A Pedagogical Approach to Create and Assess Domain Specific Data Science Learning Materials in the Biomedical and Health Sciences

ChangeMedEd 2021

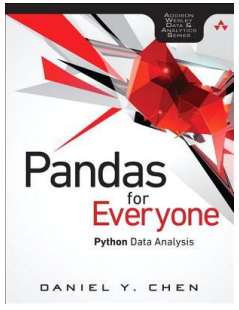
Daniel Chen, MPH

Anne Brown, PhD

2021-09-30

Hello!

- PhD **Candidate**: Virginia Tech (Winter 2021)
  - Data Science education & pedagogy
  - Medical, Biomedical, Health Sciences
- Inten at RStudio, 2019
  - `gradethis`
  - Code grader for `learnr` documents
- The Carpentries
  - Instructor, 2014
  - Trainer, 2020
  - Community Maintainer Lead, 2020
- R + Python!
- Author:



Current Data Science Education

Dedicated Course Titles in 2014 and 2015

Institution	Program	Inference	Modeling	Programming	Data Products	Data Cleaning	Reproducible Science	Exploratory Analysis
Stanford	MS Statistics	Introduction to Statistical Inference	Regression Models and Analysis of Variance	Programming Methodology	NA	NA	NA	NA
CMU	MS Statistical Practice	Advanced Methods for Data Analysis	Applied Linear Models	Statistical Computing	Statistical Practice	NA	NA	NA
NYU	MS Applied Statistics	Applied Statistical Modeling and Inference	Applied Statistical Modeling and Inference	Statistical Computing	NA	NA	NA	NA
Columbia	MA Statistics	Multivariate Statistical Inference	Regression and Multi-Level Models	Statistical Computing and Intro to Data Science	NA	NA	NA	Topics in Modern Statistics: Statistical Graphics
Harvard	AM Statistics	Statistical Inference	Linear and Generalized Linear Models	Statistical Computing	NA	NA	NA	NA
Illinois	MS Statistics	Statistical Analysis	Applied Regression and Design	Statistical Computing	NA	NA	NA	NA
Georgia Tech	MS Statistics	Math Statistics I	Regression Analysis	Computational Statistics	NA	NA	NA	NA
Indiana	MS Applied Statistics	Introduction to Statistical Theory	Applied Linear Models	Statistical Computing	NA	NA	Managing Statistical Research	Exploratory Data Analysis
Johns Hopkins	Data Science Specialization	Statistical Inference	Linear Models	R Programming	Developing Data Products	Getting and Cleaning Data	Reproducible Research	Exploratory Data Analysis
UBC	Master of Data Science	Statistical Inference and Computation I	Regression I	Programming for Data Science	Capstone Project	Data Wrangling	Data Science Workflows	Data Visualization I

- Data Science education is a **commodity**
  - Content is **not** an issue
  - Various learning platforms
- **Domain experts** can help learners improve **data literacy**
- Need more dedicated courses:
  - **Data Products**
  - **Data Cleaning**
  - **Reproducible Science**

Kross, S., Peng, R. D., Caffo, B. S., Gooding, I., and Leek, J. T. (2020). The Democratization of Data Science Education. The American Statistician, 74(1), 1–7. https://doi.org/10.1080/00031305.2019.1668849

Table 1: *Bachelor's and master's programmes in the United States (as of August 2014)*

Degree	*College/school/department offering the programme*	*No. of programmes*
Bachelor's	University/joint departments	3
	Computer Science	3
	Data Science	2
	Business	1
Master's	University/joint departments	17
	Information Science	7
	Computer Science	3
	Statistics	3
	Information Technology	1
	Operational Research	1
	Professional Studies	1

- Joint departments

Table 2: *Core courses in bachelor's programmes (as of August 2014)*

Course	*No. of universities offering the course*
Probability and Statistics	7
Data Mining	7
Programming	5
Discrete Mathematics	4
Data Structures and Algorithms	4
Database	4
Machine Learning	4
Statistical Modelling	3
Data Visualization	3
Introduction to Data Science	2
Artificial Intelligence	2
Computer Security	2

- Probability + Statistics
- Data Mining
- Programming

Song, I.-Y., and Zhu, Y. (2016). Big data and data science: What should we teach? Expert Systems, 33(4), 364–373. https://doi.org/10.1111/exsy.12130

Data Science Programs Are Too General

- Data science programs target **single broad audiences**
- Opportunity to **branch out** to different disciplines
- Democratization of data science education enables more **domain specific** learning materials

Kross, S., Peng, R. D., Caffo, B. S., Gooding, I., and Leek, J. T. (2020). The Democratization of Data Science Education. The American Statistician, 74(1), 1–7. https://doi.org/10.1080/00031305.2019.1668849

Why Domain Specificity?

