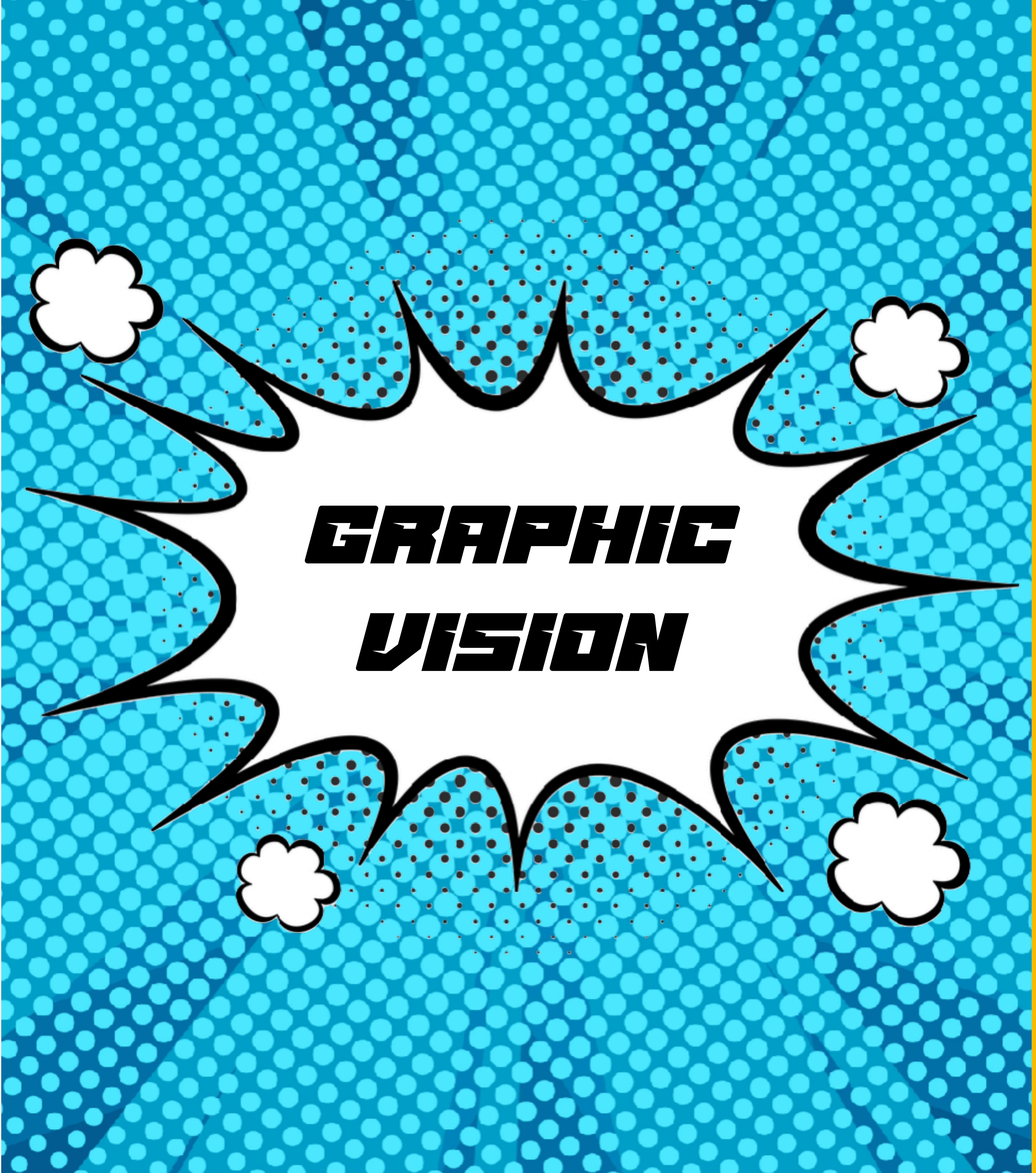
My toolbox is full of shiny tools, do I also need super powers?

🔗 bit.ly/superpowers-ecots22

Mine Çetinkaya-Rundel
Duke University / RStudio

data science

```
> super
> power
```



```
> data
> visualization
```

Data visualization

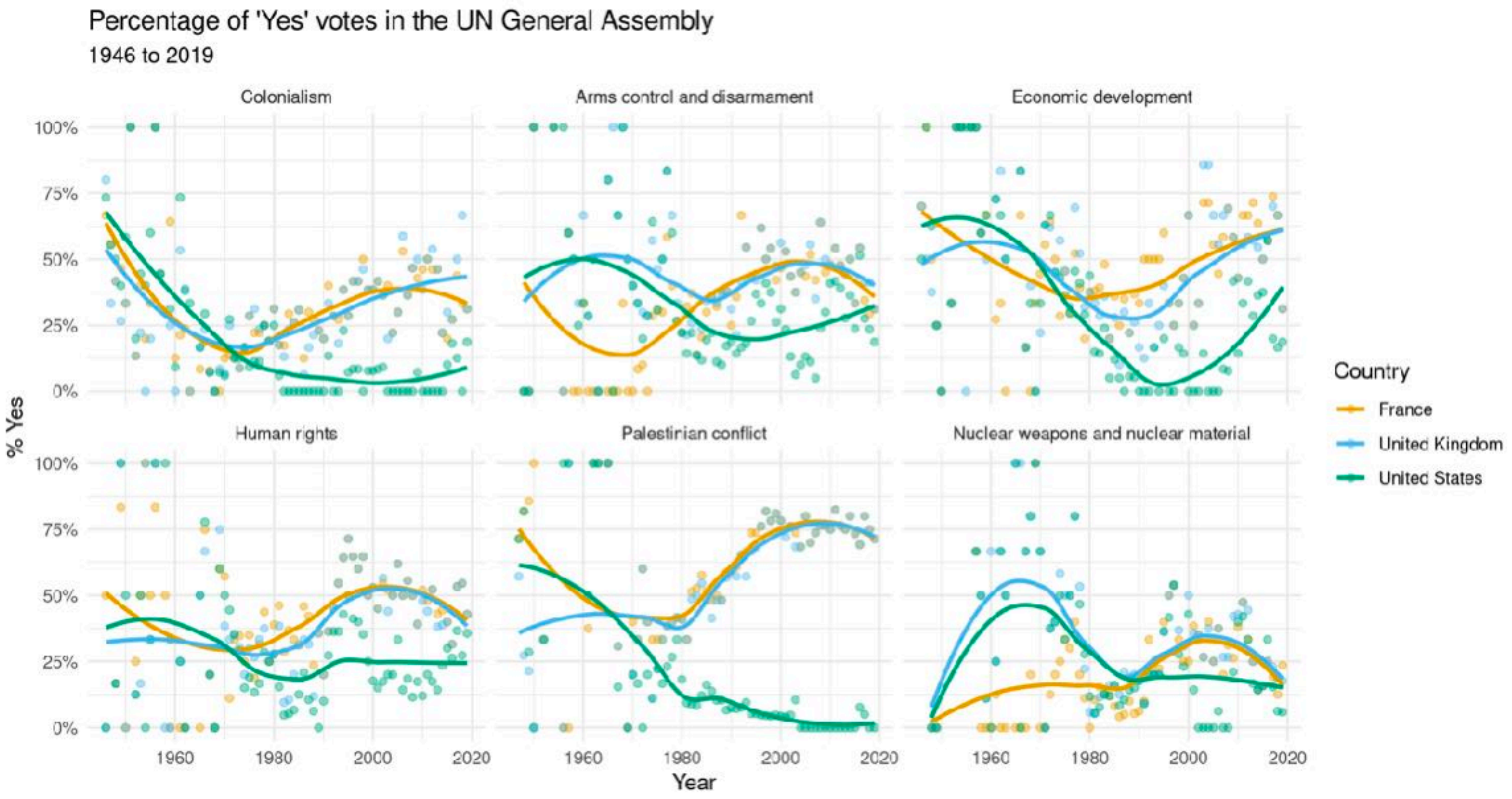
GRAPHIC VISION

- Start, literally, **on day one** and continue improving throughout the curriculum
- Teach it to
  - motivate **inquiry and exploration**
  - support **multivariate thinking**
  - effectively **communicate** of results and findings
  - advance **programming** skills
  - aid **inferential** decisions

Data visualization on day one

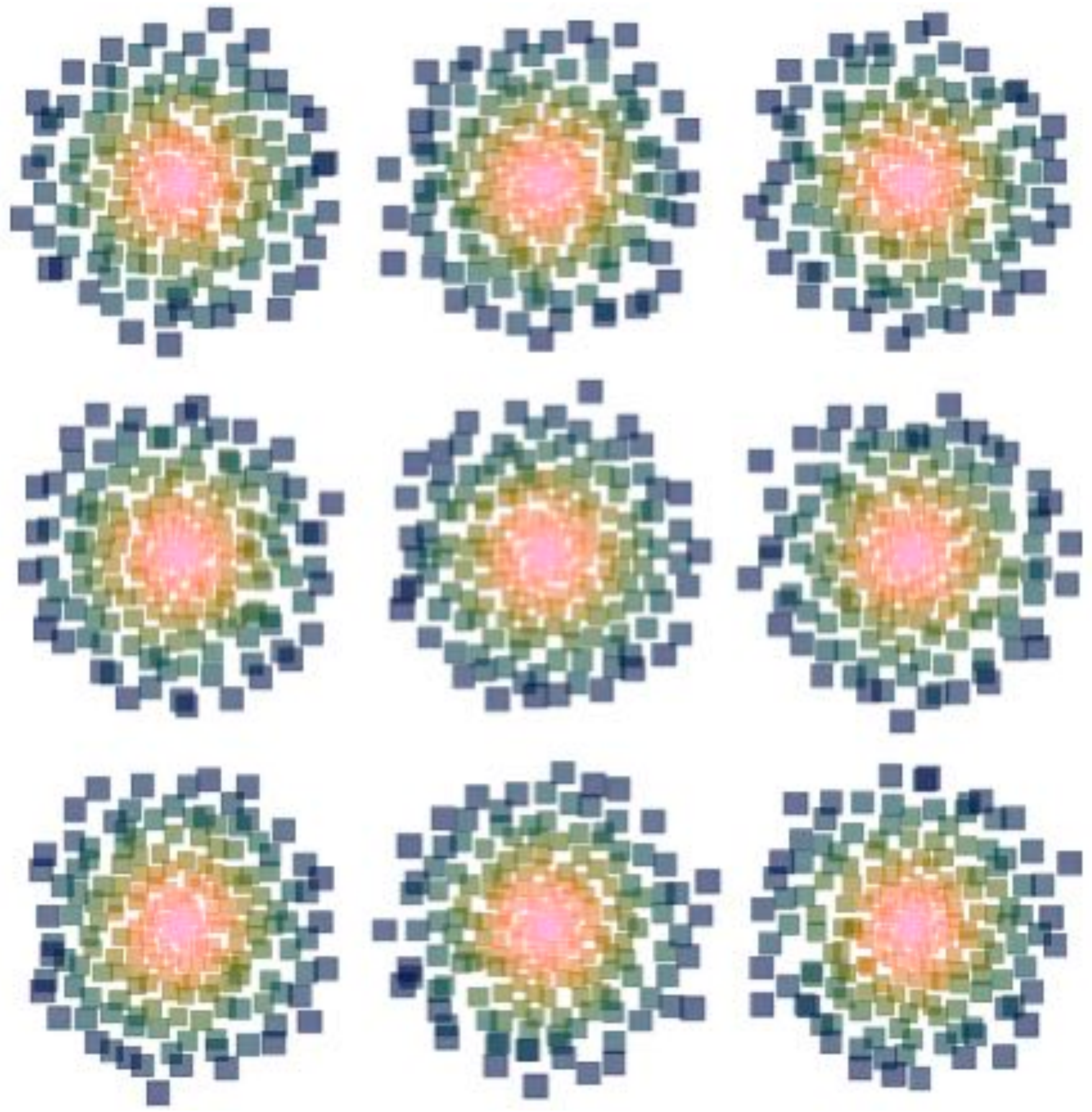
- Ready to go computing environment
- Reproducible document with code to produce the visualization
- Code that's obviously straightforward to modify for customizing the plot

```
unvotes |>
  filter(country %in% c("United Kingdom",
         "United States", "France")) |>
  ggplot(…)
```



