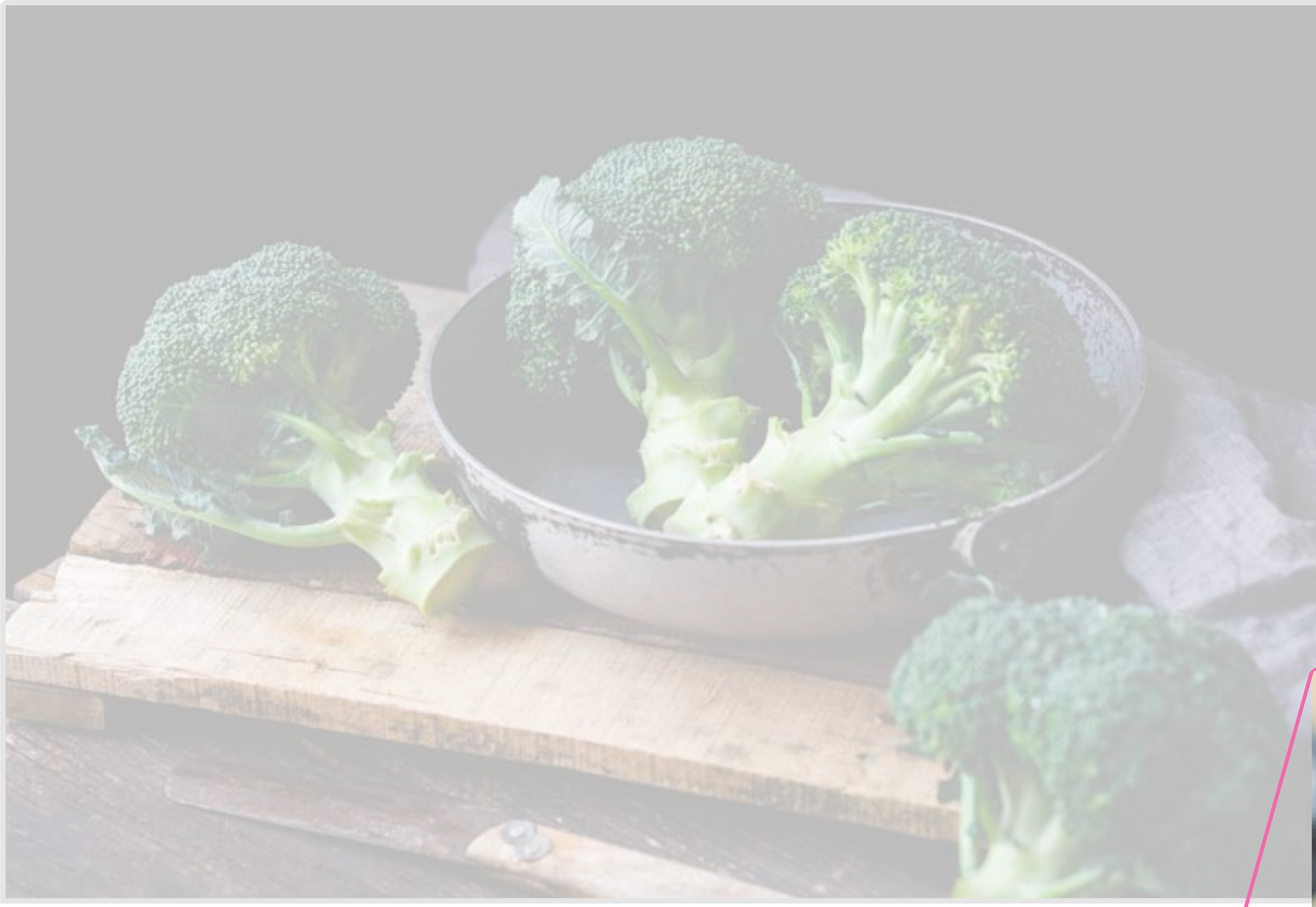
👶 Skip baby steps

But drilling through the baby steps can be useful [cite], this can happen outside of class with learnr tutorials (maybe a parson's problem here?)

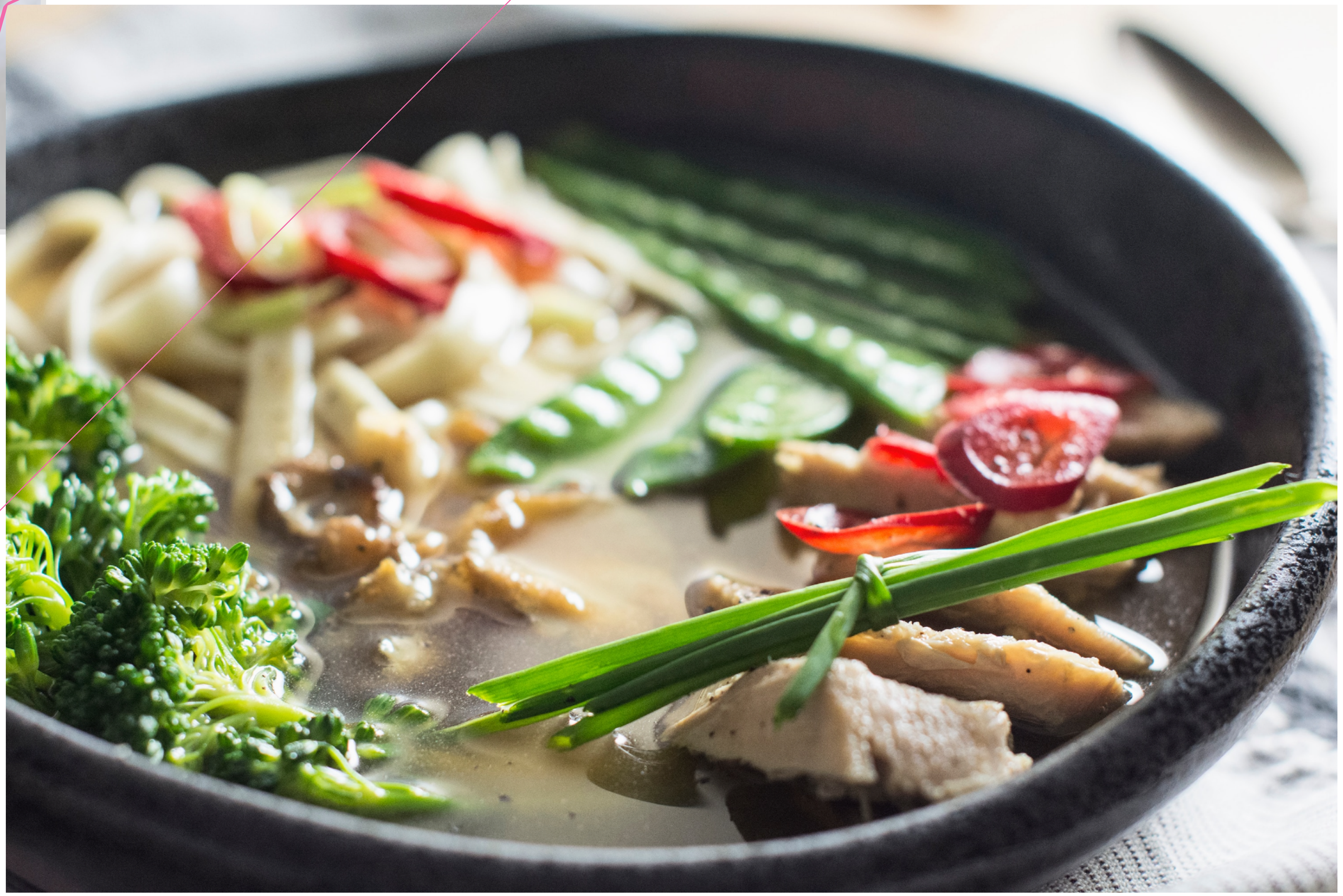
Which of the two following is more appetizing to someone who might not be a fan of broccoli?