- You learn better when things are more relevant
- Internal factors for motivation
- Learning feedback loops
- Self-directed learners

- Ambrose, S. A., Bridges, M. W., DiPietro, M., Lovett, M. C., and Norman, M. K. (2010). How learning works: Seven research-based principles for smart teaching. John Wiley & Sons.
- Koch, C., and Wilson, G. (2016). Software carpentry: Instructor Training. https://doi.org/10.5281/zenodo.57571
- Wilson, G. (2019). Teaching tech together: How to make your lessons work and build a teaching community around them. CRC Press.

NIH Strategic Plan for Data Science

Data Infrastructure	Modernized Data Ecosystem	Data Management, Analytics, and Tools	Workforce Development	Stewardship and Sustainability
•Optimize data storage and security •Connect NIH data systems	•Modernize data repository ecosystem •Support storage and sharing of individual datasets •Better integrate clinical and observational data into biomedical data science	•Support useful, generalizable, and accessible tools and workflows •Broaden utility of and access to specialized tools •Improve discovery and cataloging resources	•Enhance the NIH data-science workforce •Expand the national research workforce •Engage a broader community	•Develop policies for a FAIR data ecosystem •Enhance stewardship

Figure 2. NIH Strategic Plan for Data Science: Overview of Goals and Objectives

National Institutes of Health. (2020, September 14). NIH Strategic Plan for Data Science | Data Science at NIH. https://datascience.nih.gov/nih-strategic-plan-data-science

NIH Biomedical Research

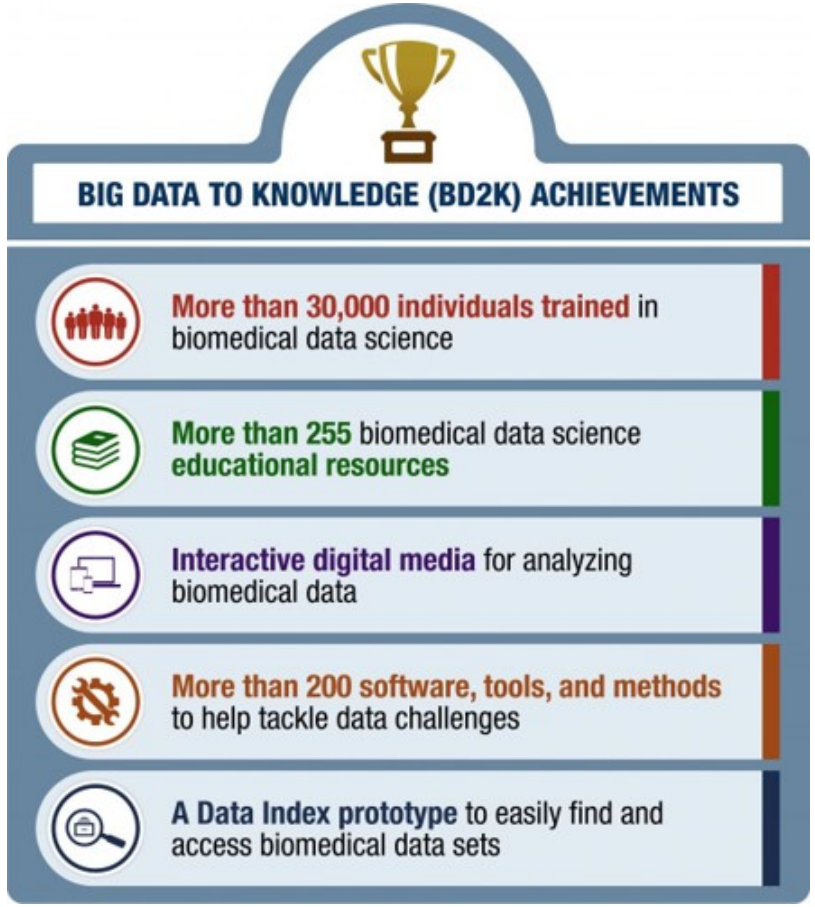
- Support substantial quantities of biomedical data and metadata
- Data is highly distributed
- Accomplished by small groups of researchers