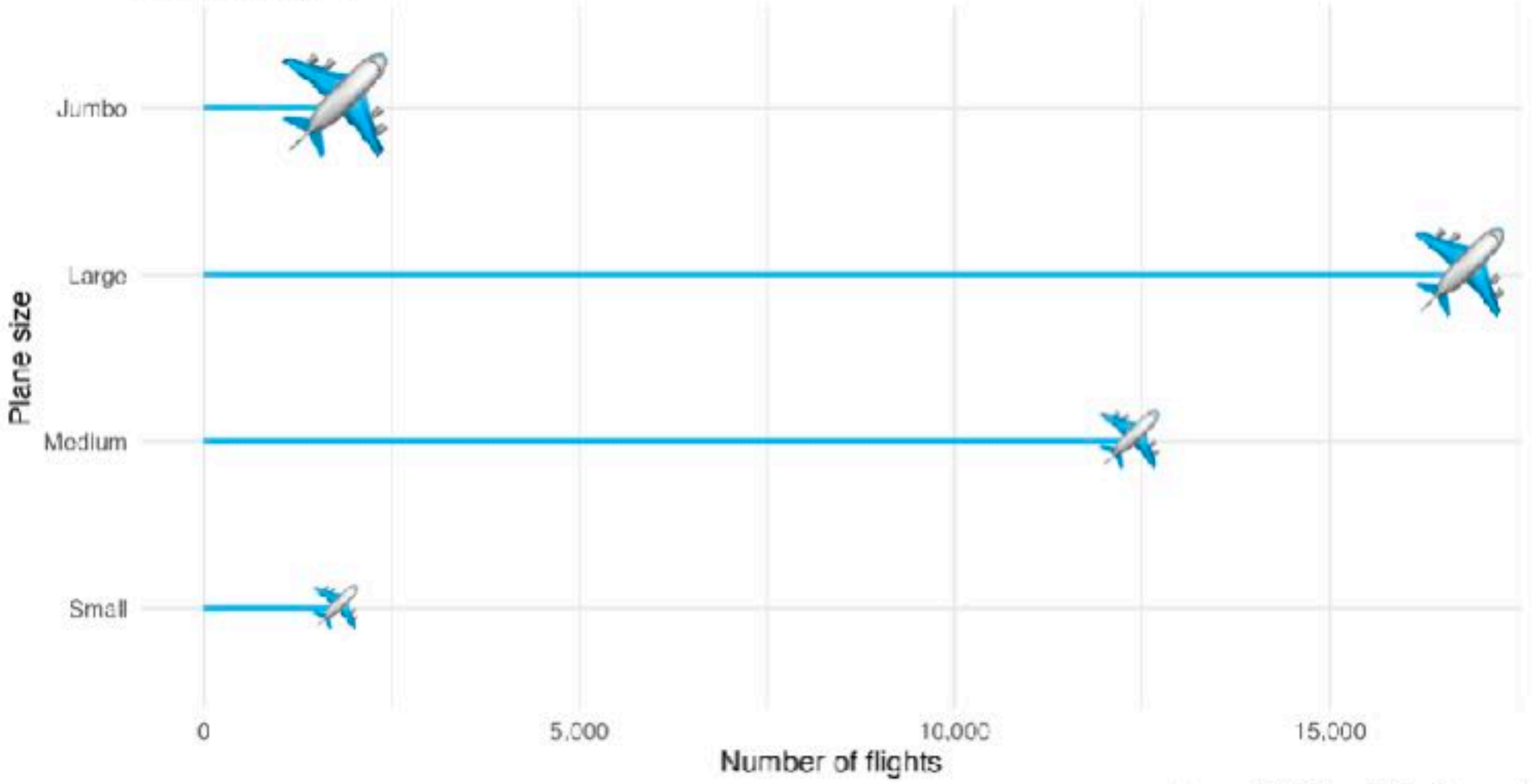
Data visualization later in curriculum

- “Recreate” to advance programming skills



Number of flights out of RDU in 2020

By size of plane

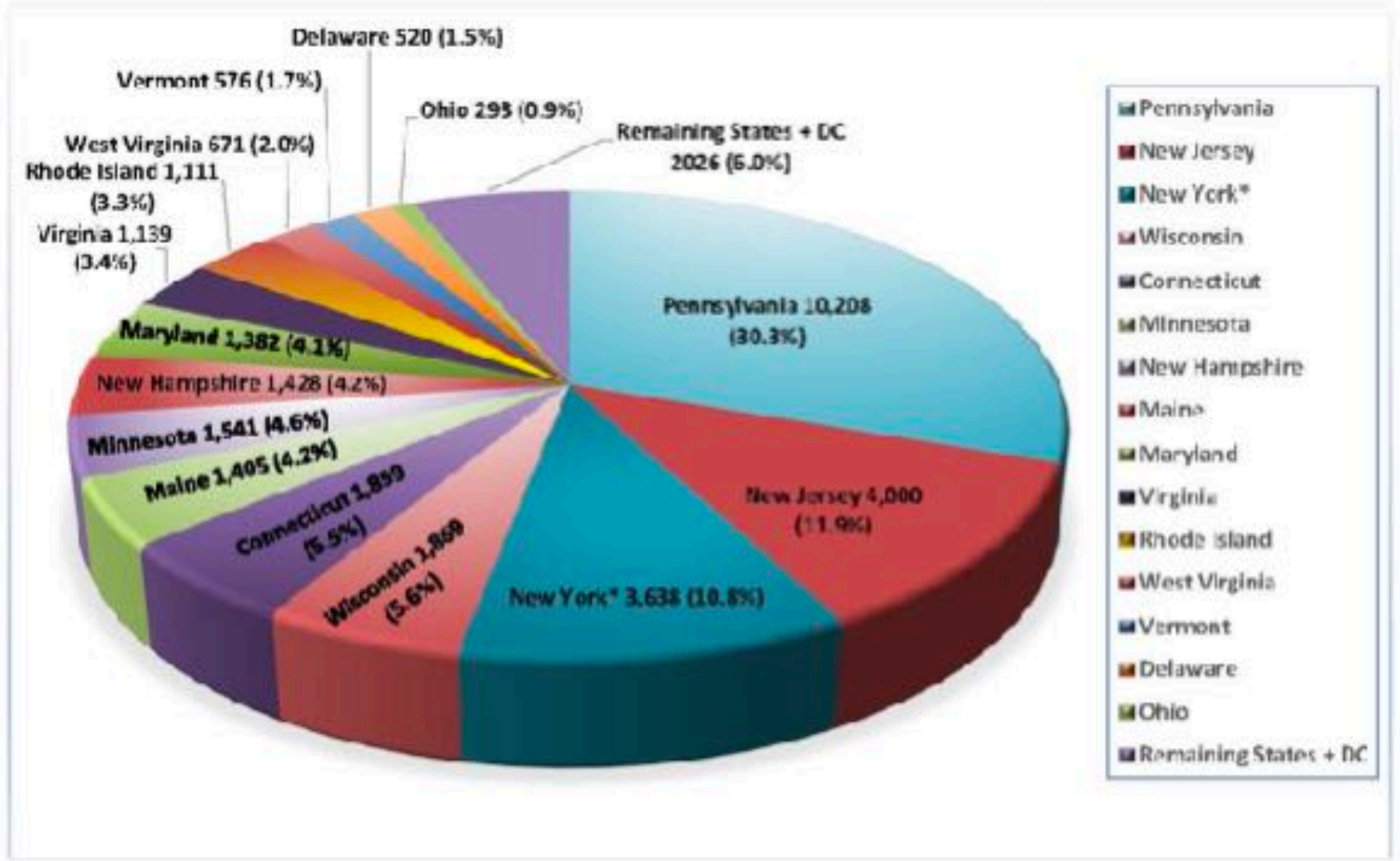


Source: FAA Aircraft Registry and Bureau of Transportation Statistics

Data visualization later in curriculum

- “Recreate” to advance programming skills
- “Recreate, then improve” to advance programming and communication skills

Lyme Disease Association Lyme Disease Analysis
2018 US Reported Lyme Disease Cases Featuring Top 15 States



U.S. TOTAL CASES 2018: 33,656

Source data compiled from CDC pub. data (DVBD)

Note1: CDC adopted a change in case definition in 2008. In addition to "Confirmed" cases, the "Probable" category was reported out for the first time

Note2 : CDC has stated, and confirmed in 2013, that only 10% of Lyme disease cases meeting the surveillance definition are reported – for example, if 30,000 cases are reported, 300,000 cases occurred (number does not include all the cases falling outside the stringent surveillance case definition).

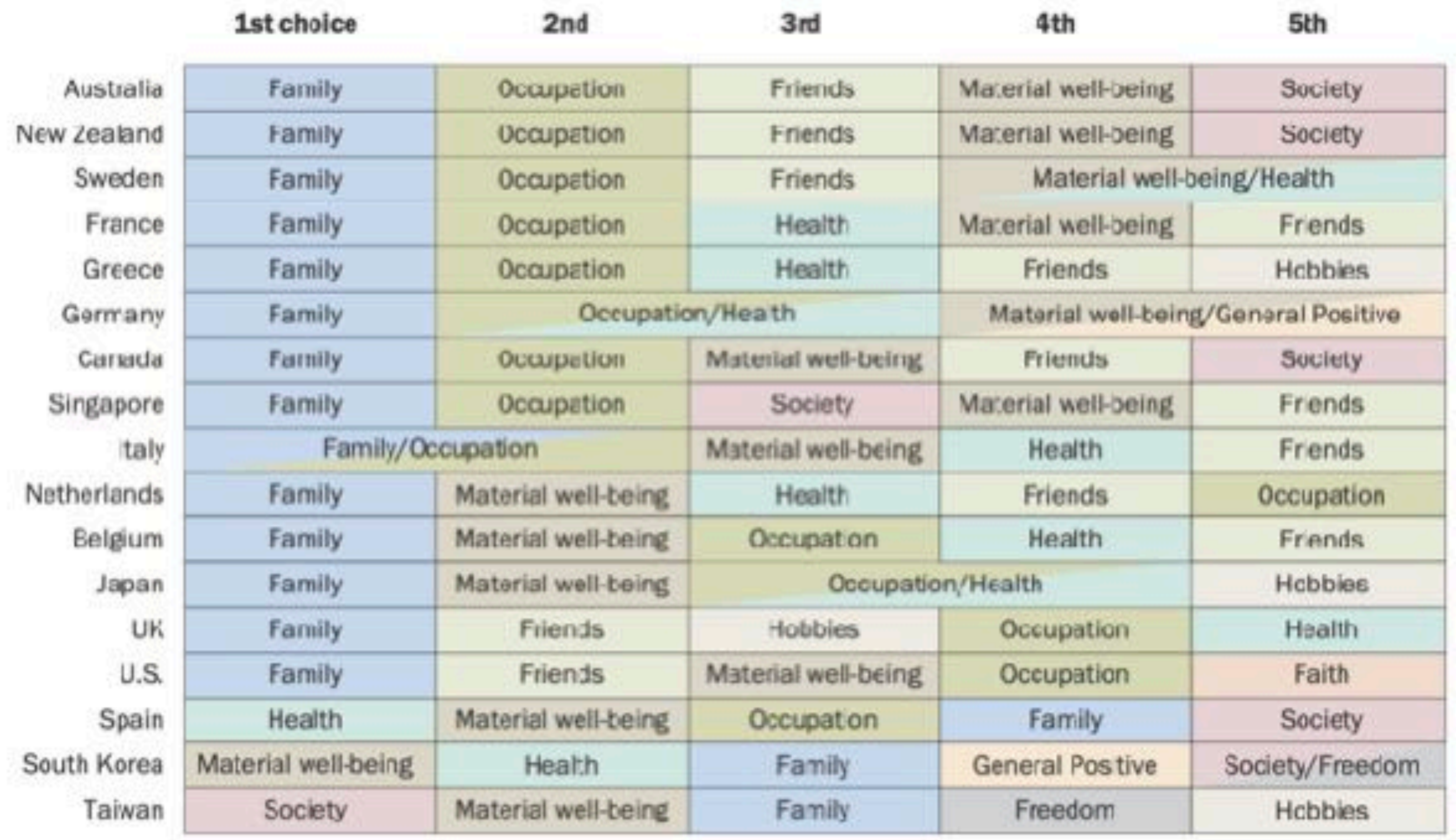
Note3: In 2016, MA changed reporting requirements and very few MA cases are now counted by CDC

*In recent years, an increasing number of NY Counties have used estimating to determine Lyme case numbers. The Council of State & Territorial Epidemiologists, in charge of surveillance, doesn't permit estimation to be reported by CDC in the national counts. In 2018, NY State reported it had 7,320 Lyme cases including those 30 estimated county numbers. CDC reported 3,638 Lyme cases for NY state, which excluded the 30 counties estimated numbers.

