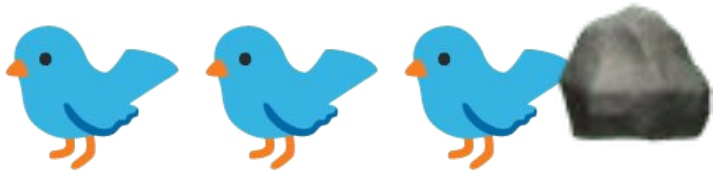
# Three birds, one stone:

## *Advancing data literacy through RDM and Carpentry instruction integration*

UC Santa Barbara



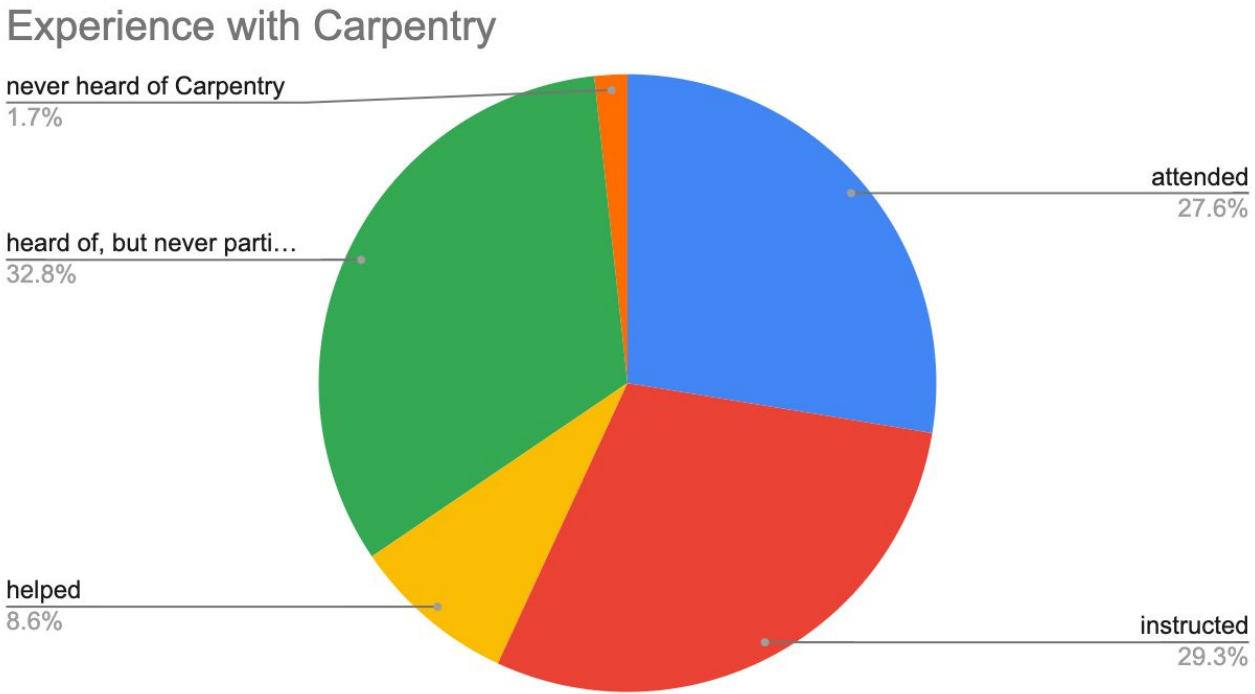
# Outline for the afternoon

- Introduction (0:35)
  - Background and motivation
  - Some concrete examples
  - Data literacy and RDM frameworks
- Breakout 1 (0:45)
  - What data literacy/RDM topics should we be teaching?
  - Report-backs
- BREAK (0:15)
- Breakout 2 (0:45)
  - Where and how can these topics be taught?
  - Report-backs
- General discussion, next steps (0:20)