- Variety of formats lead to complications in cleaning

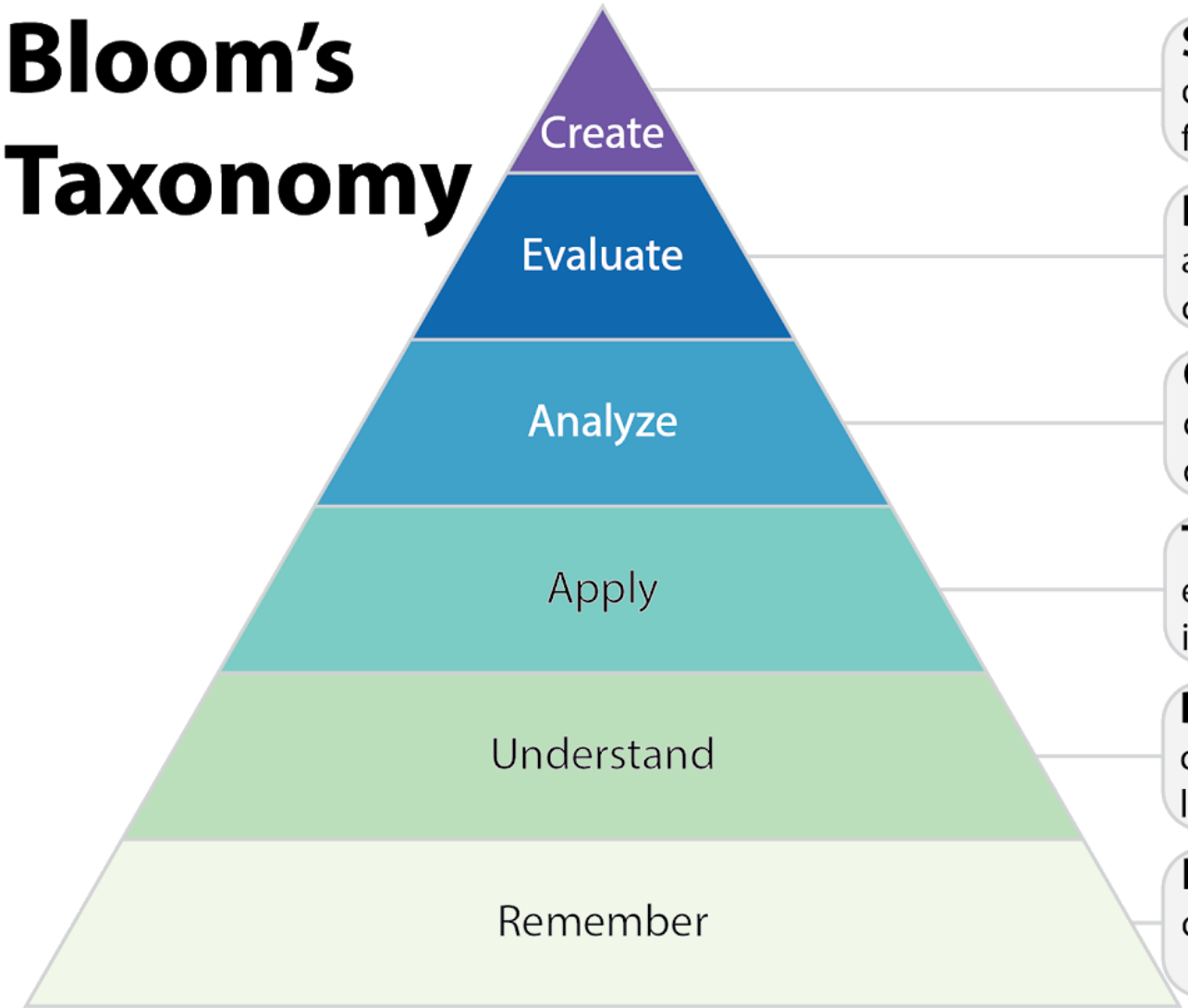
- **Develop a research workforce**

NIH The Big Data to Knowledge (BD2K)

- 2013 - 2018
- Narrow the gap in biomedical data science skills
- Train and educate workforce on analytical skills



Bloom's Taxonomy



Synthesize & Construct New Thought
design, assemble conjecture, develop, formulate, author, investigate

Defend Position
appraise, argue, judge, select, support, value, critique, weigh

Connect Between Concepts
differentiate, organize, relate, compare, contrast, distinguish, examine, experiment, question, test

Translate Knowledge to Application
execute, implement, solve, demonstrate, interpret, operate, schedule, sketch

Describe & Reaffirm Knowledge
classify, describe, discuss, explain, identify, locate, recognize, report, select

Repeat Information & Concepts
define, duplicate, list, memorize, state

Older terms: Knowledge, Comprehension, Application, Analysis, Synthesis, Evaluation

Computing + Statistics Curriculum Guidelines

Computing Education

- 2005: Knowledge-based
- 2020: Competency-based
  - Discrepancy between graduates and work ability

```
competency = knowledge + skill + disposition
           = what + how + why
```

Statistics Education

1. Teach statistical thinking
2. Focus on conceptual understanding.
3. Integrate real data with a context and a purpose.
4. Foster active learning.
5. Use technology to explore concepts and analyze data.
6. Use assessments to improve and evaluate student learning.

- Shackelford R, McGettrick A, Sloan R, et al. Computing Curricula 2005: The Overview Report. In: Proceedings of the 37th SIGCSE Technical Symposium on Computer Science Education. SIGCSE '06. Association for Computing Machinery; 2006:456-457. doi:10.1145/1121341.1121482
- CC2020 Task Force. Computing Curricula 2020: Paradigms for Global Computing Education. ACM; 2020. doi:10.1145/3467967
- GAISE College Report ASA Revision Committee. Guidelines for Assessment and Instruction in Statistics Education (GAISE) College Report 2016.