© 2019 Lyme Disease Association, Inc. (LDA), PO Box 1438, Jackson, NJ 08527 (888) 366-6611 www.LymeDiseaseAssociation.org

While family, careers, material well-being, friends and health are all top sources of meaning, they vary in importance across publics surveyed

Ranked choice among 17 topics coded as part of what gives people meaning in life



	1st choice	2nd	3rd	4th	5th
Australia	Family	Occupation	Friends	Material well-being	Society
New Zealand	Family	Occupation	Friends	Material well-being	Society
Sweden	Family	Occupation	Friends	Material well-being/Health	
France	Family	Occupation	Health	Material well-being	Friends
Greece	Family	Occupation	Health	Friends	Hobbies
Germany	Family	Occupation/Health		Material well-being/General Positive	
Canada	Family	Occupation	Material well-being	Friends	Society
Singapore	Family	Occupation	Society	Material well-being	Friends
Italy	Family/Occupation		Material well-being	Health	Friends
Netherlands	Family	Material well-being	Health	Friends	Occupation
Belgium	Family	Material well-being	Occupation	Health	Friends
Japan	Family	Material well-being	Occupation/Health		Hobbies
UK	Family	Friends	Hobbies	Occupation	Health
U.S.	Family	Friends	Material well-being	Occupation	Faith
Spain	Health	Material well-being	Occupation	Family	Society
South Korea	Material well-being	Health	Family	General Positive	Society/Freedom
Taiwan	Society	Material well-being	Family	Freedom	Hobbies

Note: Open-ended question. Rank reflects where the topic fell in a list of 17 sources of meaning that were coded. See Appendix A for more information.

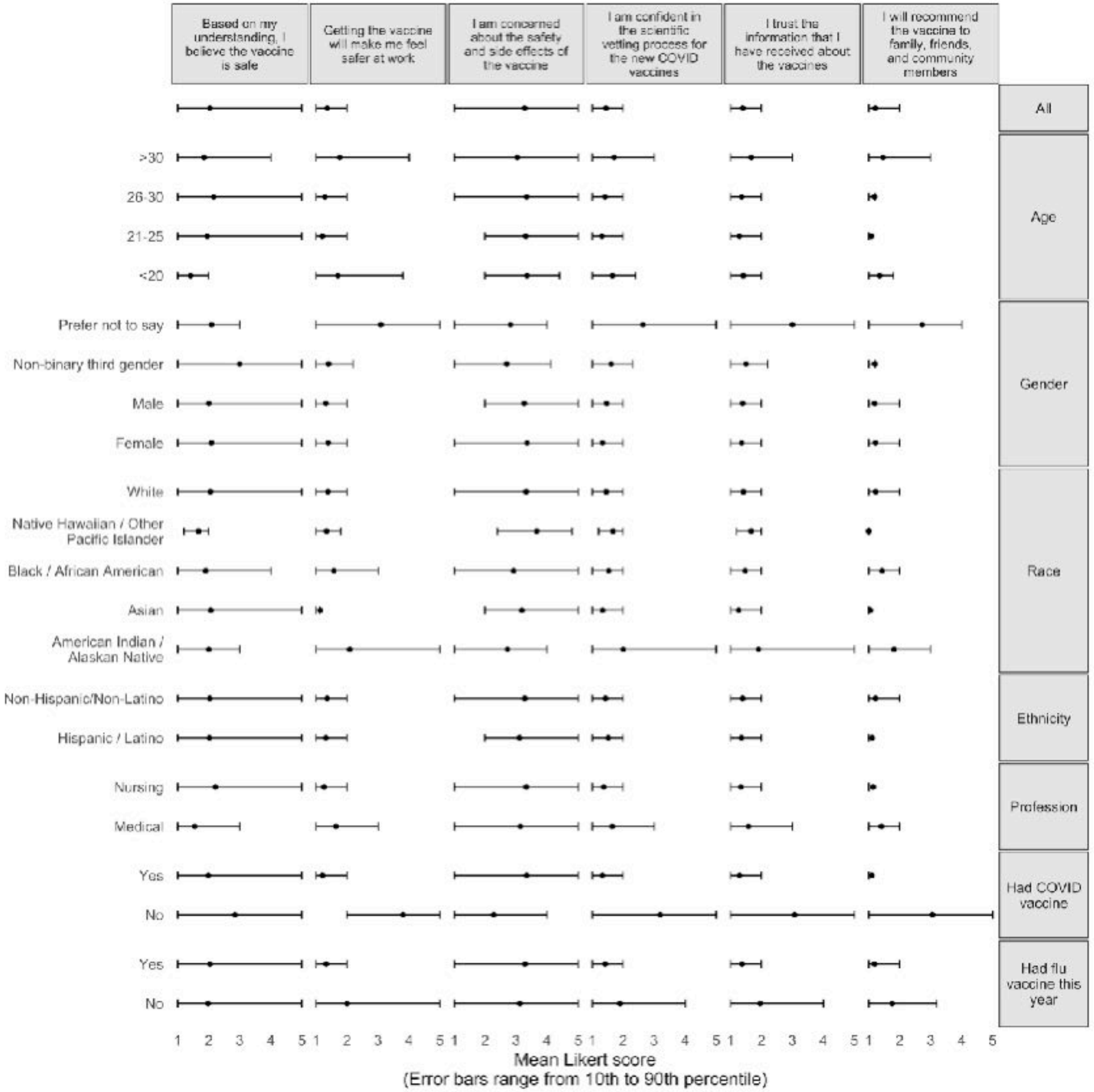
Source: Spring 2021 Global Attitudes Survey. Q36.

"What Makes Life Meaningful? Views From 17 Advanced Economies"

PEW RESEARCH CENTER

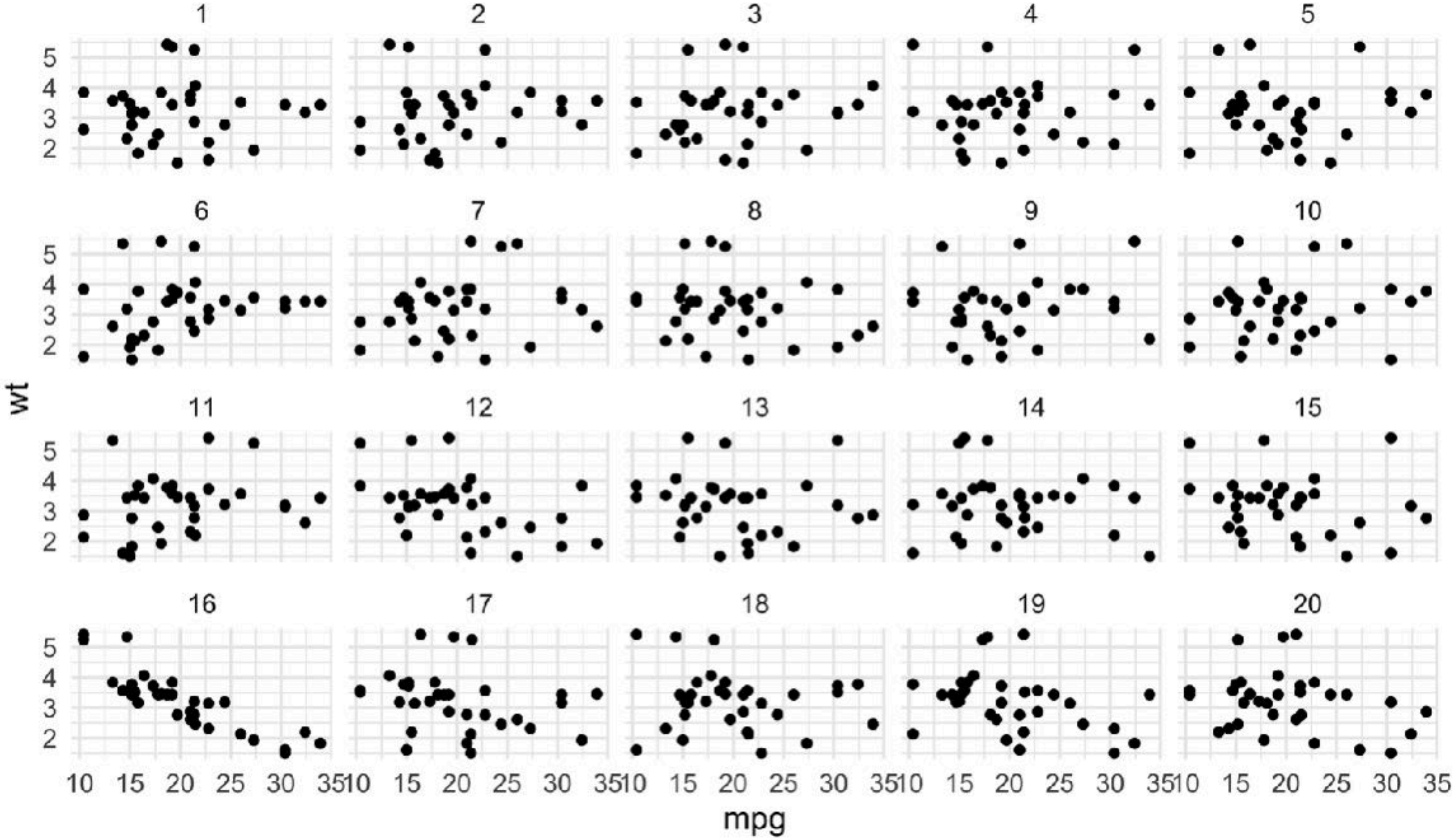
Data visualization later in curriculum

- “Recreate” to advance programming skills
- “Recreate, then improve” to advance programming and communication skills
- “Go beyond the basics” exercises to introduce commonly used visuals in scientific communication



Data visualization for inference

- Take visualizations beyond EDA
- Use them to assess significance, as an alternative method for inference



```
> data
> wrangling
```

Data wrangling

SHAPESHIFTING

- Start with data summarizing, then move on to data reshaping and tidying
- Teach it to
  - motivate **inquiry and exploration**
  - **join** data from multiple sources
  - **preprocess** data for statistical analysis

Data wrangling for summarization

- Start with the basics as early as possible

```r
penguins |>
  count(island, species)
```

```
# A tibble: 5 × 3
  island    species       n
  <fct>     <fct>     <int>
1 Biscoe    Adelie       44
2 Biscoe    Gentoo      124
3 Dream     Adelie       56
4 Dream     Chinstrap    68
5 Torgersen Adelie       52
```

Data wrangling for summarization

- Start with the basics as early as possible

- Wrangle further for better presentation

```
penguins |>
  count(island, species) |>
  pivot_wider(names_from = species, values_from = n,
              values_fill = 0)
```

```
# A tibble: 3 × 4
  island    Adelie Gentoo Chinstrap
  <fct>      <int>  <int>     <int>
1 Biscoe        44    124         0
2 Dream         56      0        68
3 Torgersen     52      0         0
```