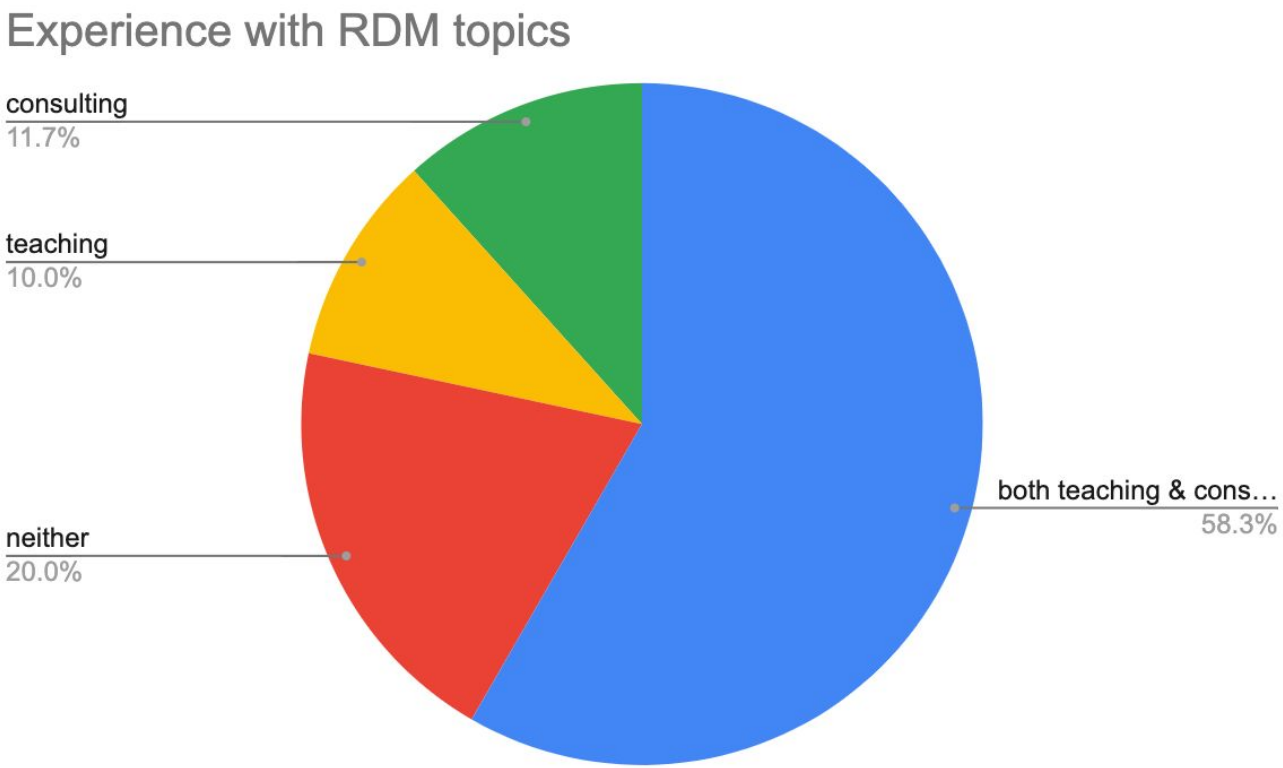
# Our background

- Past experience
  - Many Carpentry workshops
    - Heavily reorganized, adapted existing curriculum
    - Developed new modules, lessons
  - RDM workshops
    - Graduate courses, new faculty
- In the hopper
  - Campus data science initiatives
    - Focused on advanced algorithms, statistics
    - Looking for introductory, student support
  - Library undergraduate course
    - Formerly: How to use the Library
    - Being rethought as: How to be a scholar

# Your background



# Your background



# So what's the problem?

We're worried we're not covering essential topics

- Essential concepts, skills for future scientists
- Essential for correctly reasoning, working with data
- Essential for effective data management

Students request these topics, too

But topics are insufficiently addressed

# Topic 1: Accuracy and precision

What students see as they work with data:

± .0001 ?!

Filter

| sex | hindfoot_length | weight | genus | species | taxa | plot_type | weight_kg | weight_lb |
|---|---|---|---|---|---|---|---|---|
| M | 32 | 197 | Neotoma | albigula | Rodent | Control | 0.197 | 0.4334 |
| NA | NA | NA | Neotoma | albigula | Rodent | Control | NA | NA |
| M | NA | 218 | Neotoma | albigula | Rodent | Control | 0.218 | 0.4796 |
| M | 33 | 166 | Neotoma | albigula | Rodent | Control | 0.166 | 0.3652 |
| M | 32 | 184 | Neotoma | albigula | Rodent | Control | 0.184 | 0.4048 |
| M | 32 | 206 | Neotoma | albigula | Rodent | Control | 0.206 | 0.4532 |
| F | 33 | 274 | Neotoma | albigula | Rodent | Control | 0.274 | 0.6028 |
| F | 30 | 186 | Neotoma | albigula | Rodent | Control | 0.186 | 0.4092 |
| F | 33 | 184 | Neotoma | albigula | Rodent | Control | 0.184 | 0.4048 |
| F | NA | NA | Neotoma | albigula | Rodent | Control | NA | NA |
| F | 31 | 87 | Neotoma | albigula | Rodent | Control | 0.087 | 0.1914 |
| F | 33 | 174 | Neotoma | albigula | Rodent | Control | 0.174 | 0.3828 |
| F | 30 | 130 | Neotoma | albigula | Rodent | Control | 0.130 | 0.2860 |
| M | 34 | 208 | Neotoma | albigula | Rodent | Control | 0.208 | 0.4576 |

Showing 15 to 28 of 34,786 entries

# Topic 1: Accuracy and precision

Important concepts:

- Accuracy
  - Error bounds
  - Significant digits
  - Confidence interval
- Precision
  - Should not exceed accuracy
  - Often implied by number of digits printed
  - Reflects repeatability of measurement

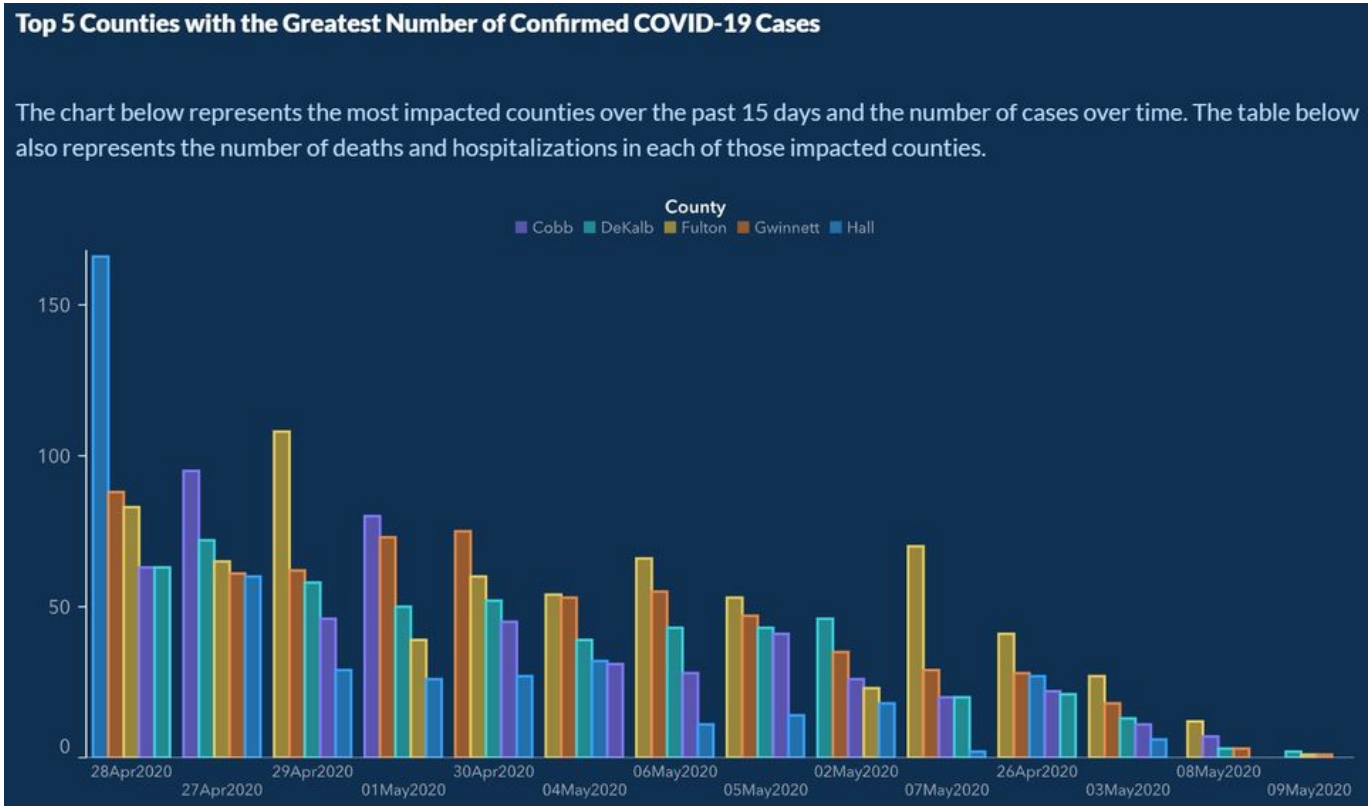
```
    meanGDPperCap <- mean(x$gdpPercap)
    print(paste(
      "The mean GDP per capita for", unique(x$continent),
      "is", format(meanGDPperCap, big.mark=",")
    ))
  }
)
```