American Medical Association

PUBLIC HEALTH

Why it's essential to improve data collection and reporting

The AMA is leading the fight against the COVID-19 pandemic. See daily video updates on how the AMA is fighting COVID-19 by discussing the need for improved data collection and reporting.

PRESS RELEASES

AMA underscores importance of science, data in COVID-19 fight

AMA president highlighted essential role of science and data in the fight against COVID-19 during today's national address.

ACCELERATING CHANGE IN MEDICAL EDUCATION

What future physicians need to know: Mastering clinical informatics

The AMA Accelerating Change in Medical Education Consortium and member schools are reshaping the way physicians are trained in many ways. Attendees at a workshop identified curricular goals and outcomes to guide what med students should learn about medical informatics and technology.

ACCELERATING CHANGE IN MEDICAL EDUCATION

How AI is driving new medical frontier for physician training

Medical students at Duke are taking part in machine-learning projects that can change how care is delivered and save lives. Find out how AI is being used in medical education.

ACCELERATING CHANGE IN MEDICAL EDUCATION

Learning to see beyond the patient in the room

The one-on-one approach is no longer sufficient. Population-health management concepts are being integrated into medical education in innovative ways.

PRESS RELEASES

AMA to unleash a new era of patient care

New collaborative initiative brings health and technology stakeholders around a common data model.

ACCELERATING CHANGE IN MEDICAL EDUCATION

At these 3 med schools, health systems science is core component

Faculty from innovative schools are incorporating med ed's "third pillar" and highlight its benefits for students, physicians and patients.

ACCELERATING CHANGE IN MEDICAL EDUCATION

Student interest in informatics outpaces opportunities: Study

A study looked at interest in training for clinical informatics—the study of health information and data to improve patient care. It found that medical students' interest in learning more about health care data outpaces the number of opportunities to do so.

American Medical Association. (2021). Accelerating Change in Medical Education. American Medical Association. https://www.ama-assn.org/education/accelerating-change-medical-education

Applies to All Clinicians

American Nursing Association

Nursing and Big Data

This brain-based learning ... course is designed to engage you in the world of Big **Data** to sharpening your critical thinking skills.

Data Makes the Difference: The Smart Nurse's Handbook for Using Data to Improve Care

Did you know that the **data** used for measuring and safety purposes actually comes directly from the information ... patient care you document daily? You generate the **data** — now it's...

Making Data Work for Nursing: Using Nursing Research and Evidence-based Practice to Affect Nursing Outcomes

Making **Data** Work for Nursing: Using Nursing Research and Evidence-based Practice to Affect Nursing Outcomes ... Outcomes Making **Data** Work for Nursing: Using Nursing Research and...

Overcome Education Challenges

- Elective courses in informatics
- Professional society incentives
- Online or in-person forums to bring interest parties together
- Informal partnerships between medical students and informatics experts

- ANA Enterprise | American Nurses Association. ANA. Accessed September 29, 2021. https://www.nursingworld.org/
- Student interest in informatics outpaces opportunities: Study. American Medical Association. Accessed September 29, 2021. https://www.ama-assn.org/education/accelerating-change-medical-education/student-interest-informatics-outpaces-opportunities

Interest in Informatics Outpace Opportunities

ACCELERATING CHANGE IN MEDICAL EDUCATION

Student interest in informatics outpaces opportunities: Study

A study looked at interest in training for clinical informatics—the study of health information and data to improve patient care. It found that medical students' interest in learning more about health care data outpaces the number of opportunities to do so.

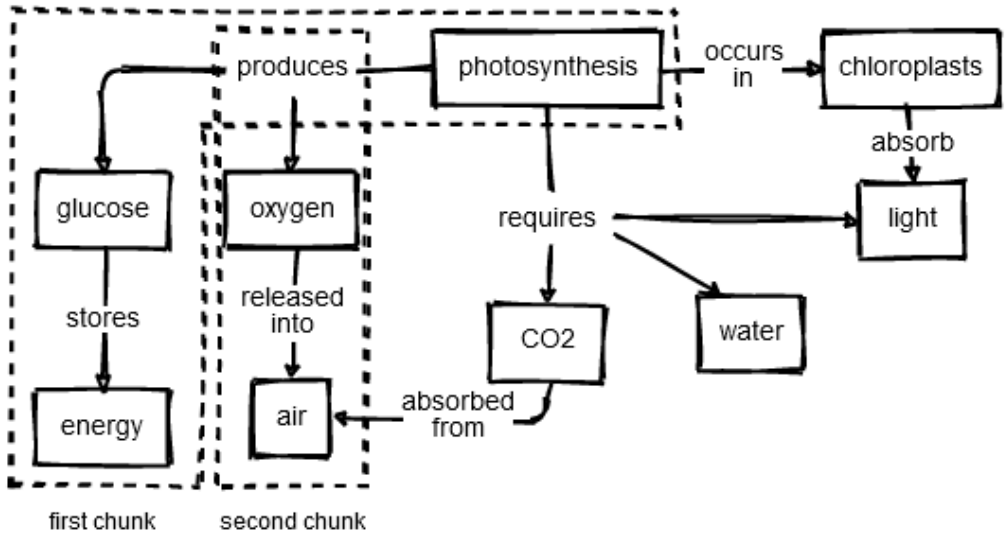
- Students who are interest in a clinical informatics related career
- Not aware of training opportunities
- Need to increase **quantity**, **quality**, and **publicity**