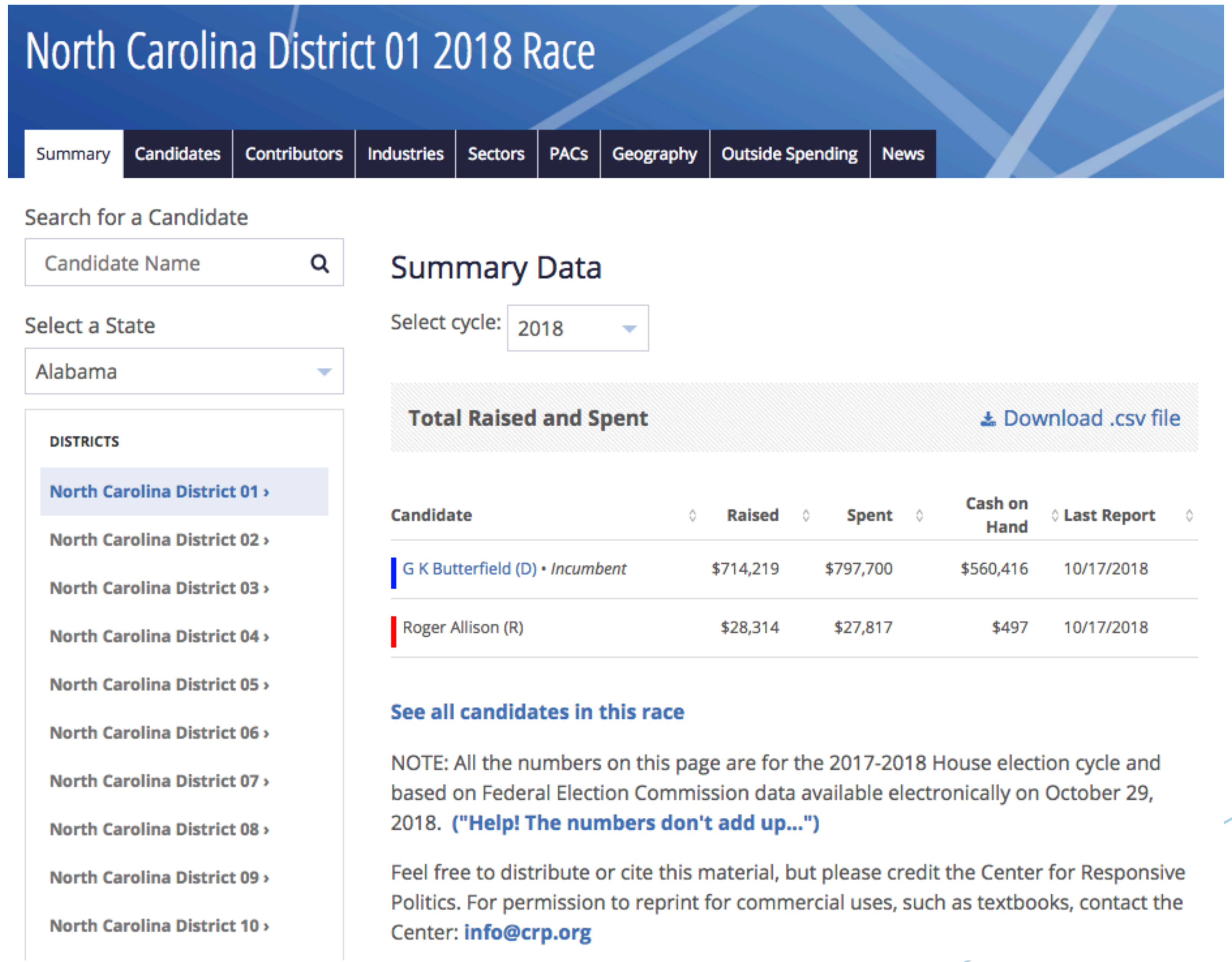


Which of the two following is more appetizing to someone who might not be a fan of broccoli?



Hide the veggies

- Today we go from this to that



North Carolina District 01 2018 Race

Search for a Candidate

Candidate Name

Select a State

Alabama

DISTRICTS

- North Carolina District 01 ›
- North Carolina District 02 ›
- North Carolina District 03 ›
- North Carolina District 04 ›
- North Carolina District 05 ›
- North Carolina District 06 ›
- North Carolina District 07 ›
- North Carolina District 08 ›
- North Carolina District 09 ›
- North Carolina District 10 ›

Summary Data

Select cycle: 2018

Total Raised and Spent

Download .csv file

Candidate	Raised	Spent	Cash on Hand	Last Report
G K Butterfield (D) • *Incumbent*	$714,219	$797,700	$560,416	10/17/2018
Roger Allison (R)	$28,314	$27,817	$497	10/17/2018

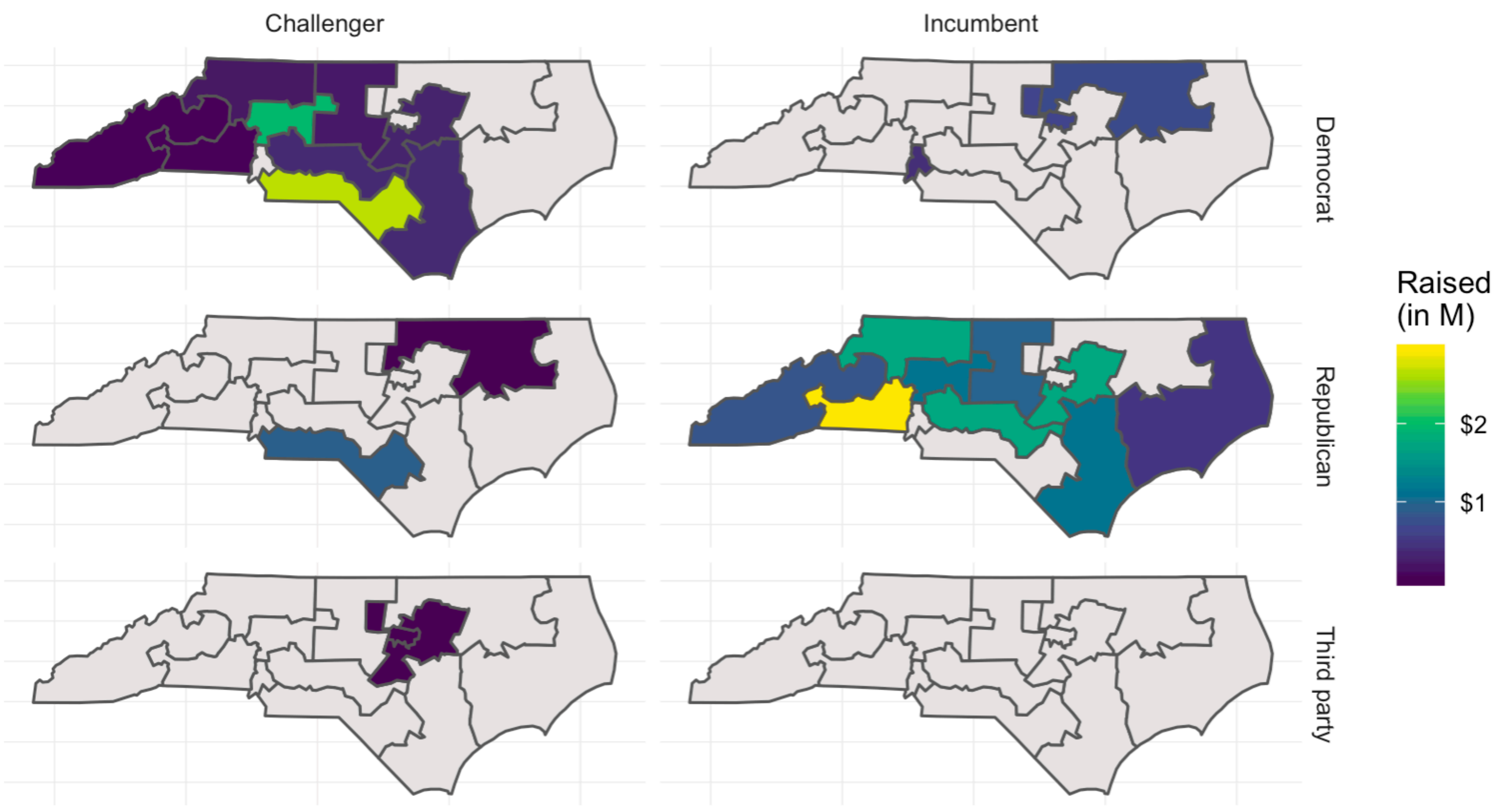
See all candidates in this race

NOTE: All the numbers on this page are for the 2017-2018 House election cycle and based on Federal Election Commission data available electronically on October 29, 2018. ("Help! The numbers don't add up...")