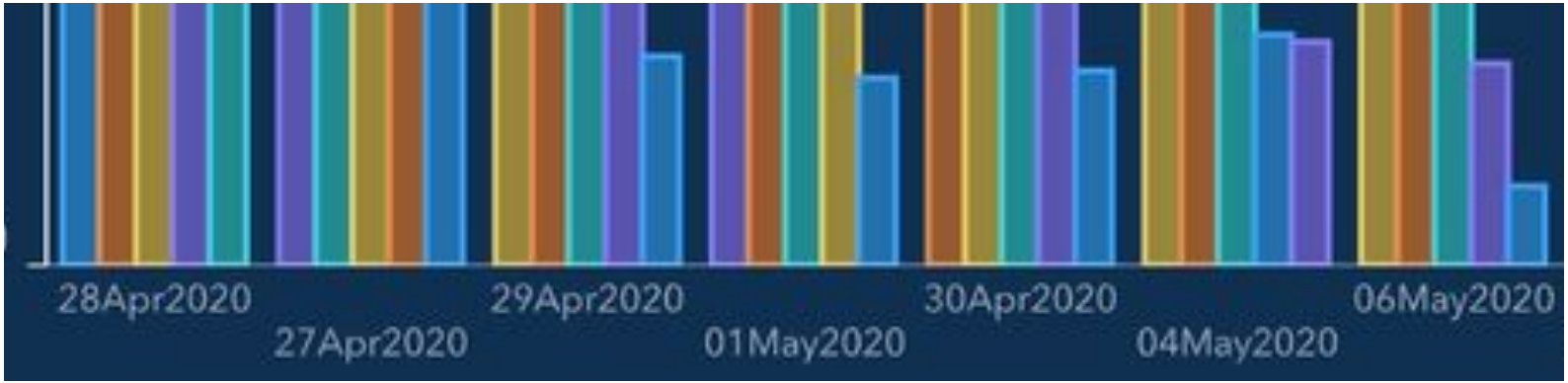
**Output**

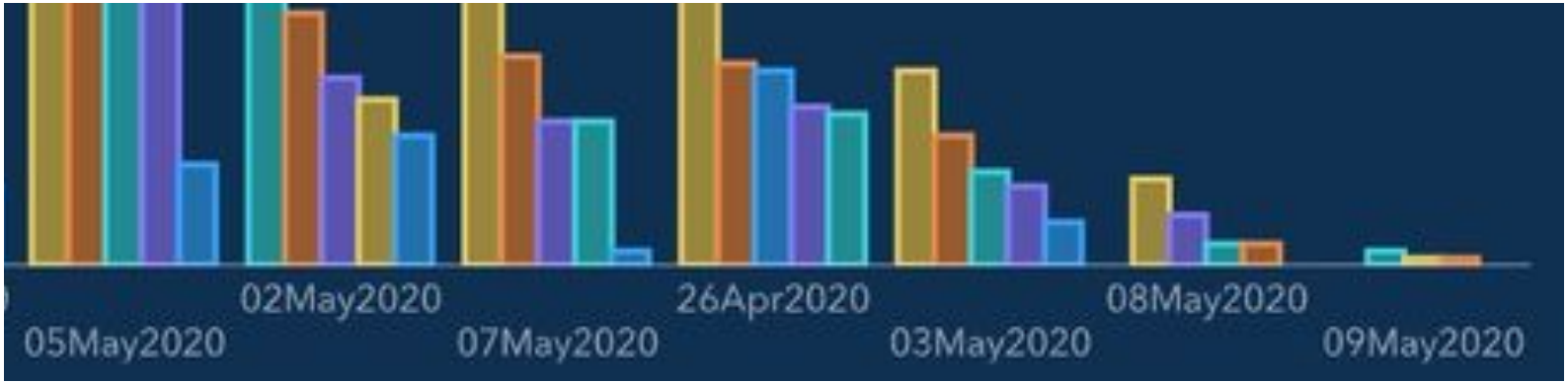
```
[1] "The mean GDP per capita for Africa is 2,193.755"
[1] "The mean GDP per capita for Americas is 7,136.11"
[1] "The mean GDP per capita for Asia is 7,902.15"
[1] "The mean GDP per capita for Europe is 14,469.48"
[1] "The mean GDP per capita for Oceania is 18,621.61"
```

?