- American Medical Association. (2021). Accelerating Change in Medical Education. American Medical Association. https://www.ama-assn.org/education/accelerating-change-medical-education
- Banerjee R, George P, Priebe C, Alper E. Medical student awareness of and interest in clinical informatics. Journal of the American Medical Informatics Association. 2015;22(e1):e42-e47. doi:10.1093/jamia/ocu046

Identifying Our Learners

What Do Our Learners Know?

Concept Maps



Using concept maps in lesson design

Dreyfus model of skill acquisition

Can also use "task deconstruction"

Novice, Competent, Proficient, Expert, Master

- Dreyfus, S. E., and Dreyfus, H. L. (1980). A five-stage model of the mental activities involved in directed skill acquisition. California Univ Berkeley Operations Research Center.
- Koch, C., and Wilson, G. (2016). Software carpentry: Instructor Training. https://doi.org/10.5281/zenodo.57571
- Wilson, G. (2019). Teaching tech together: How to make your lessons work and build a teaching community around them. CRC Press.

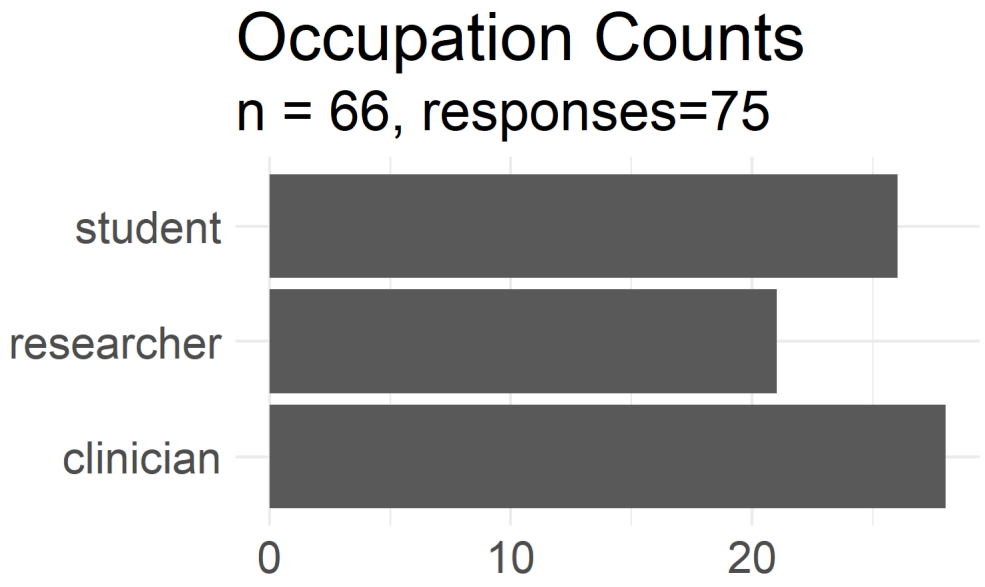
Identify Learners: Learner Self-Assessment Survey

- VT IRB-20-537
- Surveys: https://github.com/chendaniely/dissertation-irb/tree/master/irb-20-537-data_science_workshops
    - Currently working on survey validation
- Combination of:
    - **The Carpentries** surveys: https://carpentries.org/assessment/
    - **"How Learning Works: Seven Research-Based Principles for Smart Teaching"** by Susan A. Ambrose, Michael W. Bridges, Michele DiPietro, Marsha C. Lovett, Marie K. Norman
    - **"Teaching Tech Together"** by Greg Wilson

1. Demographics (6)
2. Programs Used in the Past (1)
3. **Programming Experience** (6)
4. **Data Cleaning and Processing Experience** (4)
5. **Project and Data Management** (2)
6. **Statistics** (4)
7. Workshop Framing and Motivation (3)
8. Summary Likert (7)