Feel free to distribute or cite this material, but please credit the Center for Responsive Politics. For permission to reprint for commercial uses, such as textbooks, contact the Center: info@crp.org

→

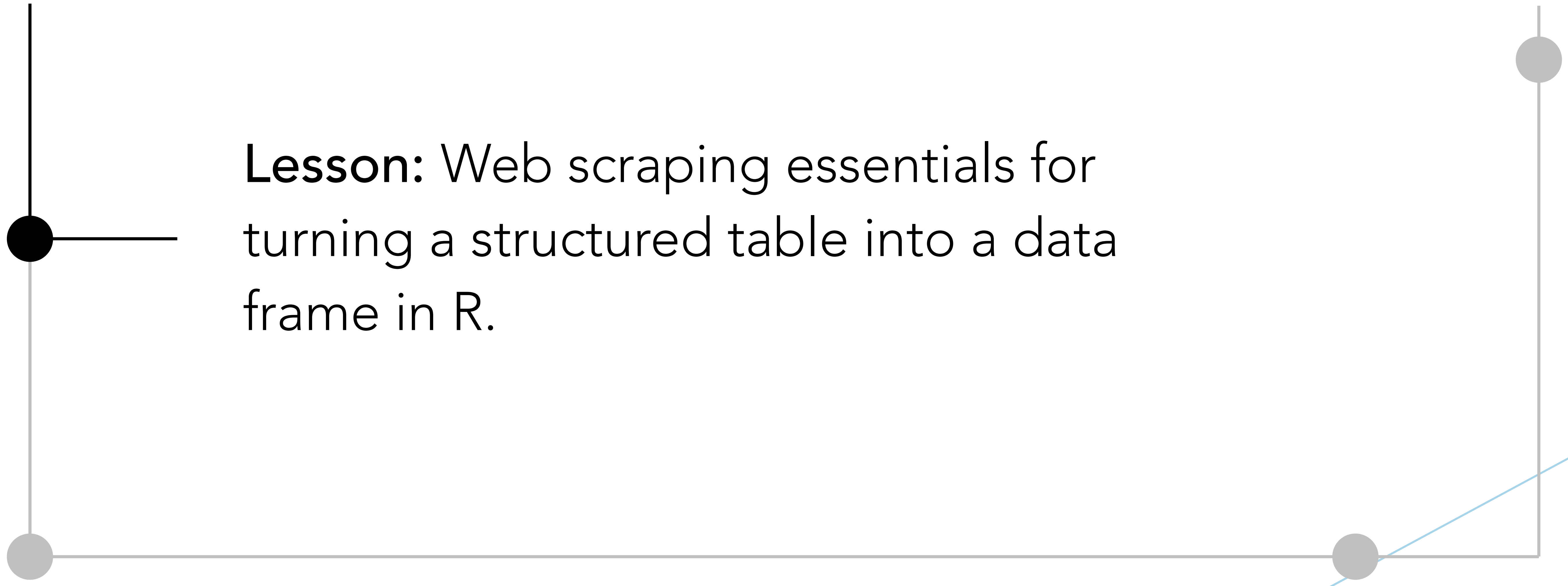
Political contributions for 2018 NC Congressional Races
as of 9/30/2018



Source: OpenSecrets.org

- And do so in a way that is easy to replicate for another state

🥦 Hide the veggies



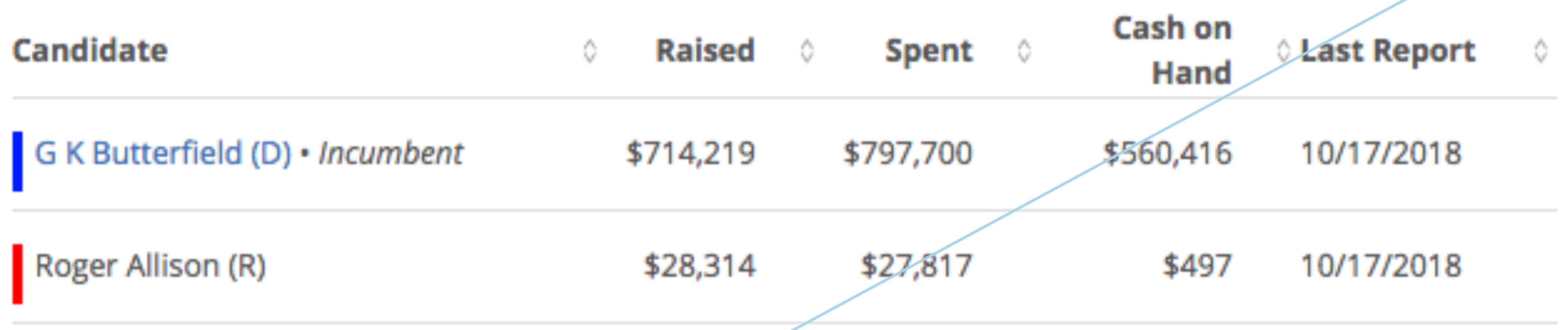
Lesson: Web scraping essentials for turning a structured table into a data frame in R.

DESIGN PRINCIPLES

Hide the veggies

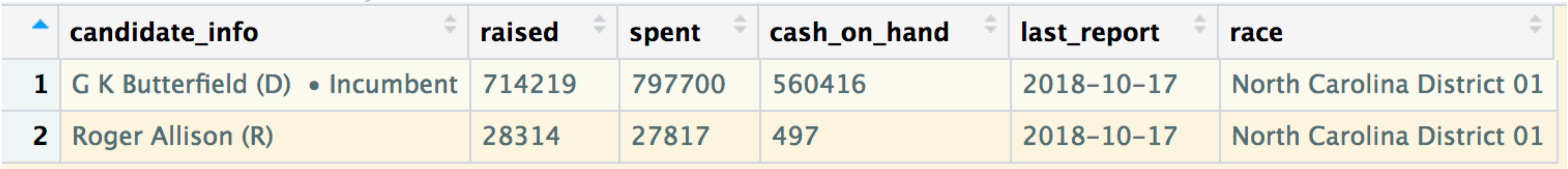
Lesson: Web scraping essentials for turning a structured table into a data frame in R.

Ex 1: Scrape the table off the web and save as a data frame.