# Topic 2: Avoiding unintentionally misleading graphs

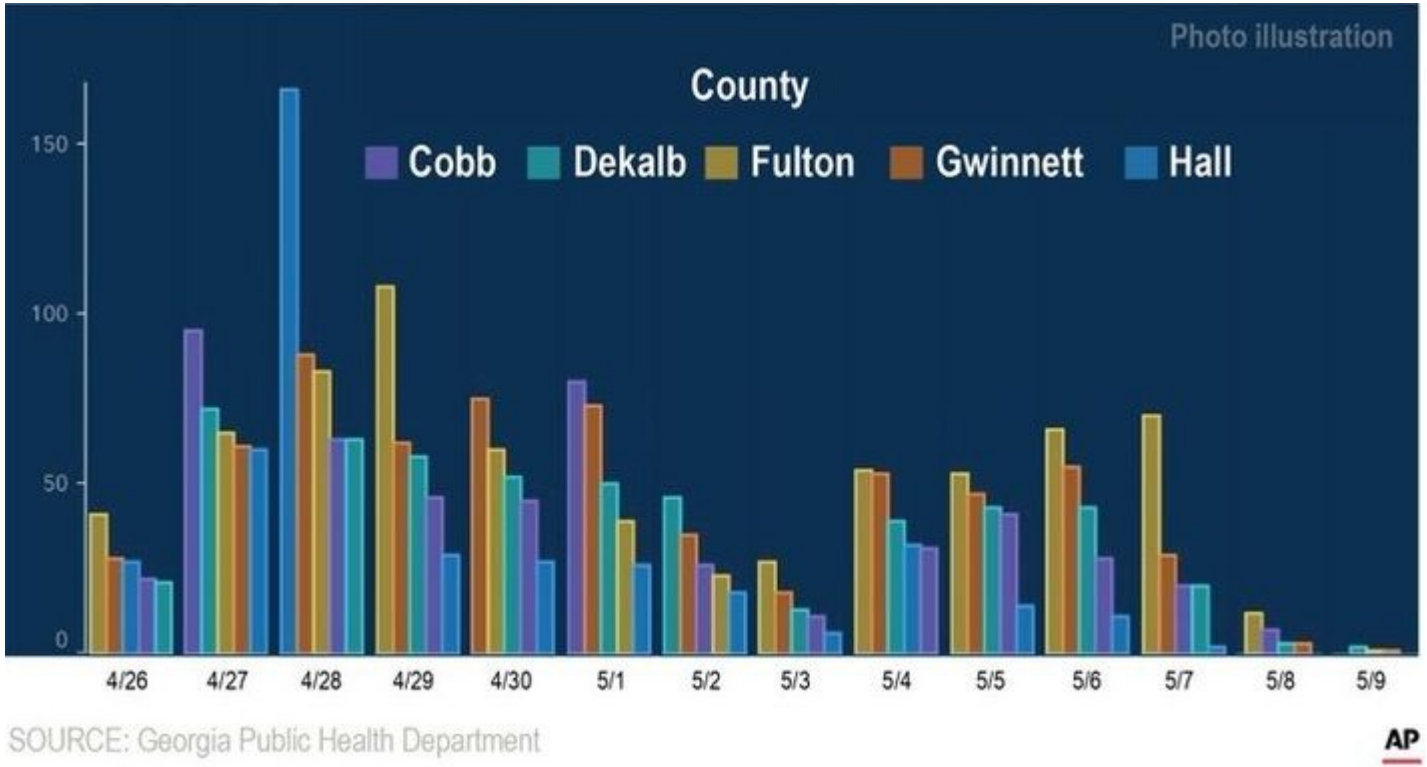




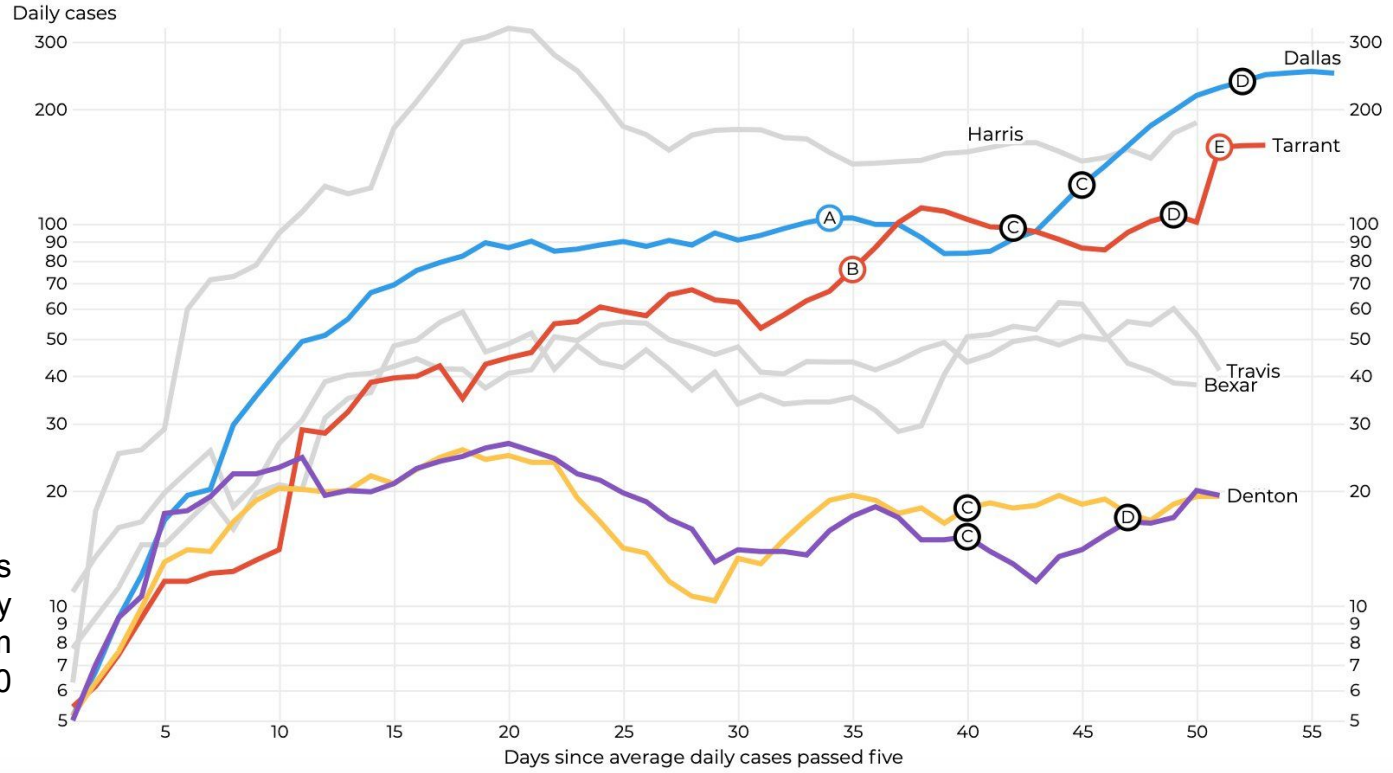




https://towardsdatascience.com/stopping-covid-19-with-misleading-graphs-6812a61a57c9

# Corrected version:



# Even clearly labeled log scales can be deceptive




Dallas Morning News graphic republished by DMagazine. Tim Rogers, May 13, 2020

# Topic 3: Personally identifiable information (PII)

Students learning Python work with some fictitious health data...

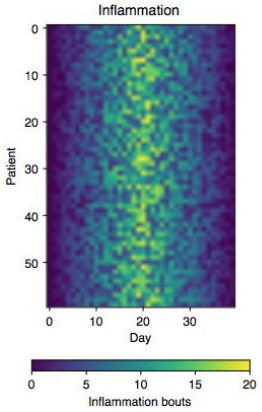
To see how the treatment is affecting the patients in general, we would like to:

1. Calculate the average inflammation per day across all patients.
2. Plot the result to discuss and share with colleagues.

## Inflammation data

| Patients | Day 1 | Day 2 | Day 3 | Day 4 | Day 5 | Day 6 | Day 7 |
|---|---|---|---|---|---|---|---|