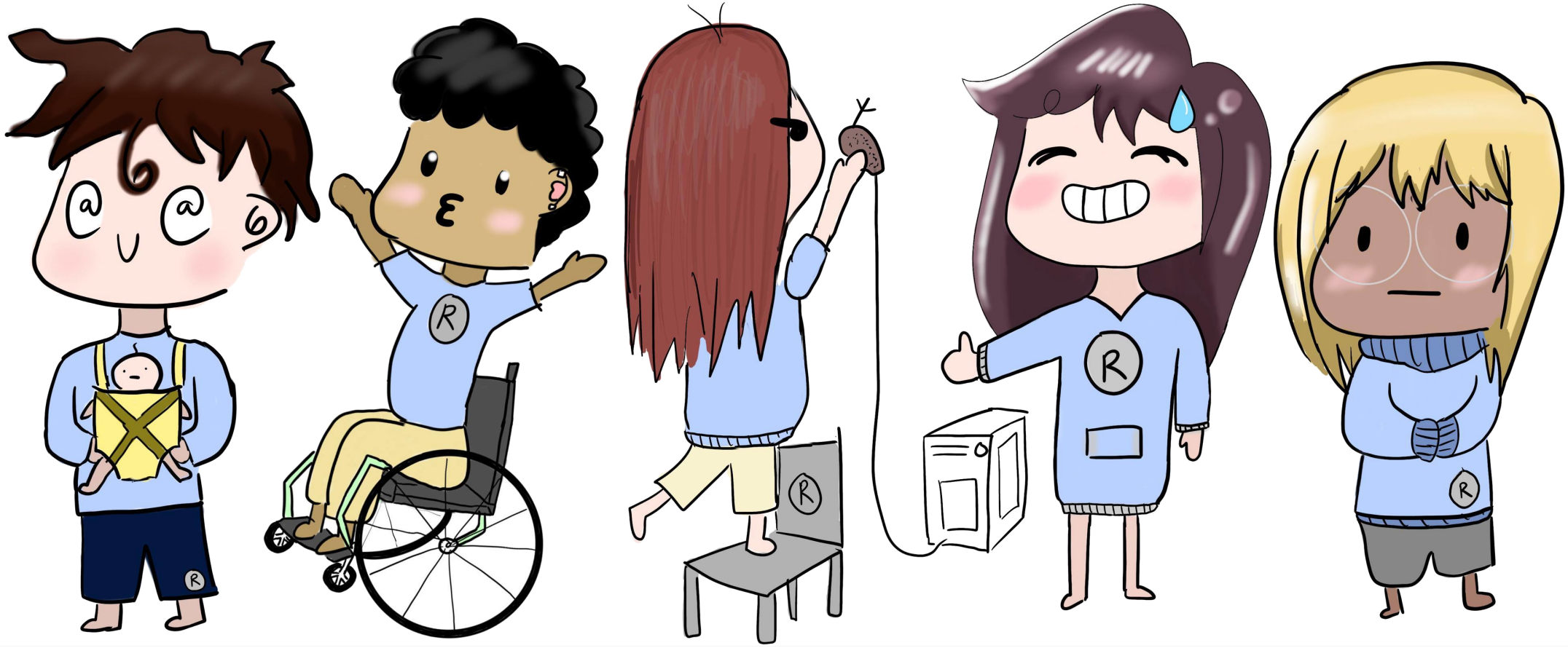
Occupations

- Grouped occupation demographic data into one of 3 groups.



The Personas

Clare Clinician, Samir Student, Patricia Programmer, Alex Academic
https://ds4biomed.tech/who-is-this-book-for.html#the-personas



Clare Clinician

Figure 0.3: Drawn by Julia Chen

Background

Clare has spent the last 6 years working in the Cardiothroasic ICU in a large medical hospital system. They read lots of gushing articles about data science, and was excited by the prospect of learning how to do it, but nothing makes sense when trying to learn it on their own. Clare has always been a good student and always excelled at things they tried to learn; they are hard on themselves when struggling to learn a new skill and would rather place blame on the long hours at work than having their peers know they could use assistance.

Relevant prior knowledge or experience

Clare keeps up with medical research, but has little to no experience in doing medical research. They use Excel for non-data related tasks (e.g., making lists), or manually inputting patient data into spreadsheets for chart reviews. Wants to be able to collect and manage data as well as learn about the process behind data analysis to perform their own analysis and study one day.

Perception of needs

Clare wants self-paced tutorials with practice exercises, plus forums where they can ask for help. They also need short overviews to orient them and introductory tutorials that include videos or animated GIFs showing exactly how to drive the tools, and that use datasets they can relate to. Clare wishes they had a community of other people in the medical field who are interested in learning how to do data work so they can learn and ask questions.