Candidate	Raised	Spent	Cash on Hand	Last Report
G K Butterfield (D) • *Incumbent*	$714,219	$797,700	$560,416	10/17/2018
Roger Allison (R)	$28,314	$27,817	$497	10/17/2018

↓

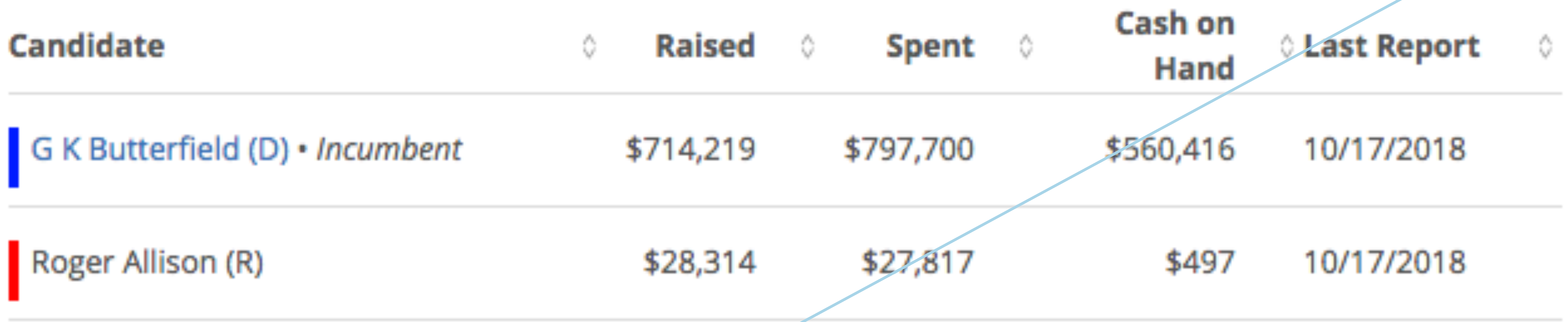


	candidate_info	raised	spent	cash_on_hand	last_report	race
1	G K Butterfield (D) • Incumbent	714219	797700	560416	2018-10-17	North Carolina District 01
2	Roger Allison (R)	28314	27817	497	2018-10-17	North Carolina District 01

Hide the veggies

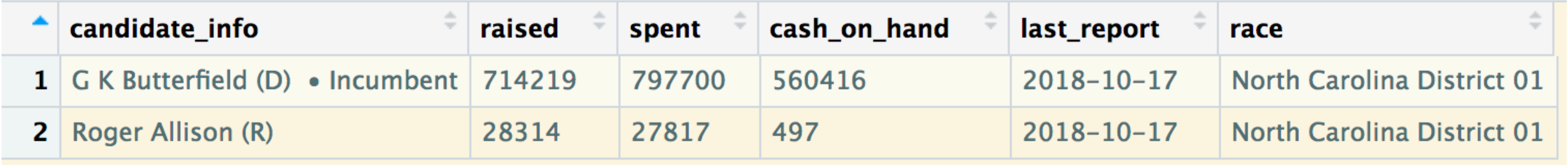
Lesson: Web scraping essentials for turning a structured table into a data frame in R.

Ex 1: Scrape the table off the web and save as a data frame.



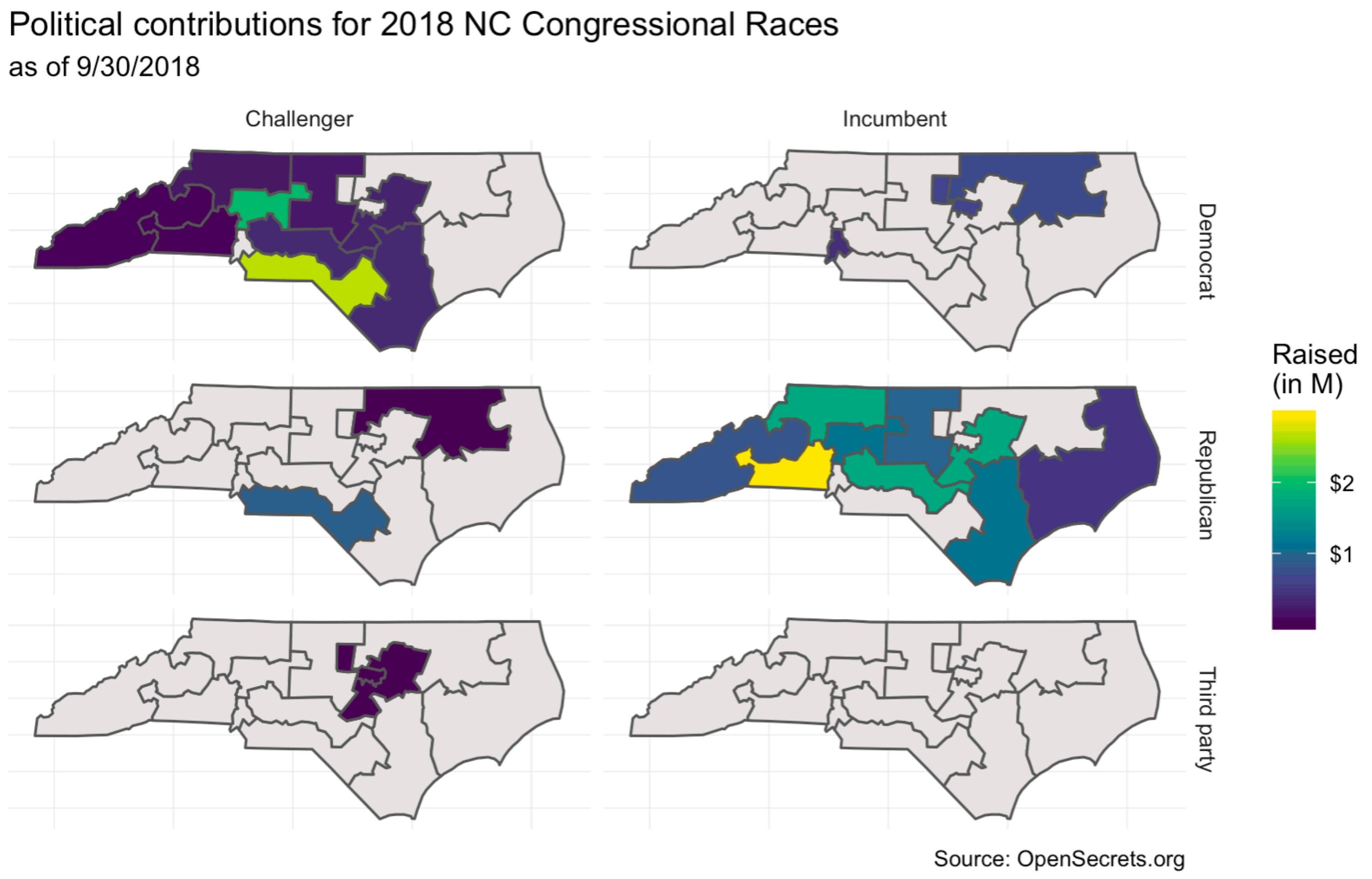
Candidate	Raised	Spent	Cash on Hand	Last Report
G K Butterfield (D) • *Incumbent*	$714,219	$797,700	$560,416	10/17/2018
Roger Allison (R)	$28,314	$27,817	$497	10/17/2018

↓



	candidate_info	raised	spent	cash_on_hand	last_report	race
1	G K Butterfield (D) • Incumbent	714219	797700	560416	2018-10-17	North Carolina District 01
2	Roger Allison (R)	28314	27817	497	2018-10-17	North Carolina District 01

Ex 2: What other information do we need represented as variables to make this figure?



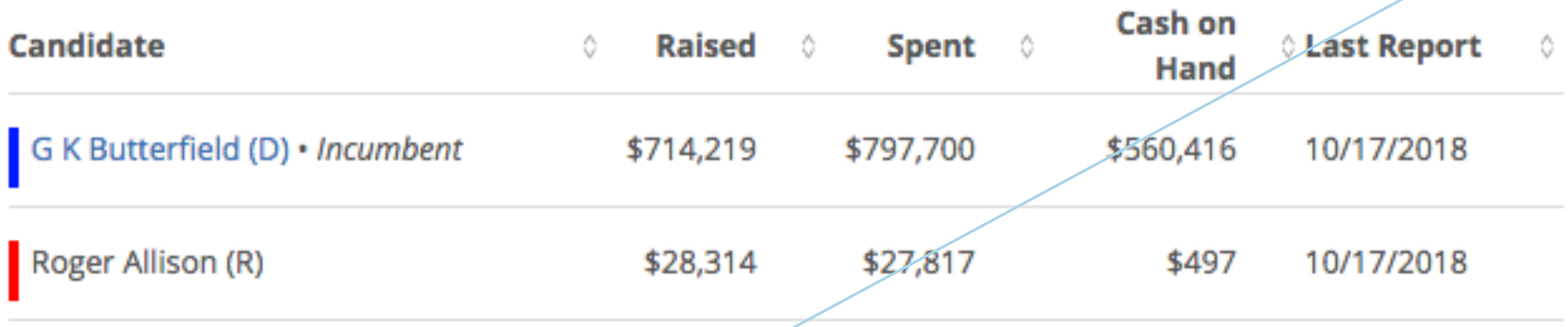
Political contributions for 2018 NC Congressional Races
as of 9/30/2018