| | 0 | 0 | 1 | 3 | 1 | 2 | 4 |
| | 0 | 1 | 2 | 1 | 2 | 1 | 3 |
| | 0 | 1 | 1 | 3 | 3 | 2 | 6 |
| | 0 | 0 | 2 | 0 | 4 | 2 | 2 |
| | 0 | 1 | 1 | 3 | 3 | 1 | 3 |
| | | | | | | | |

## Analysis




## Conclusion


**How does the medication affect patients?**

# Topic 3: Personally identifiable information (PII)

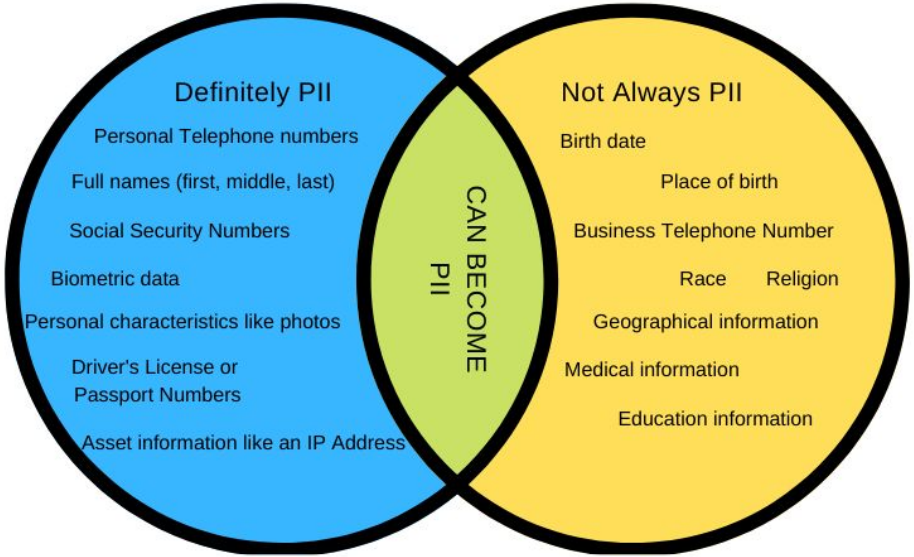
Important concepts:

- Generally PII cannot be released, ethically or legally
  - Health data is additionally protected
- Requires IRB approval and careful planning
- De-identification likely required; may be easy (use opaque IDs) or hard (identity is still inferrable)
- Original data, if it needs to be retained, will require safeguarding
- Best practice: de-identify as soon as possible

# Topic 3: Personally identifiable information (PII)

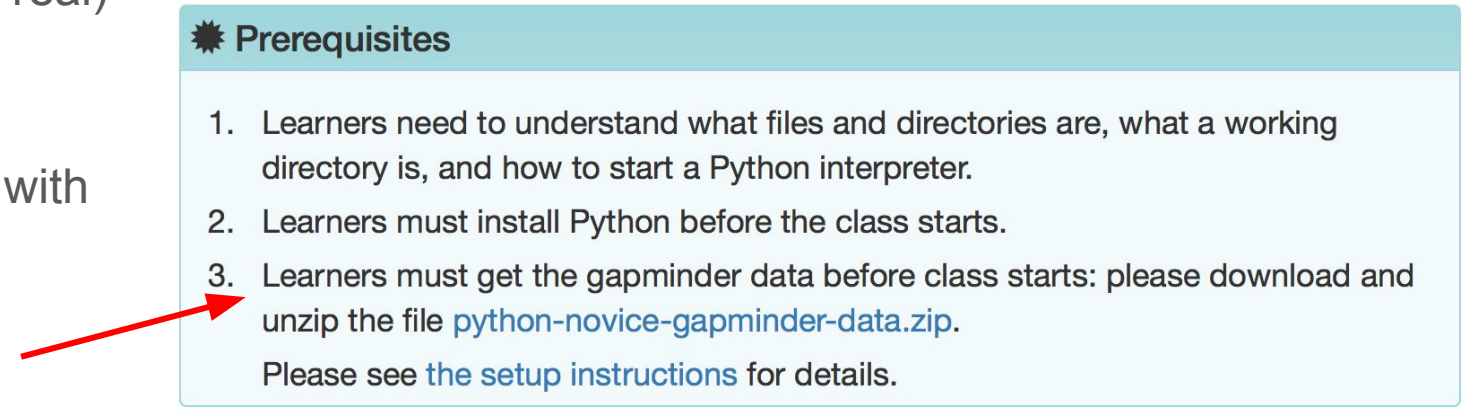
## What constitutes PII?