Data wrangling for data tidying

- Introduce more advanced data wrangling tools for joining multiple datasets into a single tidy dataset

Data wrangling for data tidying

- Introduce more advanced data wrangling tools for joining multiple datasets into a single tidy dataset

- Reshape data that comes in non-tidy format into a tidy format

```
## [
##   {
##     "gender": ["Female"],
##     "first_name": ["Kimberly"],
##     "last_name": ["Beckstead"],
##     "age": [24],
##     "phone_number": ["216-555-2549"],
##     "purchases": [
##       {
##         "SetID": [24701],
##         "Number": ["76062"],
##         "Theme": ["DC Comics Super Heroes"],
##         "Subtheme": ["Mighty Micros"],
##         "Year": [2016],
##         "Name": ["Robin vs. Bane"],
##         "Pieces": [77],
##         "USPrice": [9.99],
##         "ImageURL": ["http://images.brickset.com/sets/images/
76062-1.jpg"],
##         "Quantity": [1]
##       }
##     ]
##   }
## ]
```

```
> data
> import
```

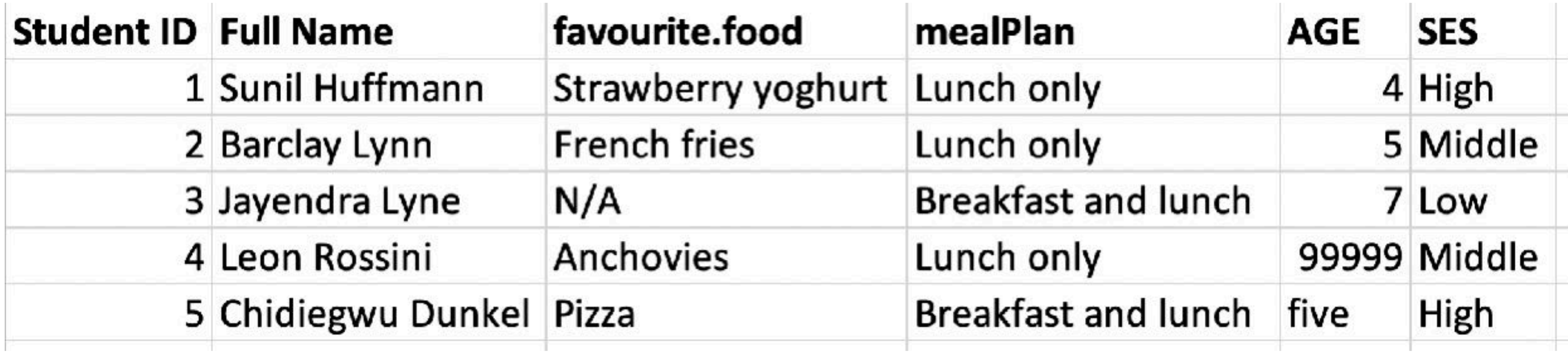
Data import

SHAPESHIFTING

- Think beyond the CSV!
- Teach it to
  - motivate discussion on **data types**
  - create an opportunity to **harvest web data**

Data types

- Discussion of data types and classes can feel dry without the right motivation
- Having to deal with unexpected data types after importing data is a very common task, hence a good motivation for this topic



Student ID	Full Name	favourite.food	mealPlan	AGE	SES
1	Sunil Huffmann	Strawberry yoghurt	Lunch only	4	High
2	Barclay Lynn	French fries	Lunch only	5	Middle
3	Jayendra Lyne	N/A	Breakfast and lunch	7	Low
4	Leon Rossini	Anchovies	Lunch only	99999	Middle
5	Chidiegwu Dunkel	Pizza	Breakfast and lunch	five	High

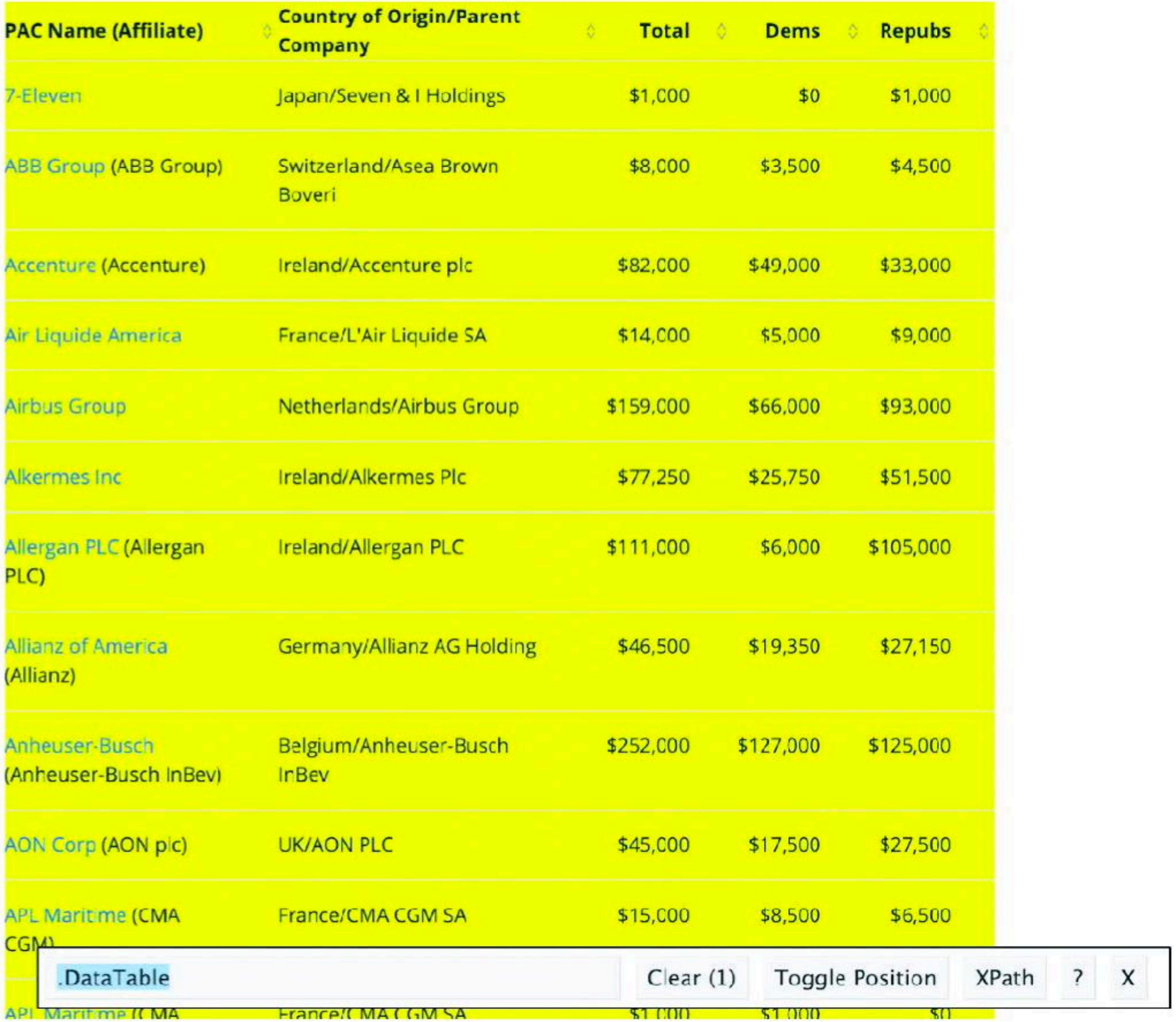
```r
fav_food <- read_excel("data/favourite-food.xlsx")

fav_food
```

```
## # A tibble: 5 x 6
##   `Student ID` `Full Name`  favourite.food  mealPlan  AGE   SES  
##          <dbl> <chr>        <chr>           <chr>     <chr> <chr>
## 1            1 Sunil Huffm… Strawberry yog… Lunch on… 4     High 
## 2            2 Barclay Lynn French fries    Lunch on… 5     Midd…
## 3            3 Jayendra Ly… N/A             Breakfas… 7     Low  
## 4            4 Leon Rossini Anchovies       Lunch on… 99999 Midd…
## 5            5 Chidiegwu D… Pizza           Breakfas… five  High 
```

Web data

- The web is an incredible source for data, but turning it into a structured format (without copy-paste or manual entry) requires learning **web scraping** skills
- Beyond screen scraping, it's useful to introduce the idea of **getting data from an API** at some point in the curriculum
- Both of these offer an opportunity for discussion on **ethics and data privacy**



PAC Name (Affiliate)	Country of Origin/Parent Company	Total	Dems	Repubs
7-Eleven	Japan/Seven & I Holdings	$1,000	$0	$1,000
ABB Group (ABB Group)	Switzerland/Asea Brown Boveri	$8,000	$3,500	$4,500
Accenture (Accenture)	Ireland/Accenture plc	$82,000	$49,000	$33,000
Air Liquide America	France/L'Air Liquide SA	$14,000	$5,000	$9,000
Airbus Group	Netherlands/Airbus Group	$159,000	$66,000	$93,000
Alkermes Inc	Ireland/Alkermes Plc	$77,250	$25,750	$51,500
Allergan PLC (Allergan PLC)	Ireland/Allergan PLC	$111,000	$6,000	$105,000
Allianz of America (Allianz)	Germany/Allianz AG Holding	$46,500	$19,350	$27,150
Anheuser-Busch (Anheuser-Busch InBev)	Belgium/Anheuser-Busch InBev	$252,000	$127,000	$125,000
AON Corp (AON plc)	UK/AON PLC	$45,000	$17,500	$27,500
APL Maritime (CMA CGM)	France/CMA CGM SA	$15,000	$8,500	$6,500
APL Maritime (CMA	France/CMA CGM SA	$1,000	$1,000	$0

Dogucu, M. & Çetinkaya-Rundel, M. "Web Scraping in the Statistics and Data Science Curriculum: Challenges and Opportunities." Journal of Statistics Education (2021): 1-11. https://doi.org/10.1080/10691898.2020.1787116.

```
> predictive
> modeling
```

Predictive modeling

CLAIRVOYANCE

- Don’t just leave it to the machine learning course, introduce it along with explanatory / inferential models
- Teach it to
  - introduce the idea of **overfitting** and mitigating it with splitting the data into testing and training sets
  - allow for **creativity** with feature engineering
  - discuss **bias-variance tradeoff** early on
  - enable those **open-ended projects** for classifying binary outcome variables