Source: OpenSecrets.org

Hide the veggies

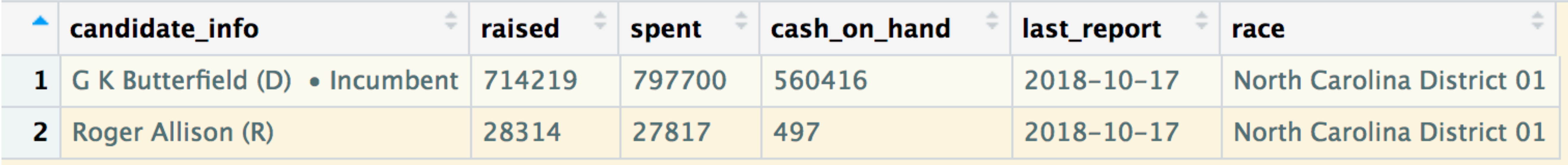
Lesson: Web scraping essentials for turning a structured table into a data frame in R.

Ex 1: Scrape the table off the web and save as a data frame.



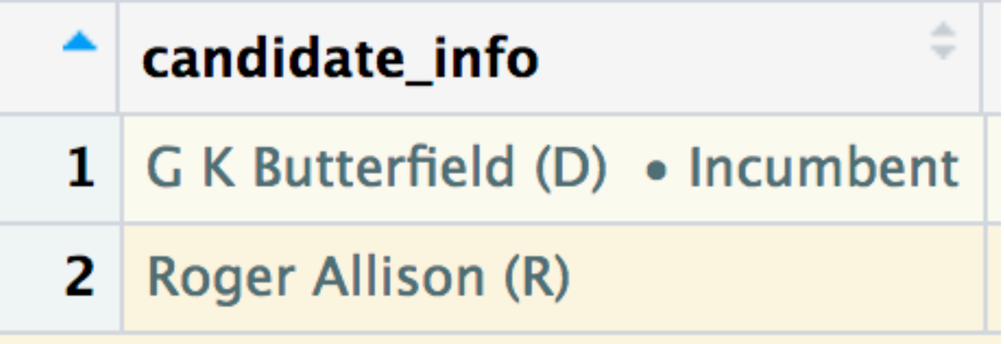
Candidate	Raised	Spent	Cash on Hand	Last Report
G K Butterfield (D) • *Incumbent*	$714,219	$797,700	$560,416	10/17/2018
Roger Allison (R)	$28,314	$27,817	$497	10/17/2018

↓

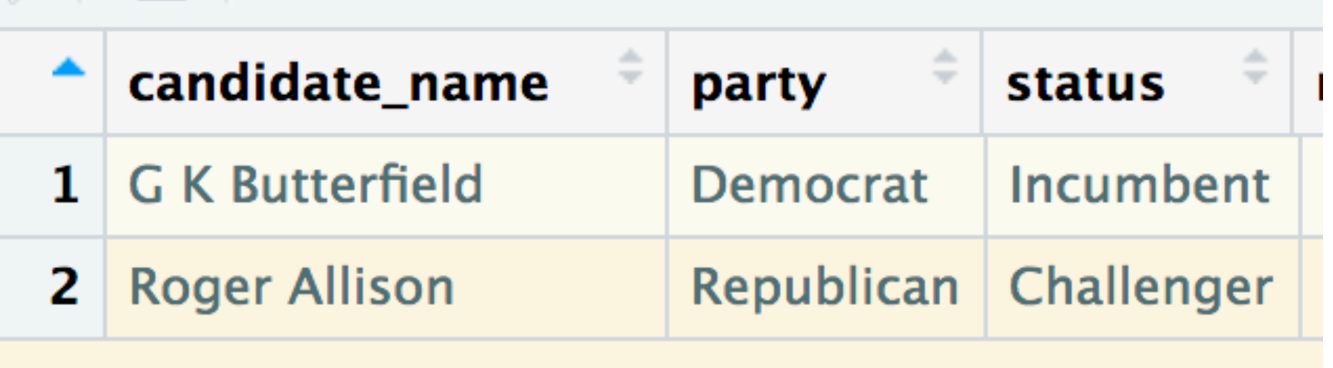


	candidate_info	raised	spent	cash_on_hand	last_report	race
1	G K Butterfield (D) • Incumbent	714219	797700	560416	2018-10-17	North Carolina District 01
2	Roger Allison (R)	28314	27817	497	2018-10-17	North Carolina District 01

Lesson: "Just enough" regex

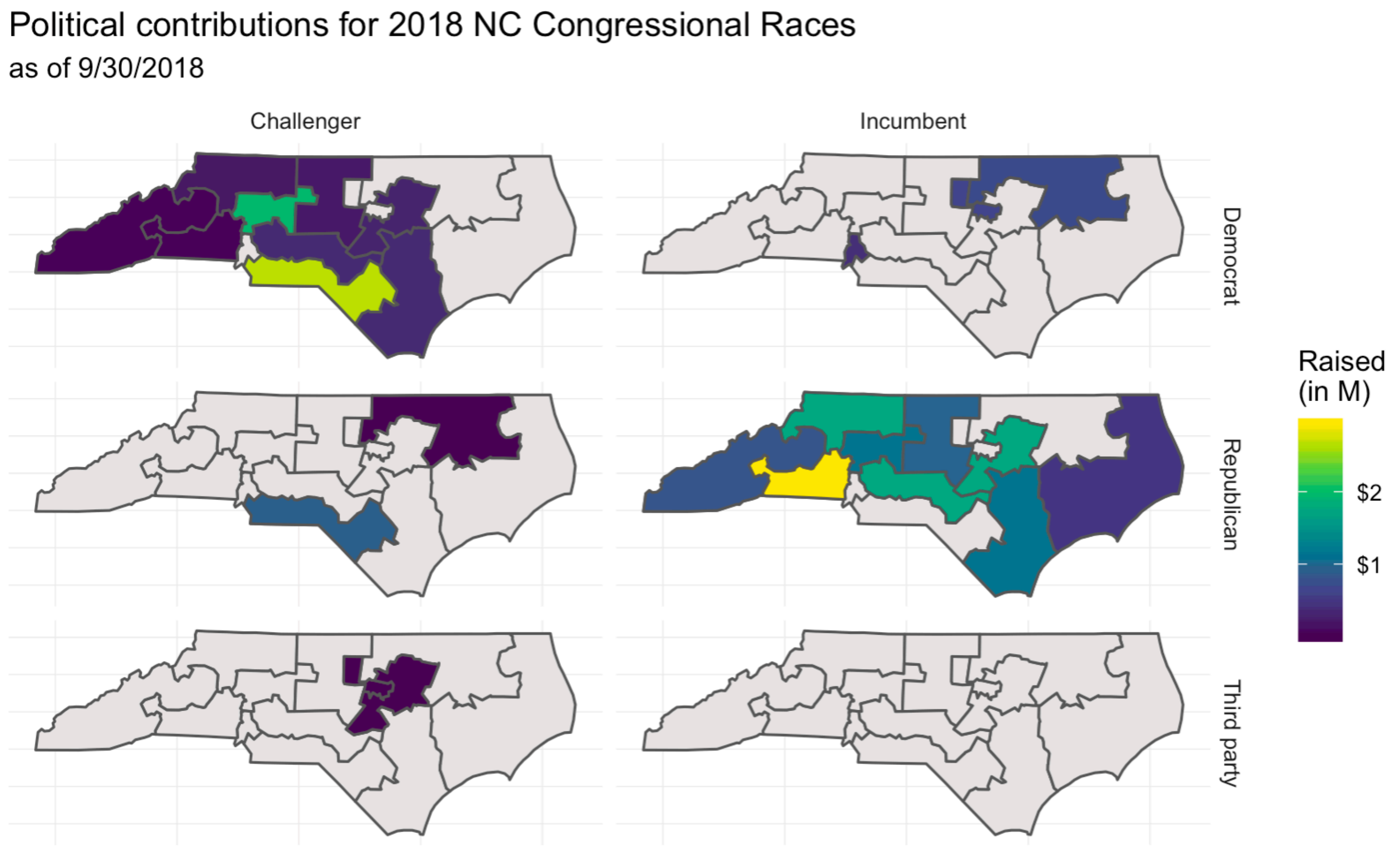


	candidate_info
1	G K Butterfield (D) • Incumbent
2	Roger Allison (R)



	candidate_name	party	status
1	G K Butterfield	Democrat	Incumbent
2	Roger Allison	Republican	Challenger

Ex 2: What other information do we need represented as variables to make this figure?