Some information on its own does not constitute as PII, but when it is linkable with other information, it can be used to identify a person




# Topic 4: Data evaluation and citation

It's common for students to work with pre-existing (and real) datasets, and valuable for the students to work with real data



# Topic 4: Data evaluation and citation

Important concepts:

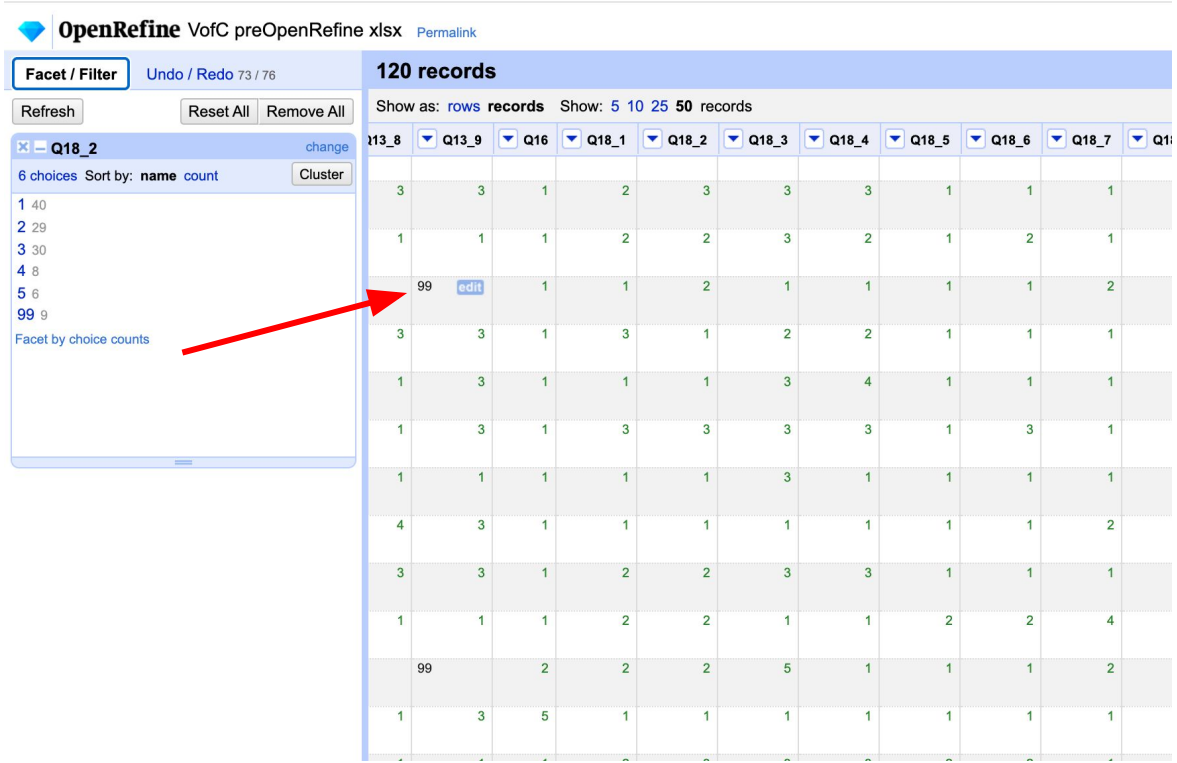
- Evaluation
  - *Who* created it?  *Why*?
  - *How* was the data gathered/created?
  - *What* are its intended uses?  Its limitations?

- Citation
  - How to cite a dataset
  - DOIs
  - Dataset versions/dating
  - Identify subset used

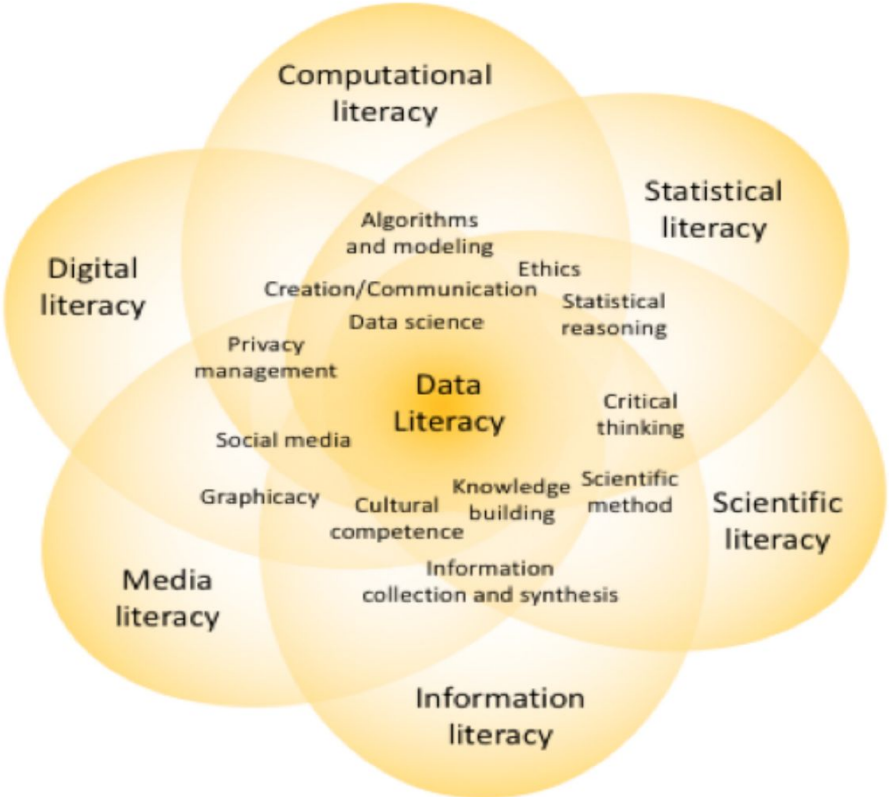
# Topic 5: Handling missing data and Likert scales