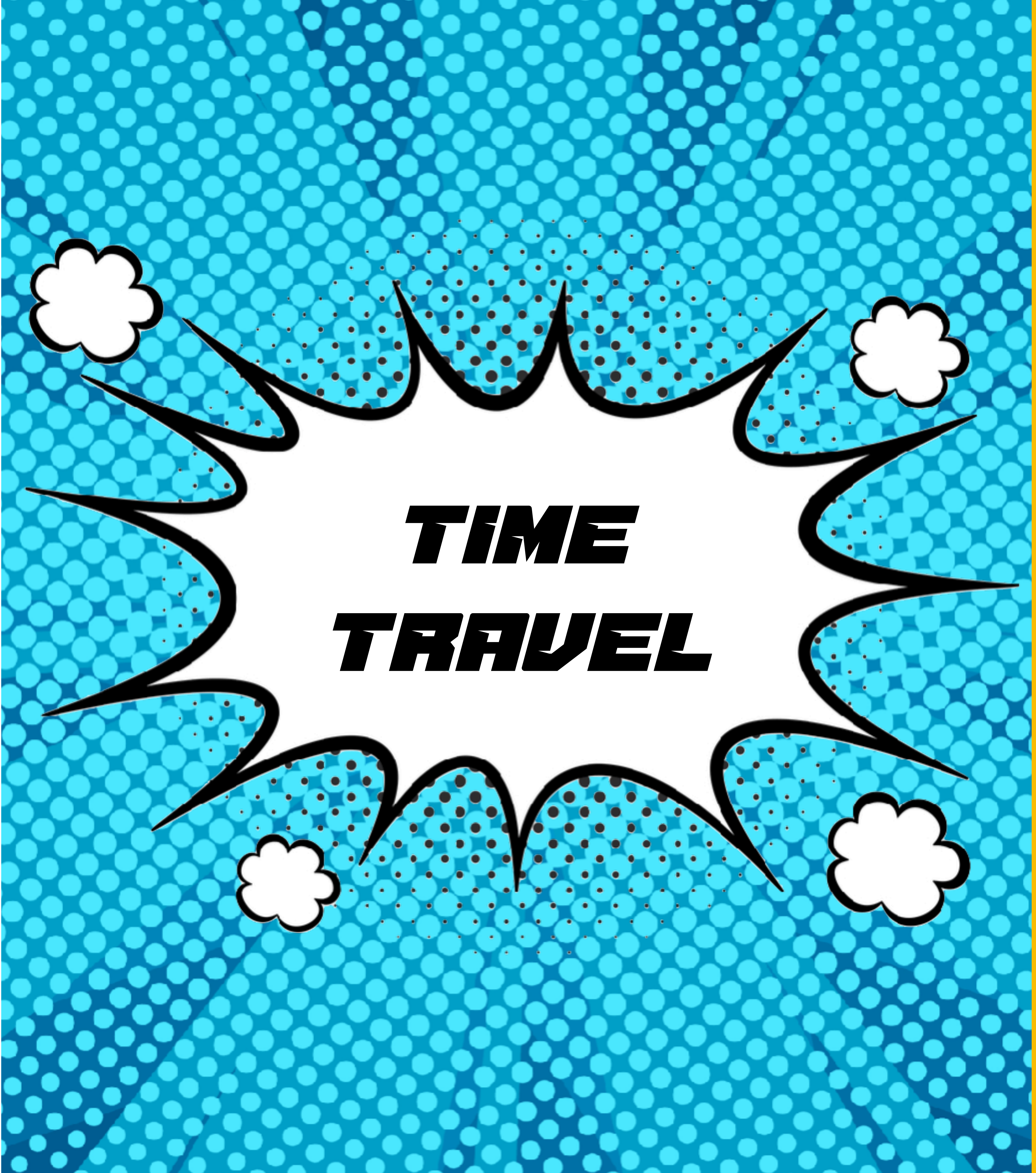
Predictive (tidy) models

- The **tidymodels** framework is a collection of packages for modeling and machine learning using tidyverse principles

- Tidymodels pipelines start with an `initial_split()` into training and testing data and the tooling provides **guard rails** to prevent prediction on the testing data at the model and feature development phase

- Functions designed specifically for **feature engineering** motivate creative thinking during model development

- eCOTS 2022 breakout session *Modernizing the undergraduate regression analysis course* — bit.ly/modern-regression



```
> version
> control
```

Version control

TIME TRAVEL

- Teach it as early as possible and as needed, but when you can make time in your curriculum and integrate it throughout the curriculum

- Teach it to

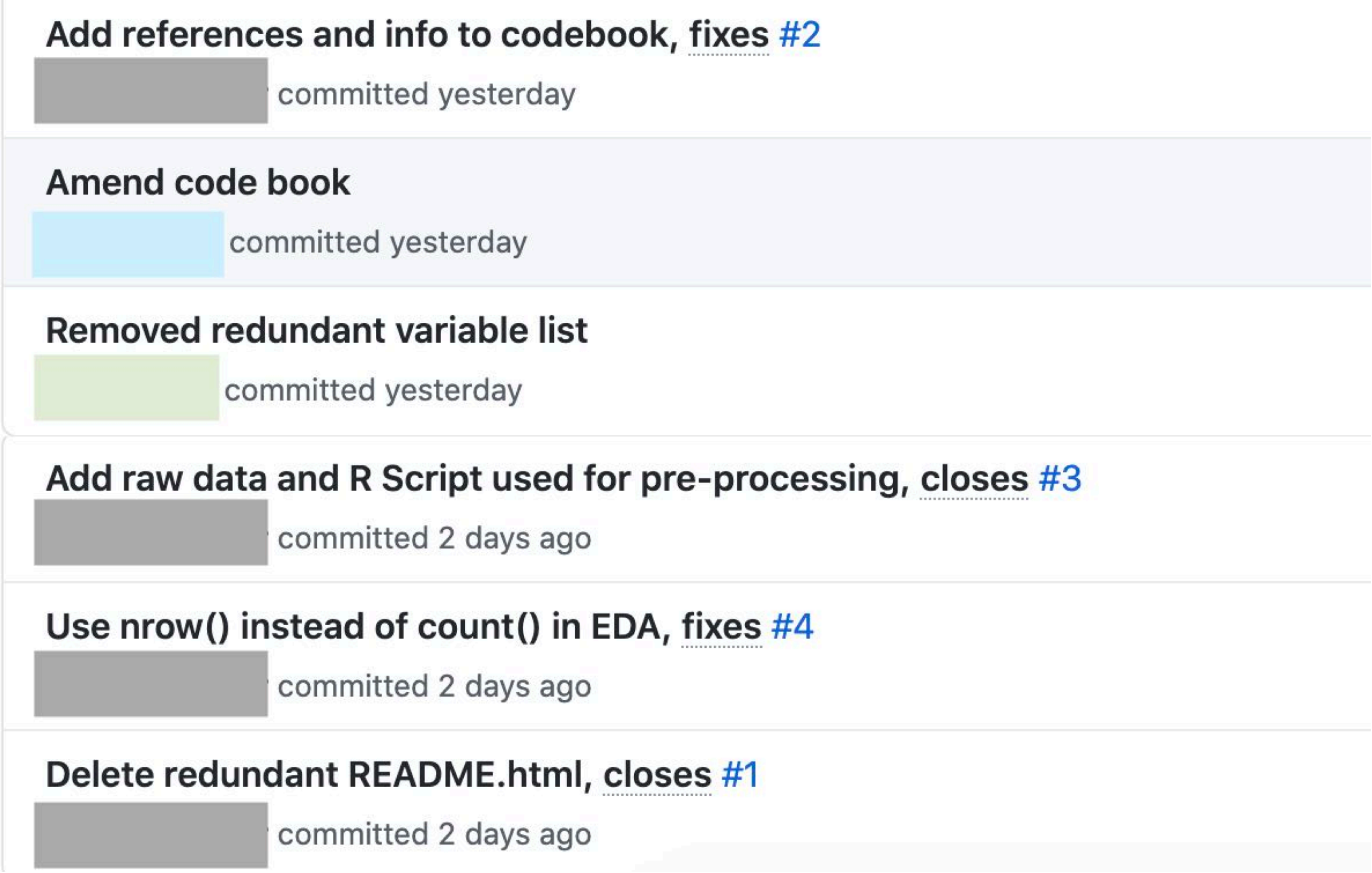
  - build **good habits** when the stakes are low

  - motivate not just reproducibility but also **collaboration**

  - instill practice of **open sharing** and start curating an **online portfolio**

Beckman, Matthew D., et al. "Implementing version control with Git and GitHub as a learning objective in statistics and data science courses." Journal of Statistics and Data Science Education 29, no. sup1 (2021): S132-S144. https://doi.org/10.1080/10691898.2020.1848485.

Reproducibility and collaboration



Add references and info to codebook, fixes #2
[redacted] committed yesterday

Amend code book
[redacted] committed yesterday

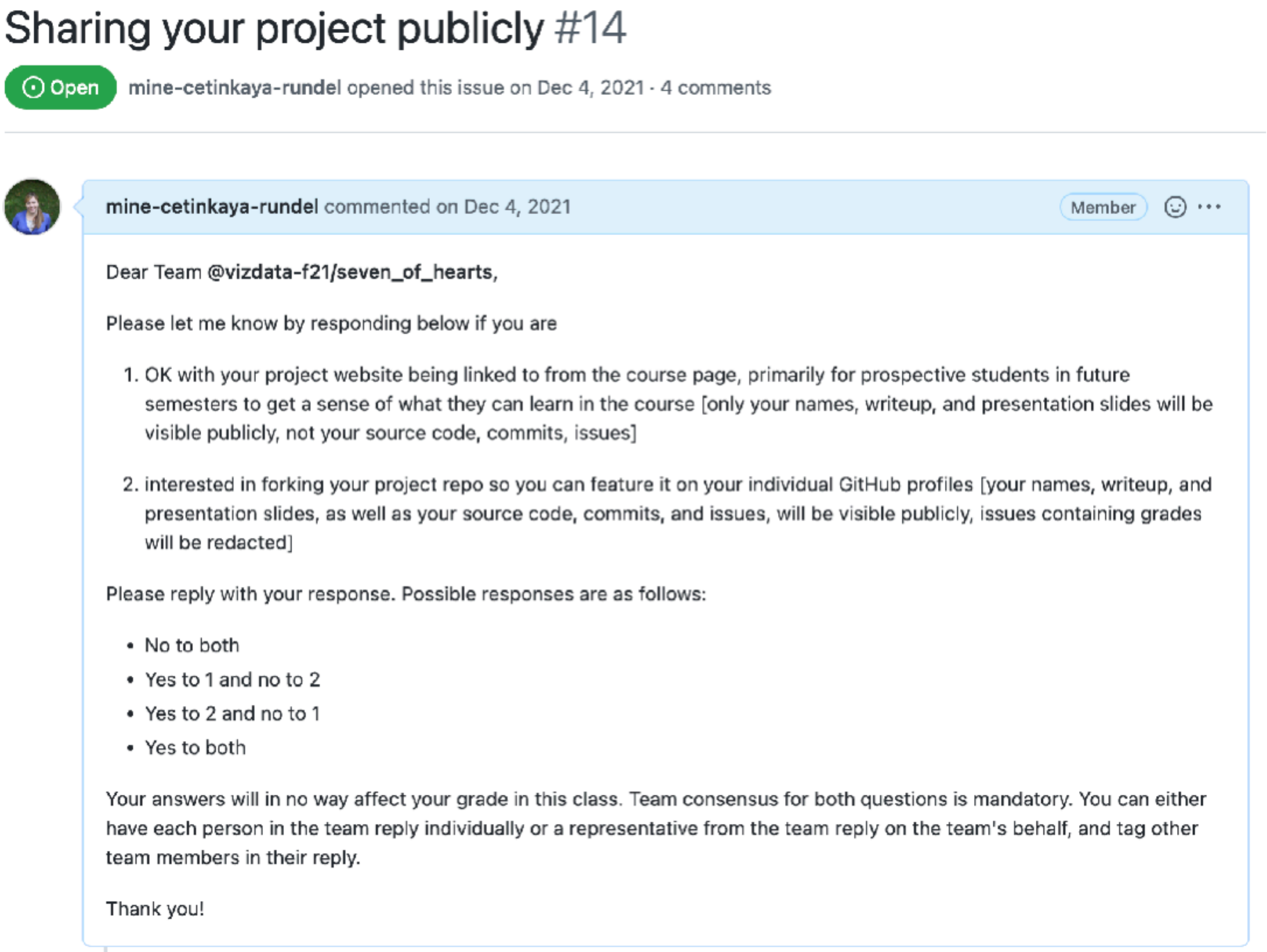
Removed redundant variable list
[redacted] committed yesterday

Add raw data and R Script used for pre-processing, closes #3
[redacted] committed 2 days ago