Source: OpenSecrets.org

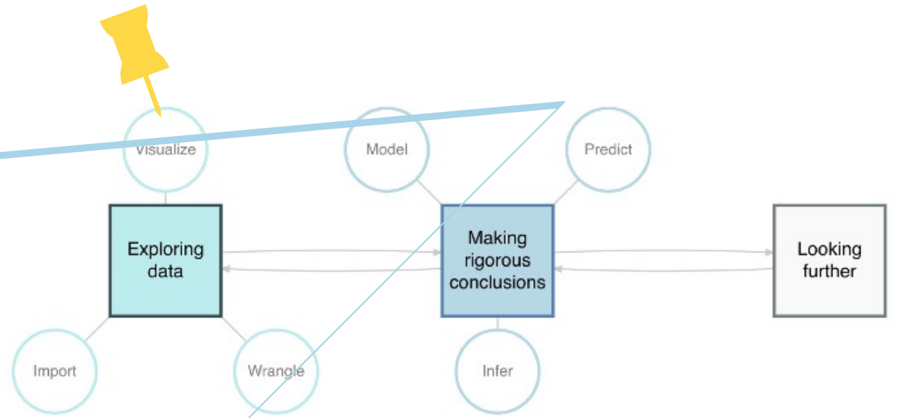
DESIGN PRINCIPLES

If you already have ingredients and tools to bake a cake, which of these will be easier to also prepare?

DESIGN PRINCIPLES

If you already have ingredients and tools to bake a cake, which of these will be easier to also prepare?

🌲 Leverage the ecosystem



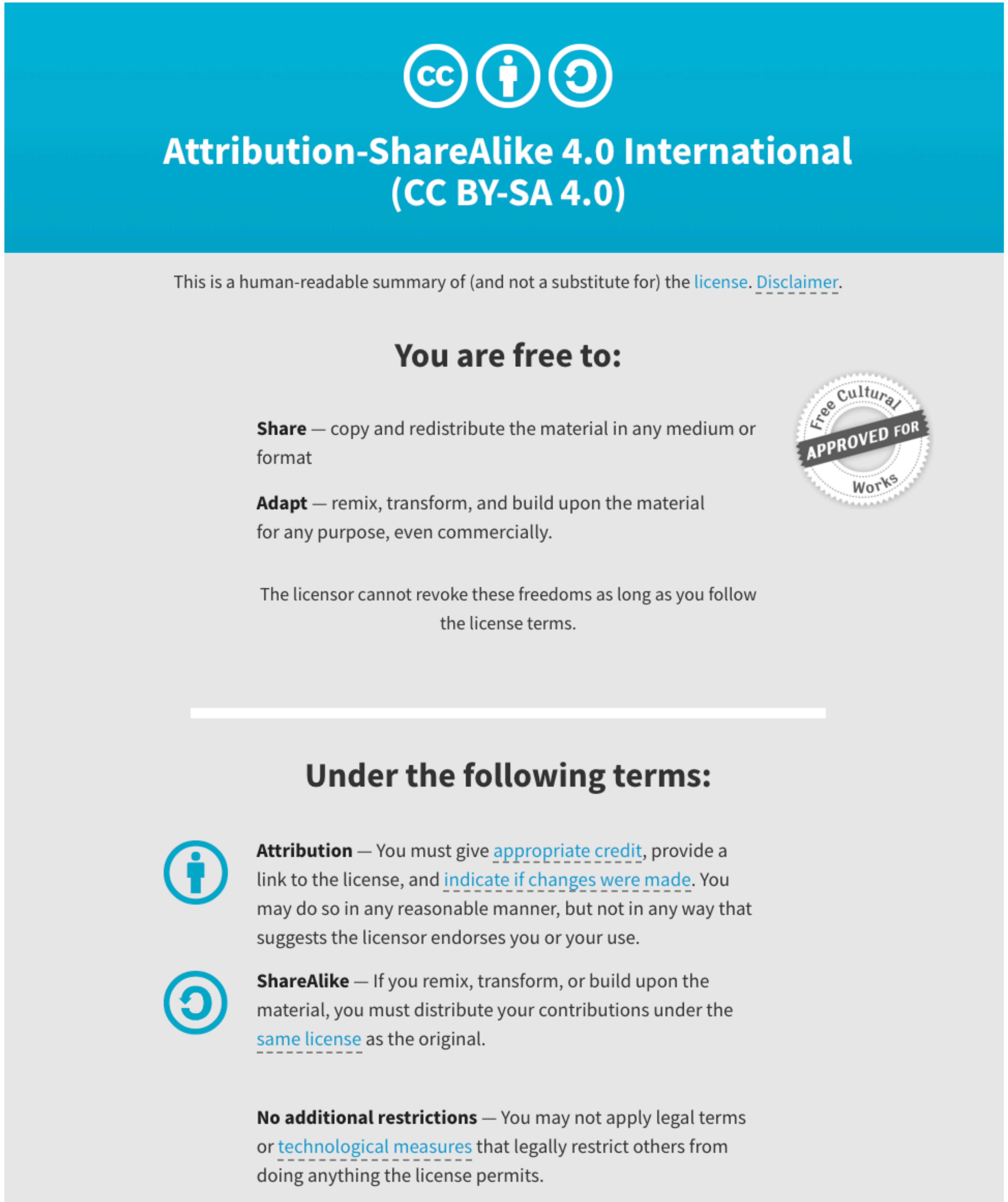
DESIGN PRINCIPLES

USAGE

1 Use it in full to jumpstart / overhaul your teaching

2 Use it in bits and pieces to supplement your teaching

LICENSE



Attribution-ShareAlike 4.0 International (CC BY-SA 4.0)

This is a human-readable summary of (and not a substitute for) the license. Disclaimer.

You are free to:

Share — copy and redistribute the material in any medium or format

Adapt — remix, transform, and build upon the material for any purpose, even commercially.

The licensor cannot revoke these freedoms as long as you follow the license terms.

Under the following terms:

Attribution — You must give appropriate credit, provide a link to the license, and indicate if changes were made. You may do so in any reasonable manner, but not in any way that suggests the licensor endorses you or your use.

ShareAlike — If you remix, transform, or build upon the material, you must distribute your contributions under the same license as the original.

No additional restrictions — You may not apply legal terms or technological measures that legally restrict others from doing anything the license permits.