While teaching data transformations and values replacement on Spreadsheets and OpenRefine, instructors can:

- Provide tips on best practices for handling missing data
- Cover the advantages and disadvantages of using odd versus even Likert scales in survey design



# Data Literacy: An Interdisciplinary & Multi-Literacy Construct




Many Existing Conceptions:

- *A **life skill** for everyday problem-solving – enabling community engagement, citizen empowerment, activity tracking, and personal health management*
- *A **research skill** for students and professionals – accessing existing data sets to produce and communicate new knowledge, making scientific experiments robust and reproducible*
- *A building block and critical success factor for rolling out data science in **business, government, and research***


Bhargava, R., Deahl, E., Letouzé, E., Noonan, A., Sangokoya, D., & Shoup, N. (2015, October) *Beyond data literacy*: Reinventing community engagement and empowerment in the age of data [White Paper]. Cambridge, MA: Data-Pop Alliance. Retrieved from https://datapopalliance.org/item/beyond-data-literacy-reinventing-community-engagement-and-empowerment-in-the-age-of-data/

Corrall, Sheila (2019) *Repositioning Data Literacy as a Mission-Critical Competence.* In: ACRL 2019: Recasting the Narrative, April 10-13, 2019, Cleveland, OH. Retrieved from http://d-scholarship.pitt.edu/36975/

# Data Literacy

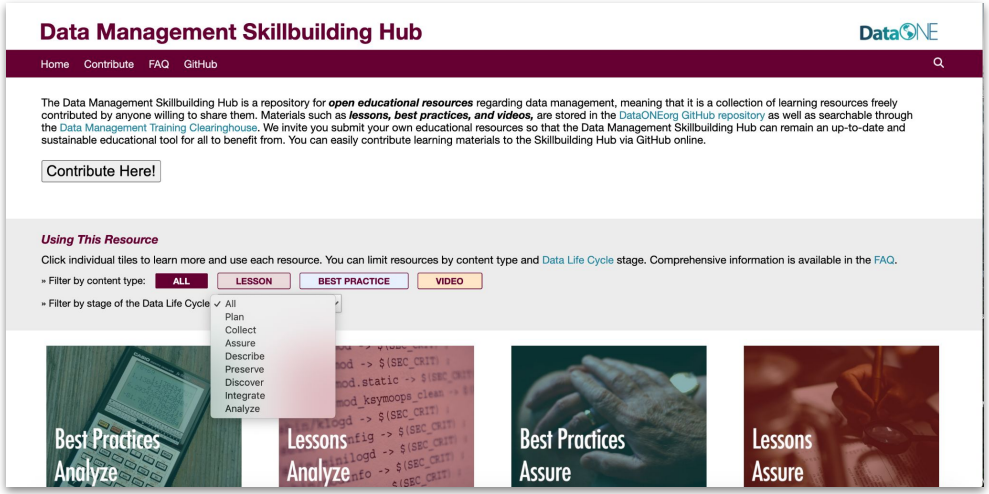
Accounts for individuals' capability to understand, explain, and document the utility and limitations of data by becoming **critical producers and consumers of data**, controlling their personal data trail, finding meaning in data, and taking action based on data.

***The data-literate individual can identify, collect, evaluate, analyze, interpret, present, document, preserve, correctly attribute, and protect data.***

Adapted from: Oceans of Data Institute (2016). *Building global interest in data literacy*: a dialogue. Waltham, MA: Educational Development Center. Retrieved from: http://oceansofdata.org/our-work/building-global-interest-data-literacy-dialogue-workshop-report

# RDM

Research Data Management (RDM) concerns about processes undertaken to create and/or gather organized, documented, accessible, and reusable quality research data.




American Library Association. (April 17, 2018). *Keeping Up With… Research Data Management*. Retrieved from:http://www.ala.org/acrl/publications/keeping_up_with/rdm

https://dataoneorg.github.io/Education/

# Research Data Literacy & RDM

Research Data Literacy can be concretely achieved through the development of competencies and skills connected to different stages of the research data lifecycle