Use nrow() instead of count() in EDA, fixes #4
[redacted] committed 2 days ago

Delete redundant README.html, closes #1
[redacted] committed 2 days ago

Web hosting to online portfolio



Sharing your project publicly #14

mine-cetinkaya-rundel opened this issue on Dec 4, 2021 · 4 comments

mine-cetinkaya-rundel commented on Dec 4, 2021

Dear Team @vizdata-f21/seven_of_hearts,

Please let me know by responding below if you are

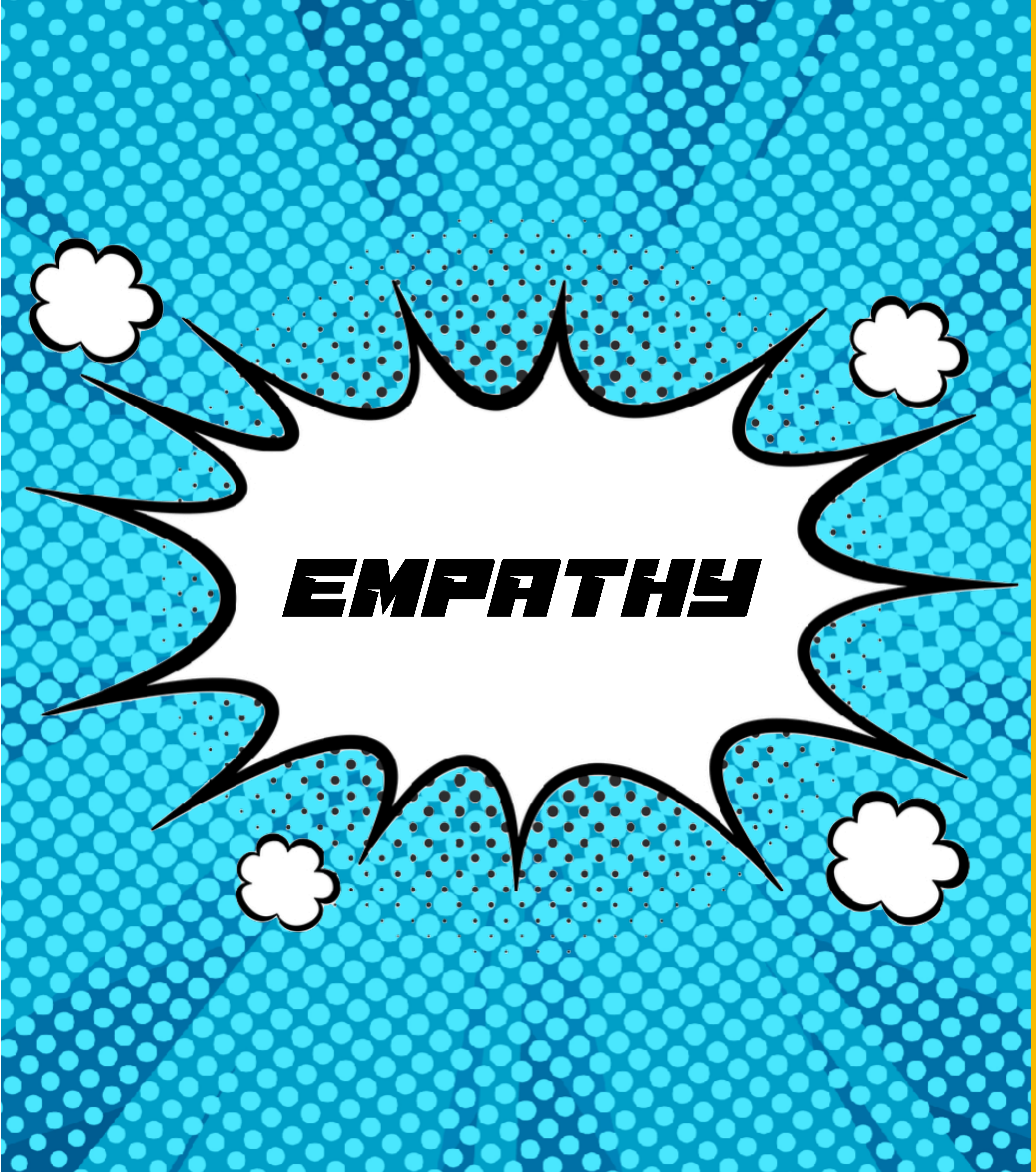
1. OK with your project website being linked to from the course page, primarily for prospective students in future semesters to get a sense of what they can learn in the course [only your names, writeup, and presentation slides will be visible publicly, not your source code, commits, issues]
2. interested in forking your project repo so you can feature it on your individual GitHub profiles [your names, writeup, and presentation slides, as well as your source code, commits, and issues, will be visible publicly, issues containing grades will be redacted]

Please reply with your response. Possible responses are as follows:

- No to both
- Yes to 1 and no to 2
- Yes to 2 and no to 1
- Yes to both

Your answers will in no way affect your grade in this class. Team consensus for both questions is mandatory. You can either have each person in the team reply individually or a representative from the team reply on the team's behalf, and tag other team members in their reply.

Thank you!



```
> empathy
```

Empathy

EMPATHY

- Strive to introduce the **story** with the dataset

- Couple each dataset with a **datasheet**:
  - For what purpose was the dataset created?
  - Does the dataset contain data that might be considered confidential (for example, data that is protected by legal privilege or by doctor–patient confidentiality, data that includes the content of individuals’ non-public communications)?
  - Is it possible to identify individuals (that is, one or more natural persons), either directly or indirectly (that is, in combination with other data) from the dataset?
  - Were the individuals in question notified about the data collection?
  - ...

- Use this practice to motivate discussion around wider data science **ethics** issues like algorithmic bias, privacy and re-identification, etc.

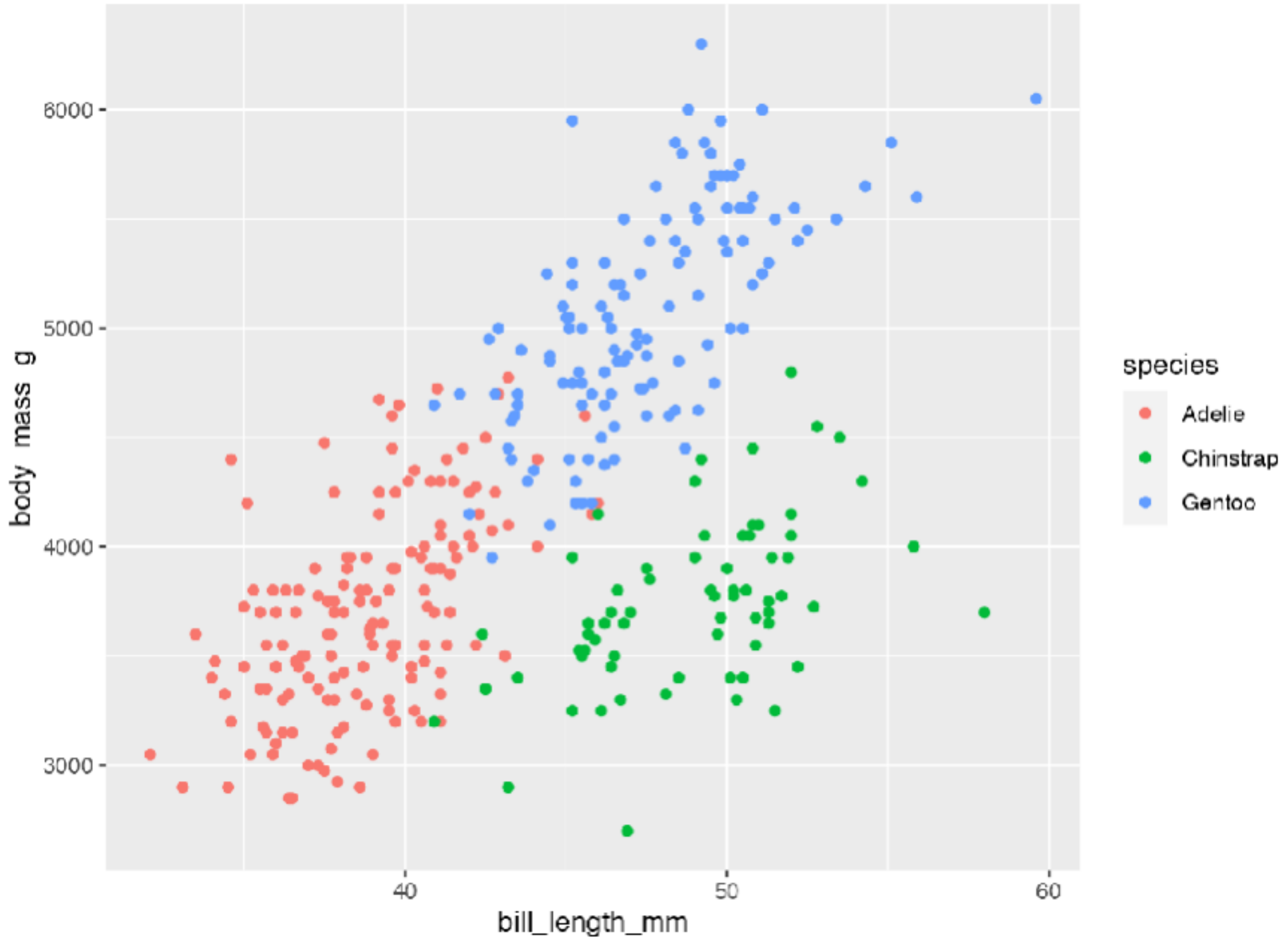
Gebru, Timnit, et al. "Datasheets for datasets." Communications of the ACM 64.12 (2021): 86-92. DOI: http://dx.doi.org/10.1145/3458723.

Accessibility

- You could teach a whole course or even a whole curriculum on accessibility…
- At a minimum, your students shouldn’t graduate without ever thinking / learning about it!
- Tooling exists to accomplish the bare minimum and that can go a long way in raising the next generation of data scientists who consider accessibility in their work

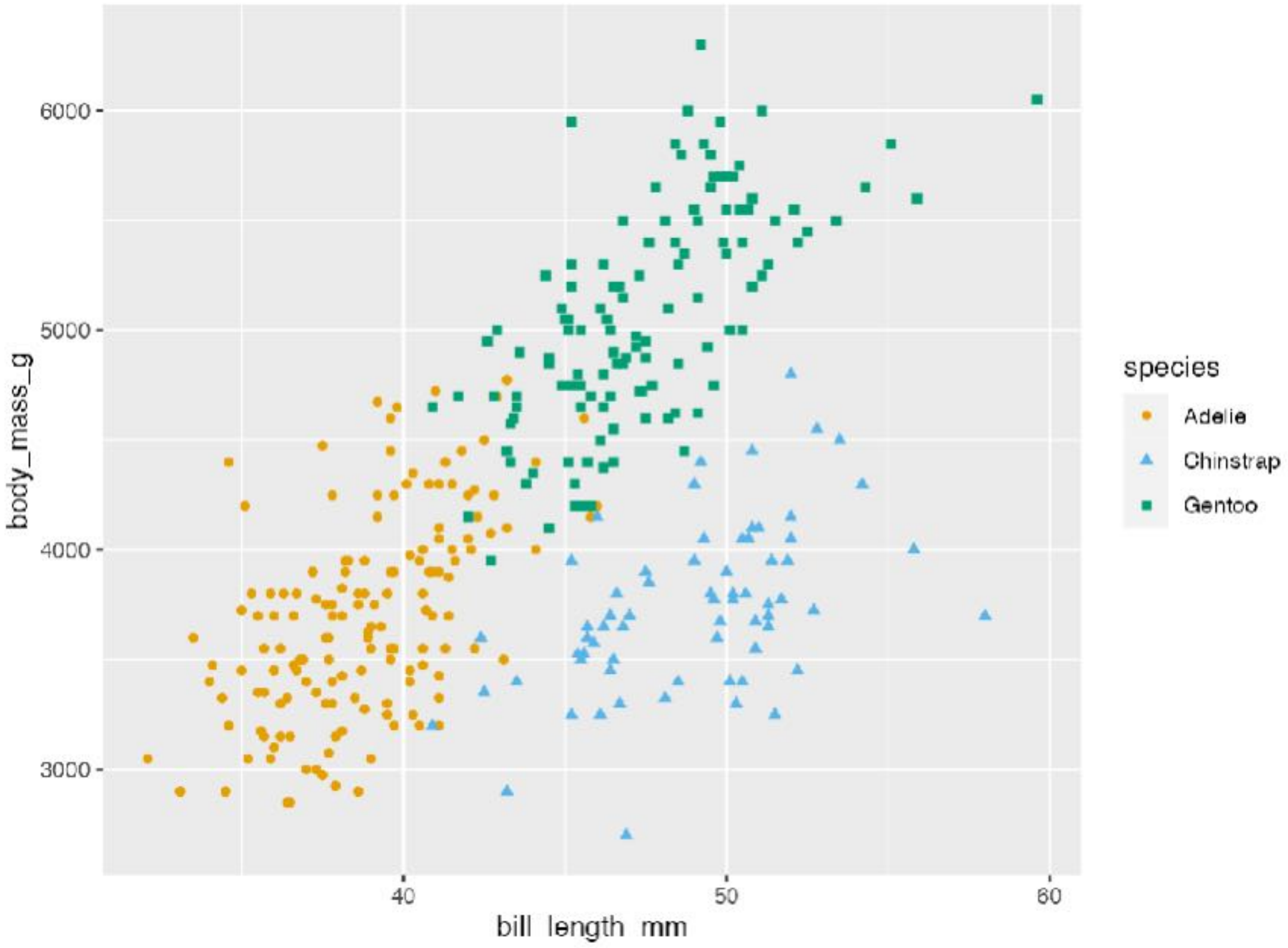
~~~
```{r}
#| fig-cap: Body mass vs. bill length of penguins.