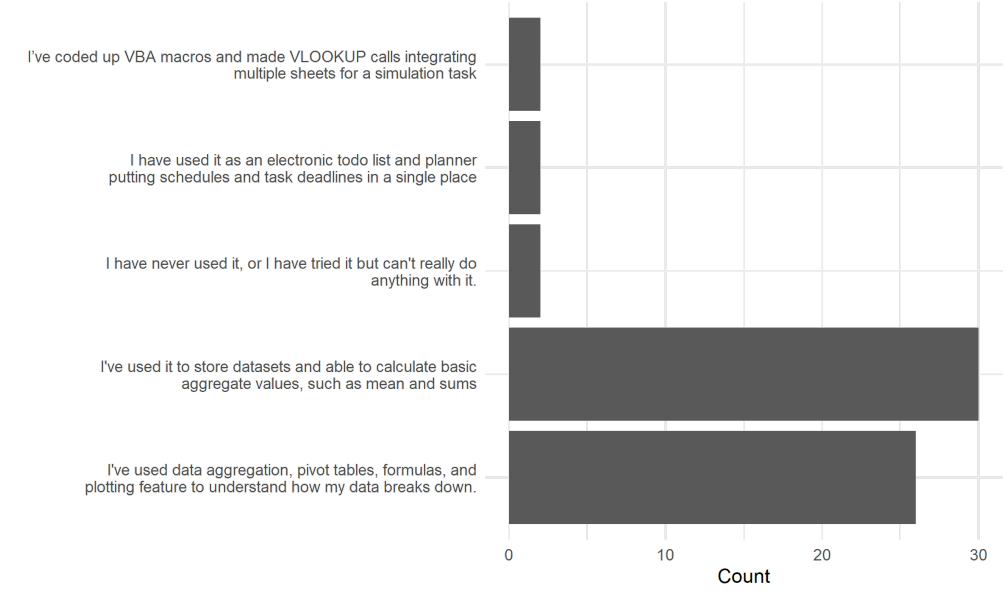
Special considerations

Clare is a single parent who juggle their time at work and at home who are strapped for time to learn a new skill.

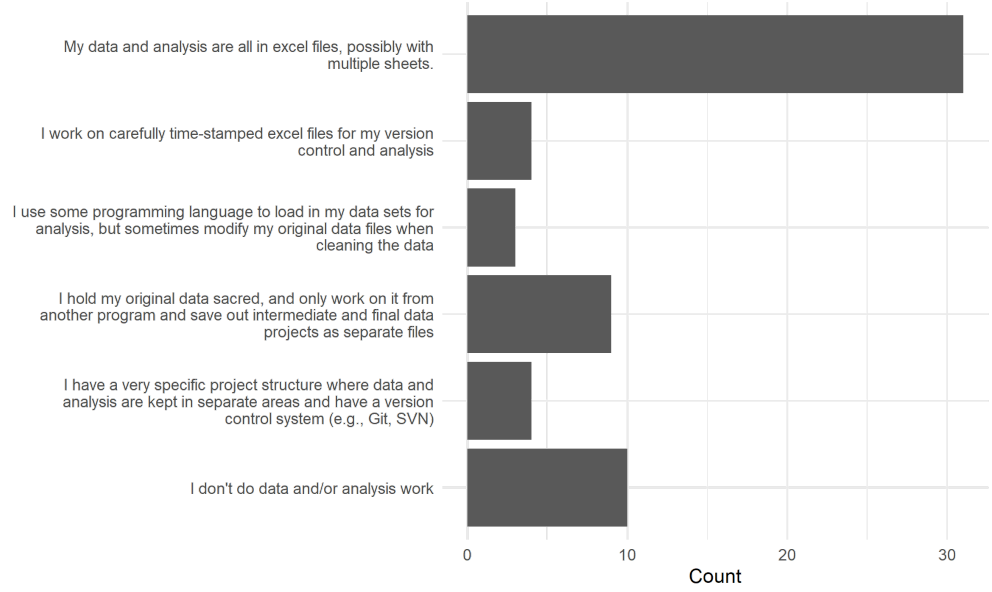
Plan the Learning Materials

Survey Responses: Excel

How familiar are you with Microsoft Excel?