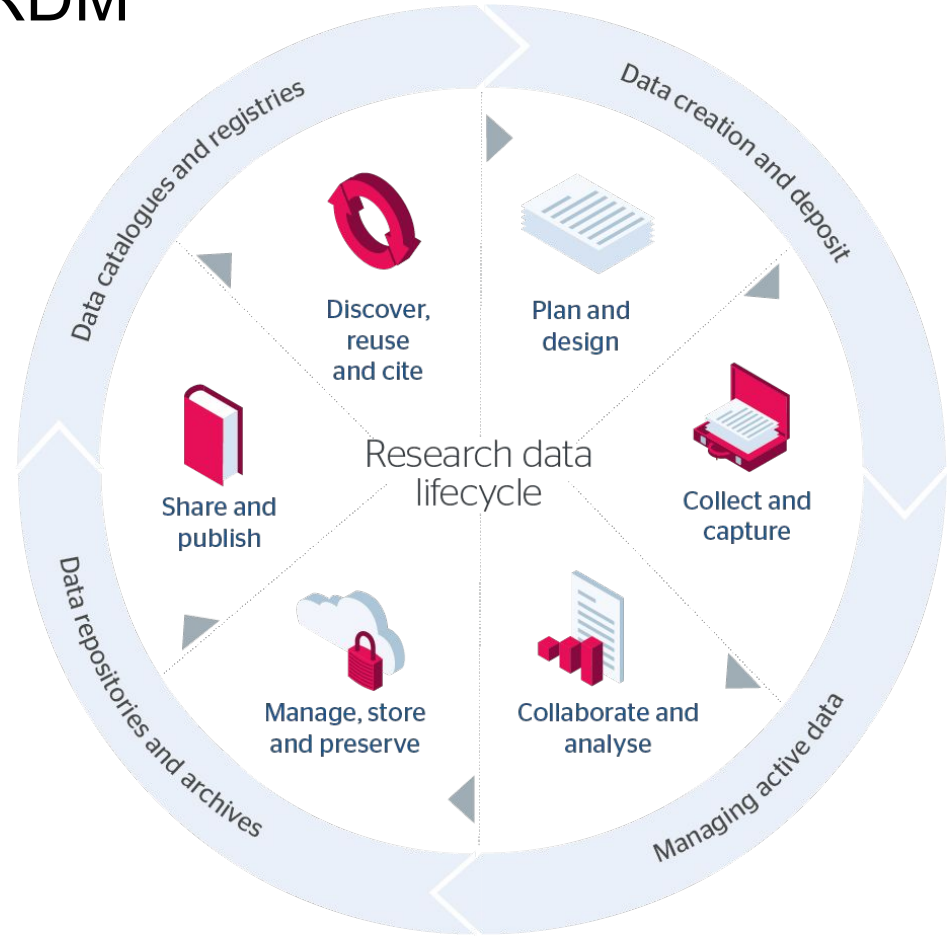
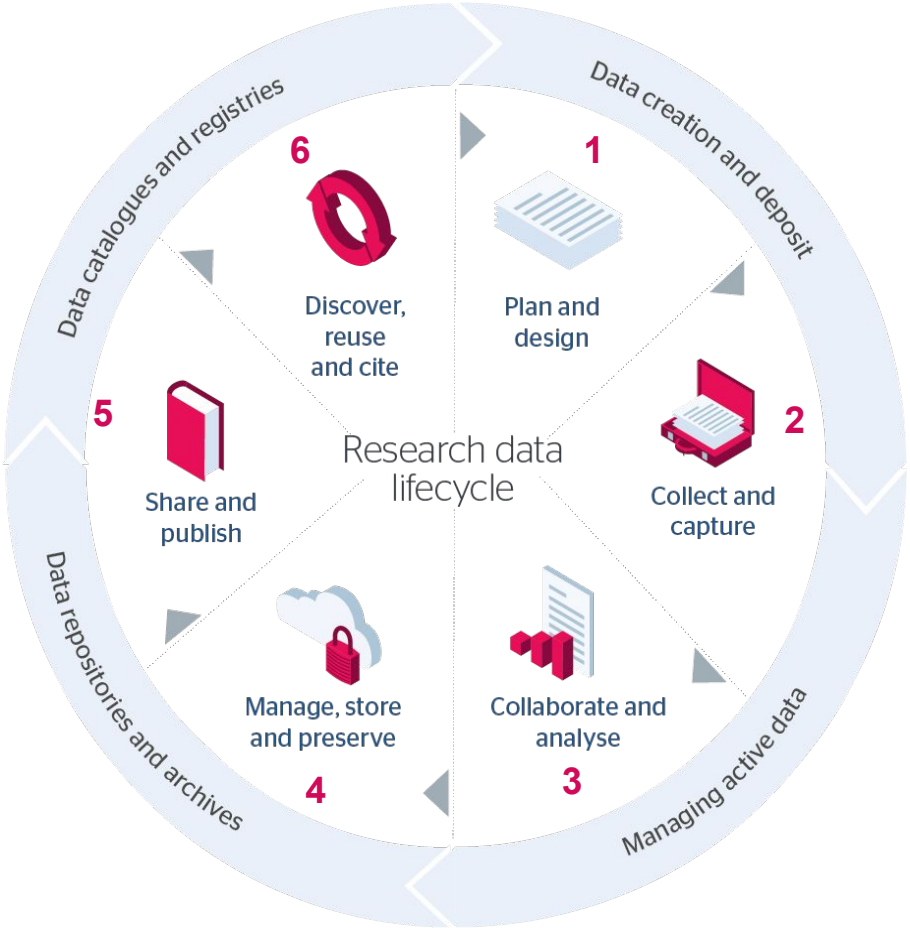



Diagram:
Jisc ([2021]). *Research Data Management Toolkit*. Retrieved from: https://rdmtoolkit.jisc.ac.uk/research-data-lifecycle/

# Topic Examples

1. IRB, DMPs, instrument design...
2. Data entry/input, handling missing data...
3. Data transformation, data cleaning, data visualization, version control...
4. File naming and organization, data documentation and metadata...
5. Licensing, de-identification, FAIR and CARE principles...
6. Data source identification, selection and attribution...



# Breakouts

- We have arranged breakout rooms to concentrate people who expressed familiarity with these Carpentries:


https://medium.datadriveninvestor.com/

- Room 1: Python
  - Software: Plotting and Programming in Python
- Room 2: R
  - Data: Data Analysis and Visualization with R for Social Scientists
- Room 3: Bash / Unix
  - Software: The UNIX Shell
- Room 4: Open Refine
  - Library: Open Refine
- Room 5: Version Control / Git
  - Version Control with Git

# Breakout 1

*What data literacy/RDM topics can and should we be teaching?*

- *We = librarians, curators, facilitators*

# BREAK

Return at 1:55pm PST

# Challenges

Specifically with Carpentry lessons:

- Too much material for the time allotted (always)
- Heavy emphasis on hands-on activity
- Cognitive load; saturation

# Breakout 2

*Where* are *how* can these topics be taught?

- Within Carpentries?
- In other curricula?
- How to add, augment, inject?

# Thank you!

Renata Curty <rcurty@ucsb.edu>
Jon Jablonski <jonjab@ucsb.edu>
Greg Janée <gjanee@ucsb.edu>
Kristi Liu <ksliu@ucsb.edu>
Torin White <whitet@ucsb.edu>