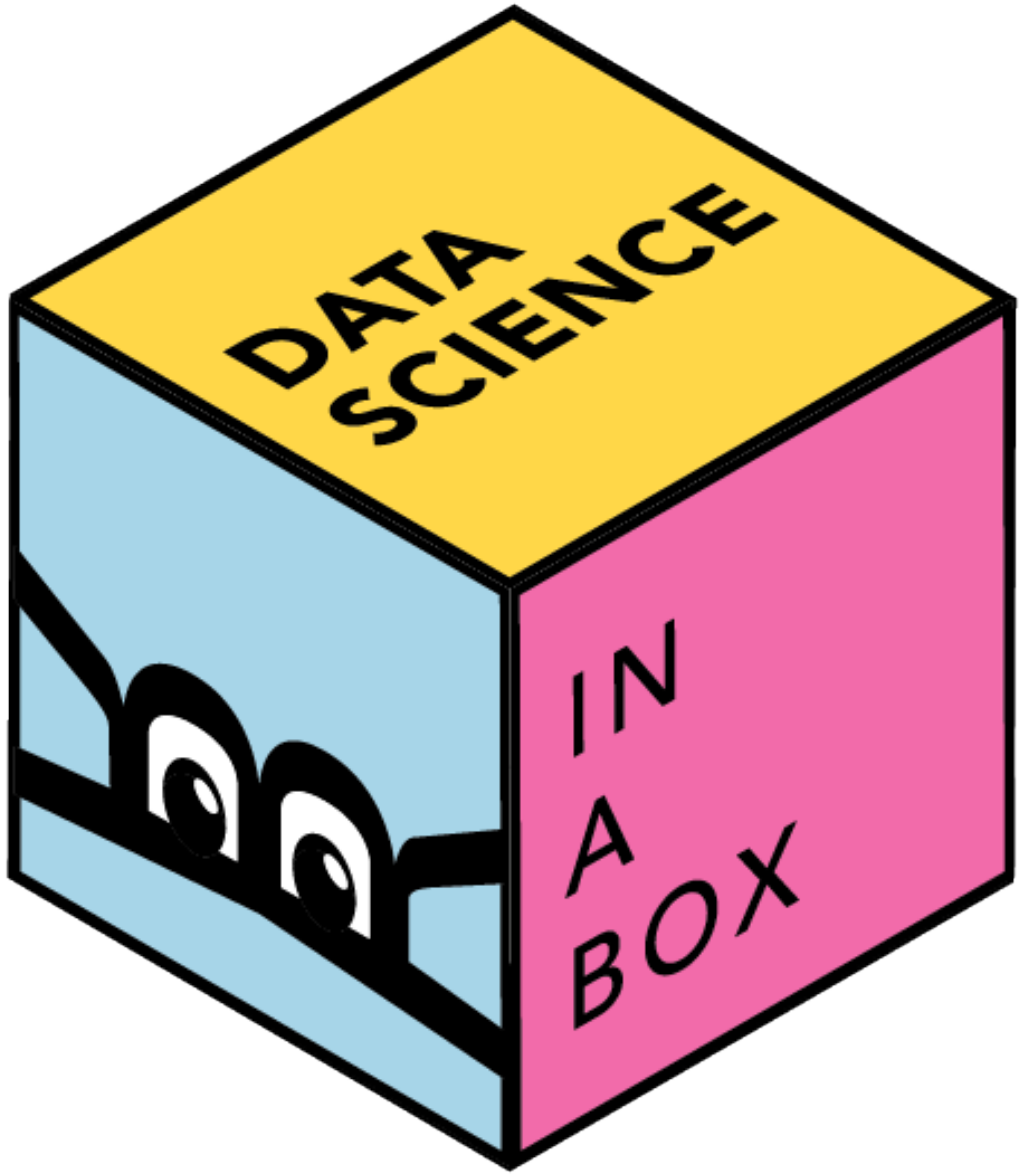
FUTURE

1 Scalability

- More formative assessments via **learnr**
- Automated feedback
- Peer review

2 Assessment

- Curriculum: How are students learning?
- Impact: How are these resources being used?



Add link

MINE ÇETINKAYA-RUNDEL

UNIVERSITY OF EDINBURGH + RSTUDIO

@minebocek

mine-cetinkaya-rundel

cetinkaya.mine@gmail.com

validated

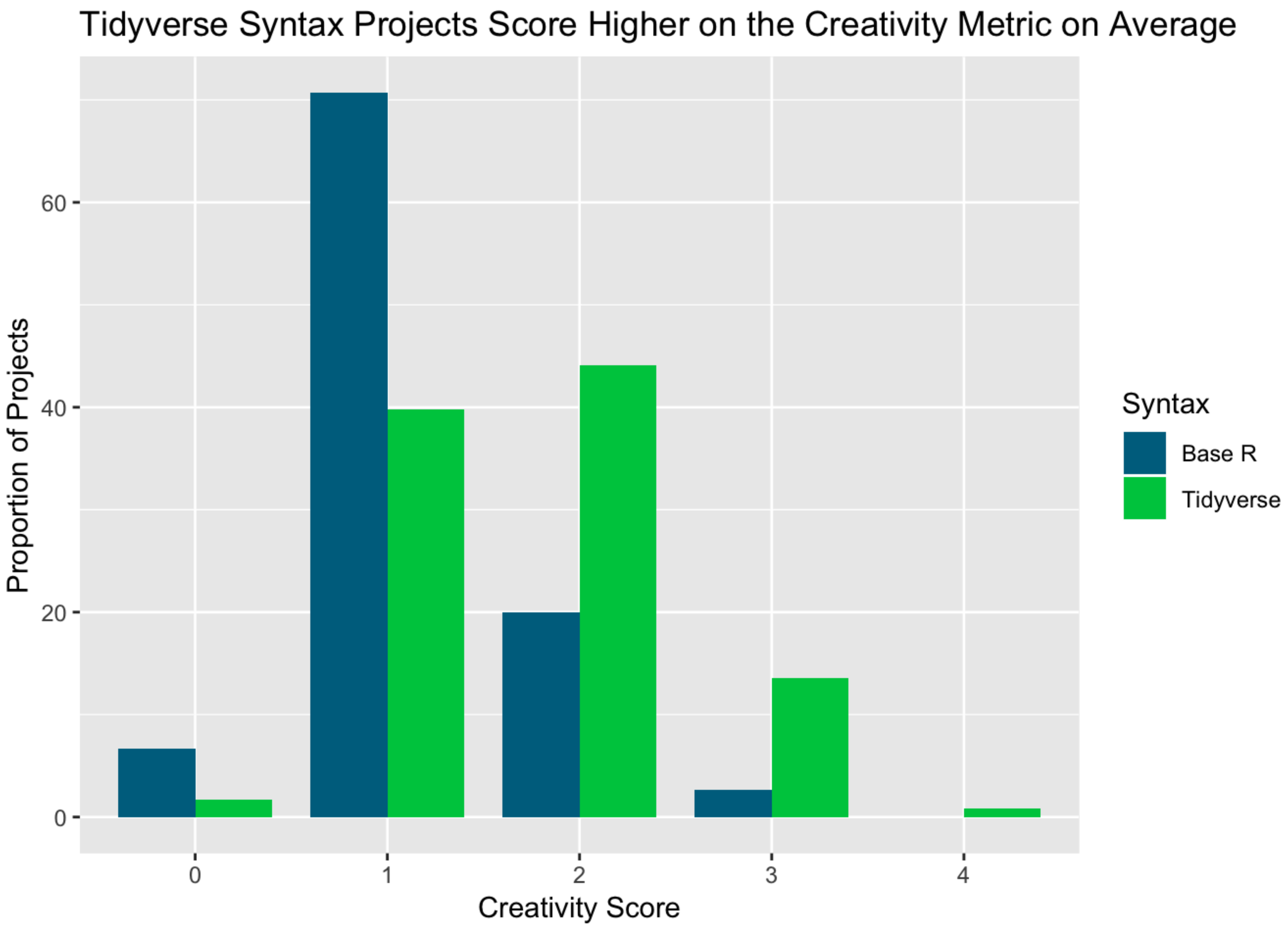
Retrospective study of 205 open ended student projects

- on **creativity**, **depth** and the complexity of **multivariate visualizations**
- compared across students who learned R using **base R** syntax vs. **tidyverse**

validated

Creativity:

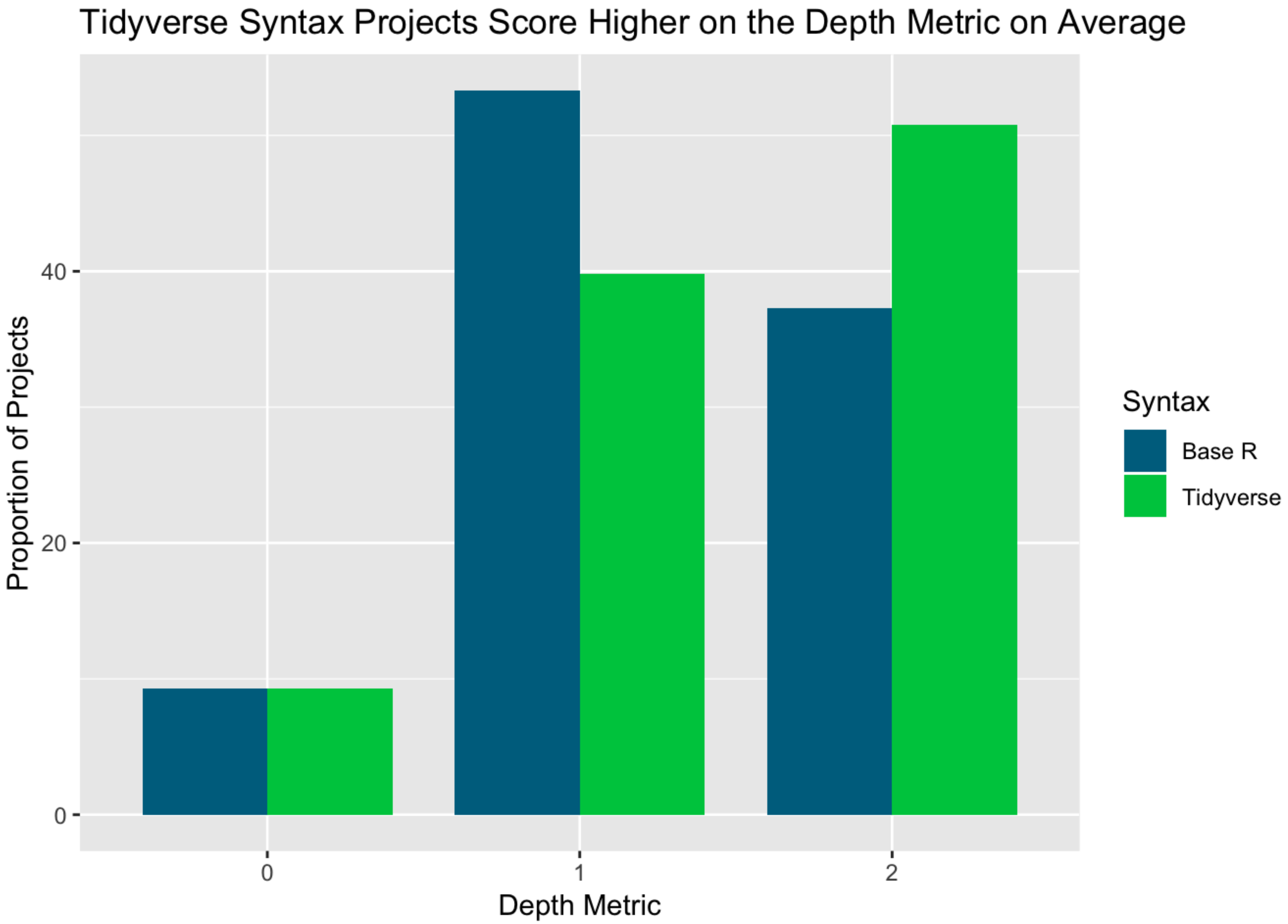
1. Creation of new variable(s) based on existing variables
2. Transformation of existing variables
3. Existence of a subgroup analysis
4. Use of a subset of the dataset for all steps of the project



validated

Depth:

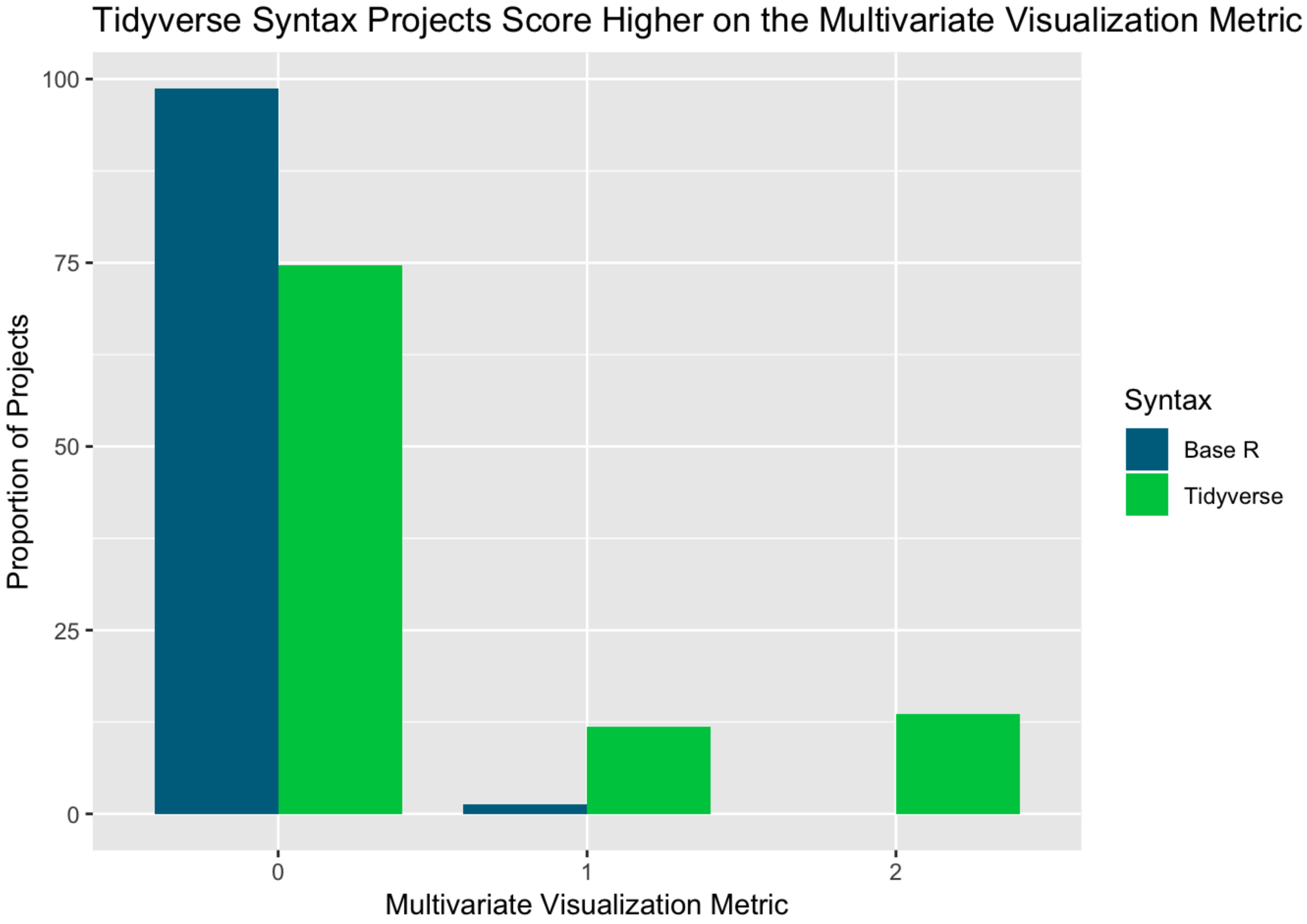
1. Presence of consistent theme throughout the project
2. Use of relevant data



validated

Multivariate visualizations:

1. Presence of a visualization with 3+ variables
2. Interpretation of the multivariate visualization


Tidyverse Syntax Projects Score Higher on the Multivariate Visualization Metric
Proportion of Projects
Multivariate Visualization Metric
Syntax
Base R
Tidyverse