ggplot(penguins,
       aes(x = bill_length_mm, y = body_mass_g,
           color = species)) +
  geom_point()
```
~~~



~~~~
```{r}
#| fig-cap: Body mass vs. bill length of penguins.
#| fig-alt: >
#|   A scatterplot showing positive, relatively strong
#|   relationship between body mass and bill length. The
#|   points representing each of the three species are
#|   clustered with Adelies with lowest typical bill length
#|   and body mass, Chinstraps with higher typical bill
#|   length and similar body mass, and Gentoos with typical
#|   bill length between the other two but higher typical
#|   body mass.

ggplot(penguins,
       aes(x = bill_length_mm, y = body_mass_g,
           color = species, shape = species)) +
  geom_point() +
  colorblindr::scale_color_OkabeIto()
```
~~~~



SELF-SUFFICIENCY

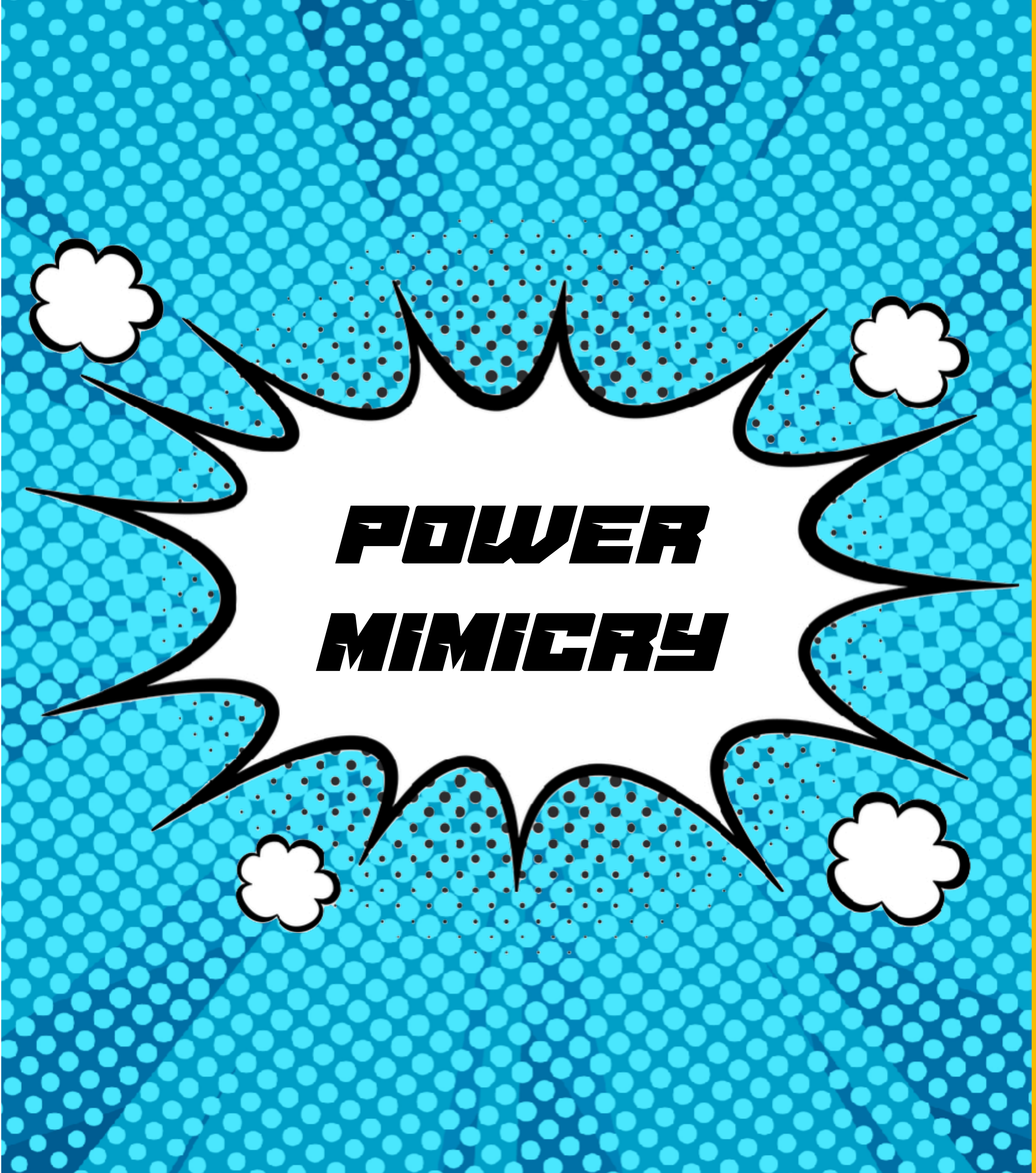
> learning
> on one’s own

Learning on one’s own

SELF SUFFICIENCY

- Share with students
  - **how** you learn, and be specific: books, blog posts, Twitter accounts you follow, etc.
  - how you choose **what** to learn
- Demonstrate how you solve problems — e.g., via live coding
- Encourage them to take active part in the **community**

AND A FEW SUPERPOWERS FOR THE EDUCATORS...



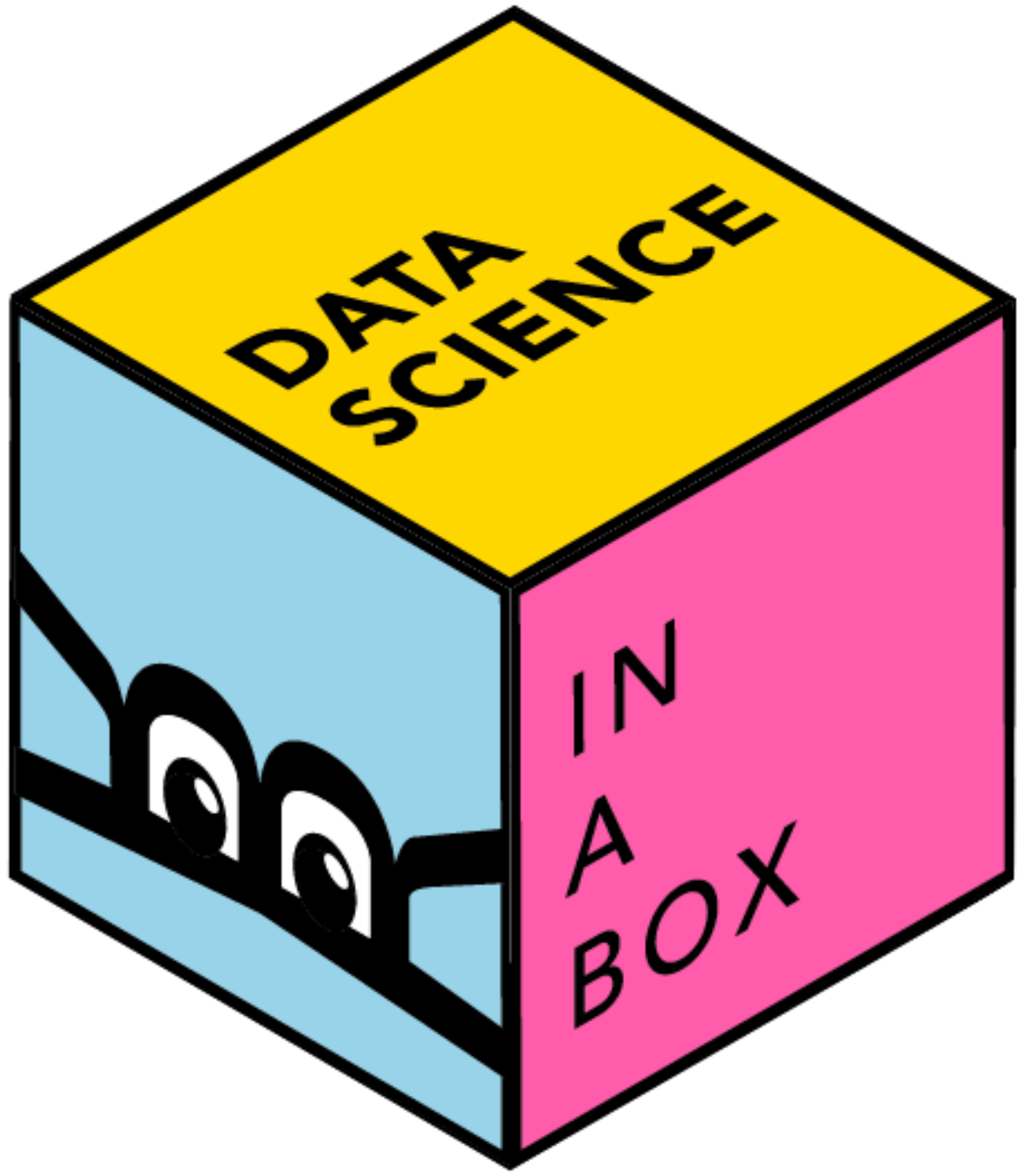
POWER MIMICRY

```
> leveraging
> open resources
```

Leveraging open resources

POWER MIMICRY

Introductory data science



datasciencebox.org

Stat 2 / Regression

sta210-s22.github.io/website

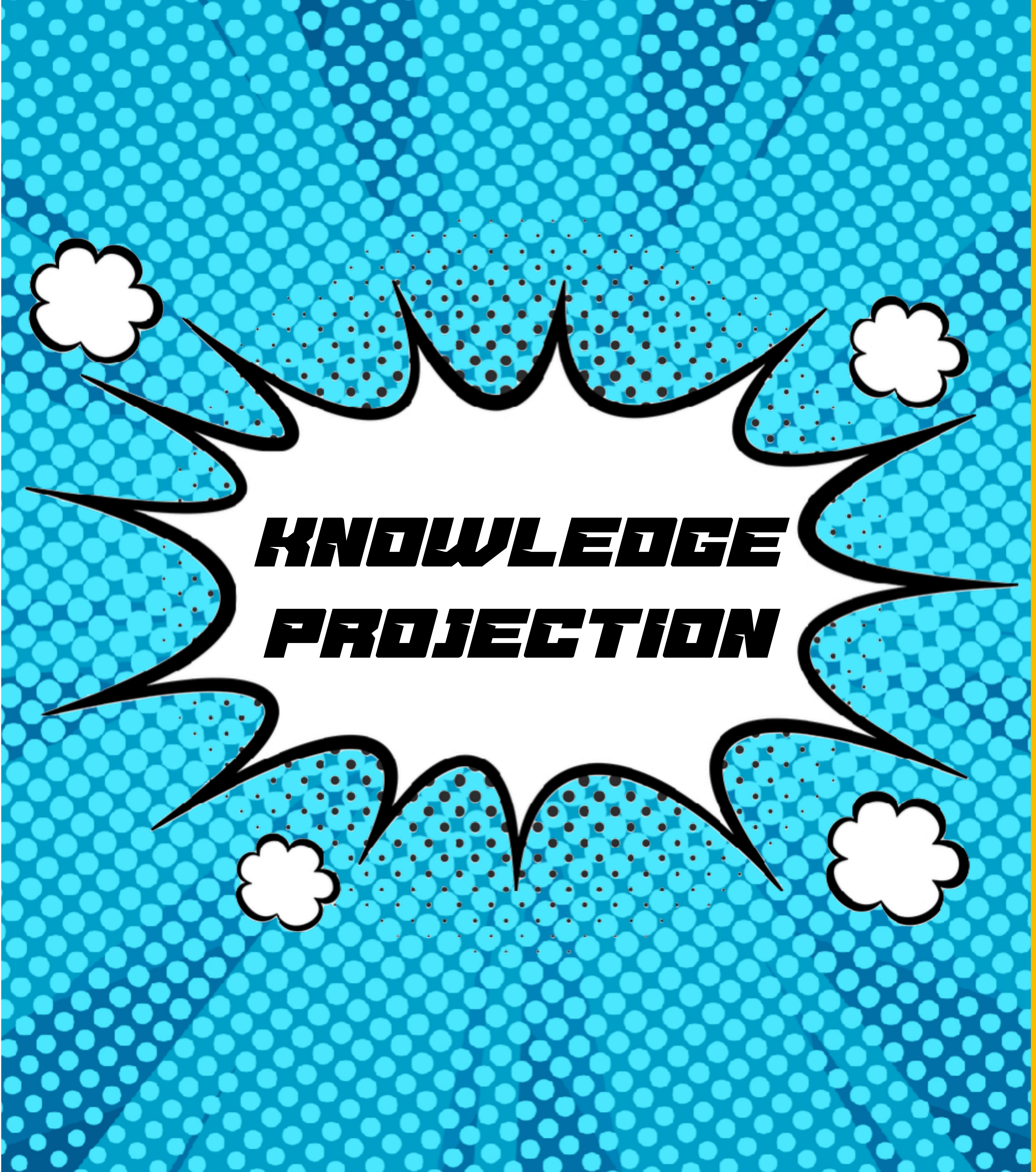
Data visualization

vizdata.org

CALL TO ACTION

In the chat, share a open educational resource you’ve created or reused. Please don’t be shy!

Image by DONT SELL MY ARTWORK AS IS Pixabay.



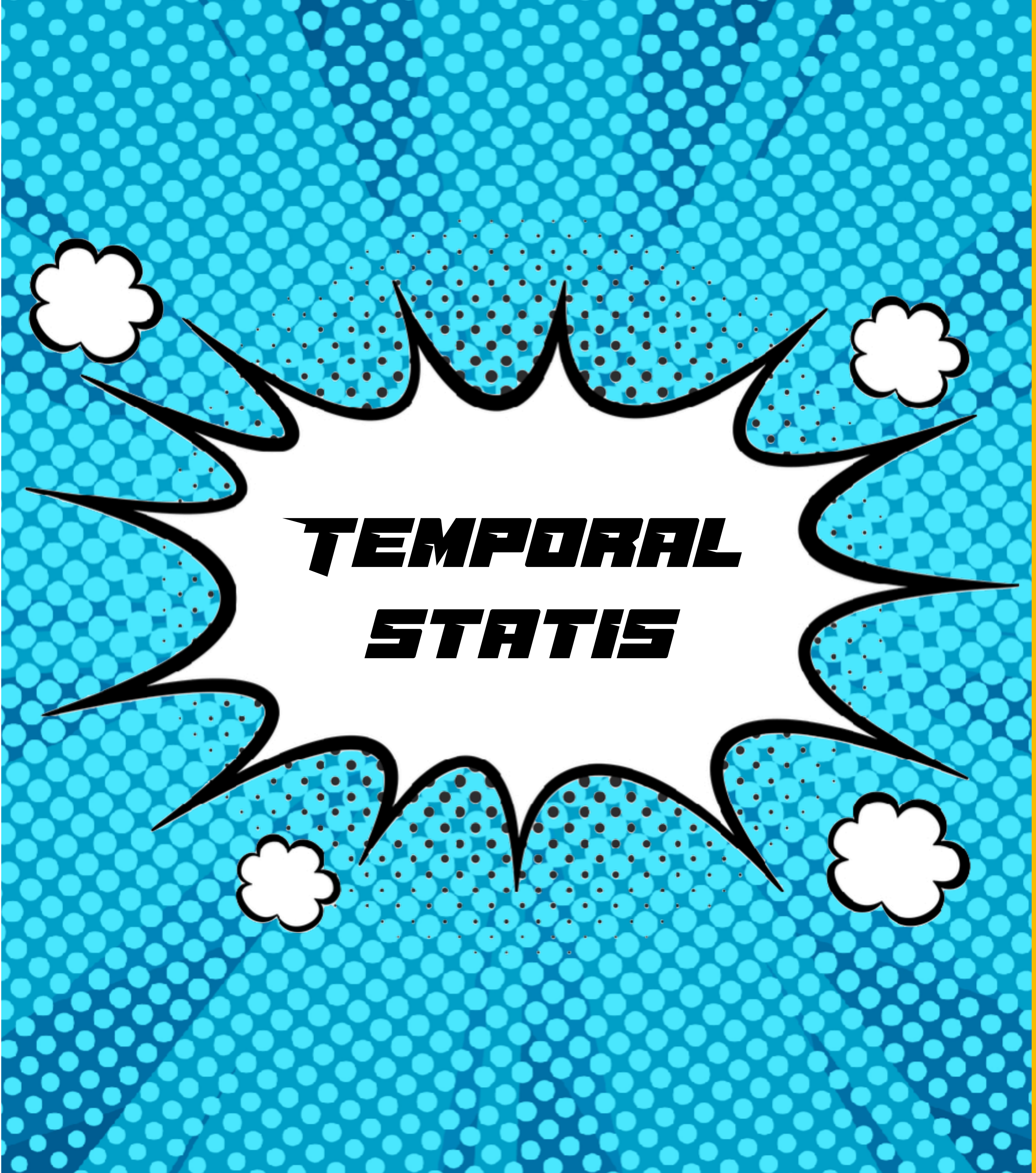
KNOWLEDGE PROJECTION

> sharing knowledge
> with others

Sharing with others

KNOWLEDGE PROJECTION

- Open-source your course materials
- Write about your experiences
  - Blog posts
  - Journal articles - not just for empirical studies but also reflective essays, datasets and stories, brief communications, etc.



TEMPORAL STATIS

```
> making time
> to keep current
```

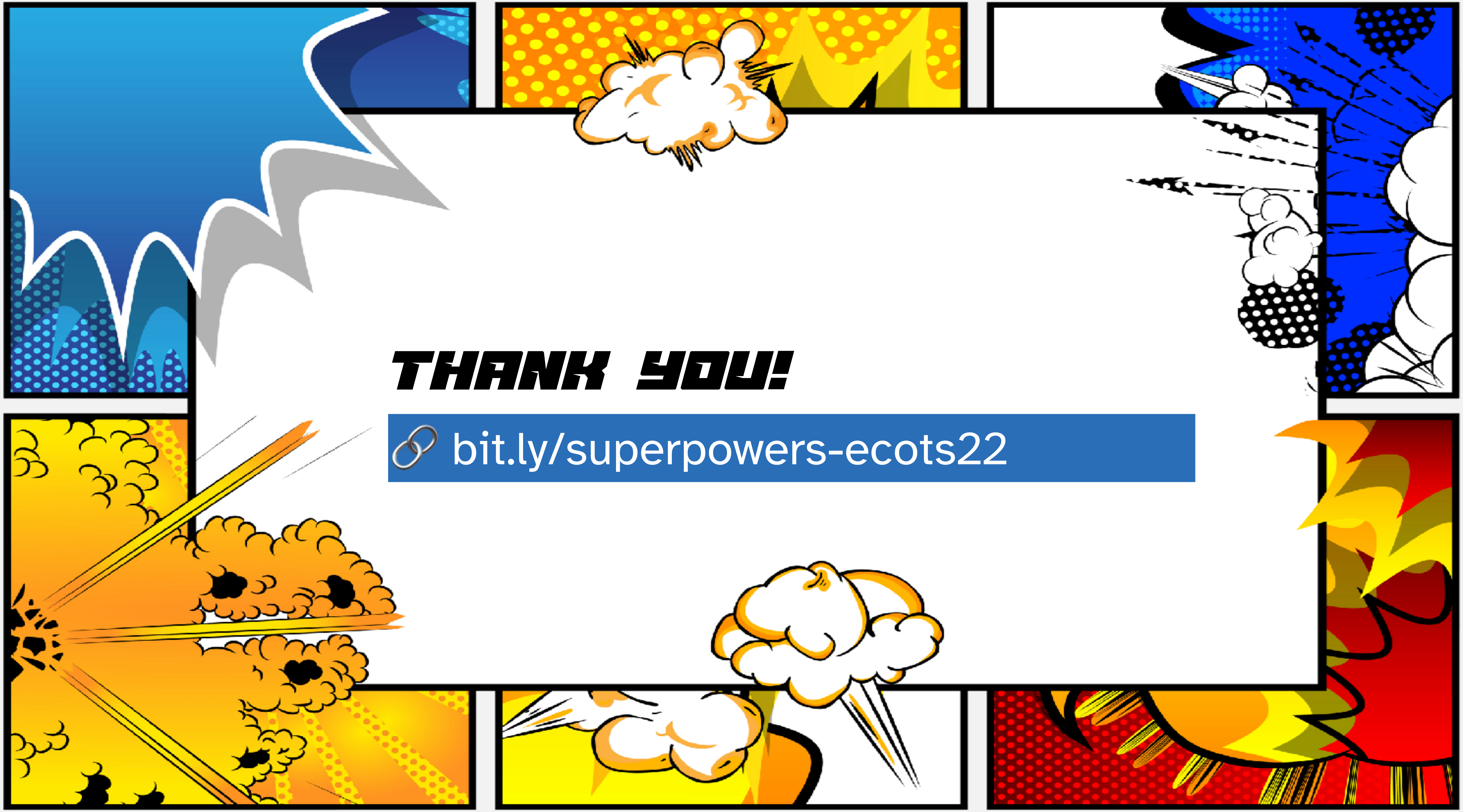
Making time to keep current

TEMPORAL STATIS

- Probably impossible, but you can try 😜
- A few things I'm learning / playing with nowadays to keep current:
  - Transitioning to the **native R pipe** |>
    - Recommended reading: Blog post by Isabella Velásquez
  - **Quarto**: Open-source scientific and technical multi-lingual publishing system, aka next generation R Markdown that supports multiple programming languages
    - Recommended reading: Get Started tutorials at quarto.org
  - **Databases** / SQL 😬
  - The wealth of **resources from eCOTS 2022**, particularly those on Diversity, Inclusion and Social Justice in data science!

NORMALIZE BEING HUMAN ❤️

- You don’t have to learn everything / you don’t have to teach everything
- Incremental changes over time more than fine!
- New “things” (features, packages, tools) being discussed / hyped in the community can be a good indication of their importance but doesn’t mean you have to adopt them right away



THANK YOU!

🔗 bit.ly/superpowers-ecots22

References

- Gebru, Timnit, et al. "Datasheets for datasets." Communications of the ACM 64.12 (2021): 86-92. DOI: http://dx.doi.org/10.1145/3458723.
- Çetinkaya-Rundel et al. “An educator’s perspective of the tidyverse.” Technology Innovations in Statistics Education (2022): 14(1). http://dx.doi.org/10.5070/T514154352.
- Dogucu, M. & Çetinkaya-Rundel, M. “Web Scraping in the Statistics and Data Science Curriculum: Challenges and Opportunities.” Journal of Statistics Education (2021): 1-11. https://doi.org/10.1080/10691898.2020.1787116.
- Beckman, Matthew D., et al. "Implementing version control with Git and GitHub as a learning objective in statistics and data science courses." Journal of Statistics and Data Science Education 29, no. sup1 (2021): S132-S144. https://doi.org/10.1080/10691898.2020.1848485.