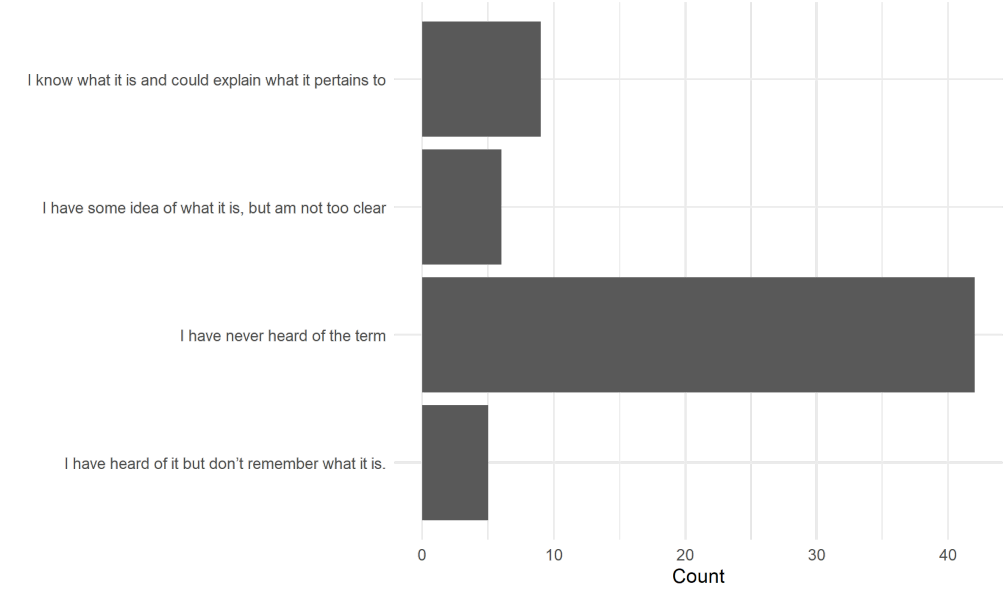


Which of the following best describes how do you manage your data and analysis?

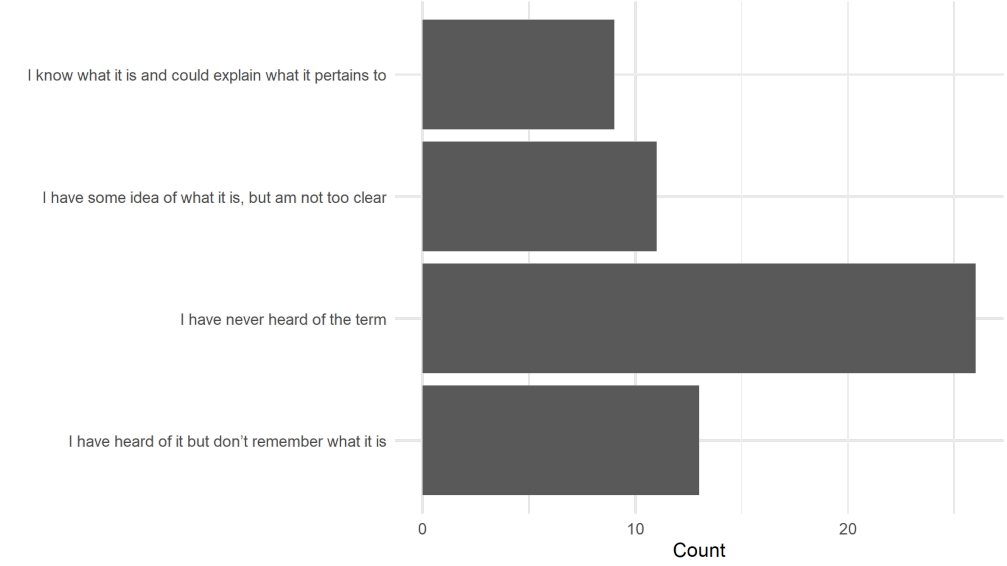


Survey Responses: Data Literacy

Do you know what ""long"" and ""wide"" data are?



Are you familiar with the term ""dummy variable""? It is sometimes also called ""one-hot encoding"".



Planning the Learning Materials

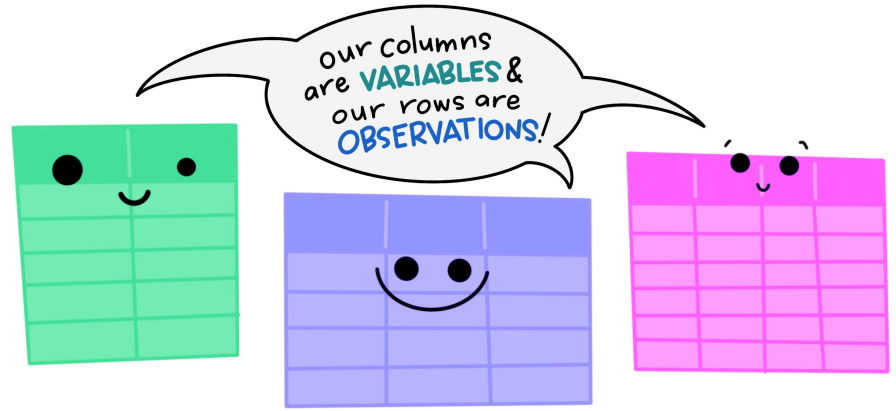
Learning objectives:

1. **Name** the features of a tidy/clean dataset
2. **Transform** data for analysis
3. **Identify** when spreadsheets are useful
4. **Assess** when a task should not be done in a spreadsheet software
5. **Break down** data processing into smaller individual (and more manageable) steps
6. **Construct** a plot and table for exploratory data analysis
7. **Build** a data processing pipeline that can be used in multiple programs
8. **Calculate, interpret, and communicate** an appropriate statistical analysis of the data

Tidy Data

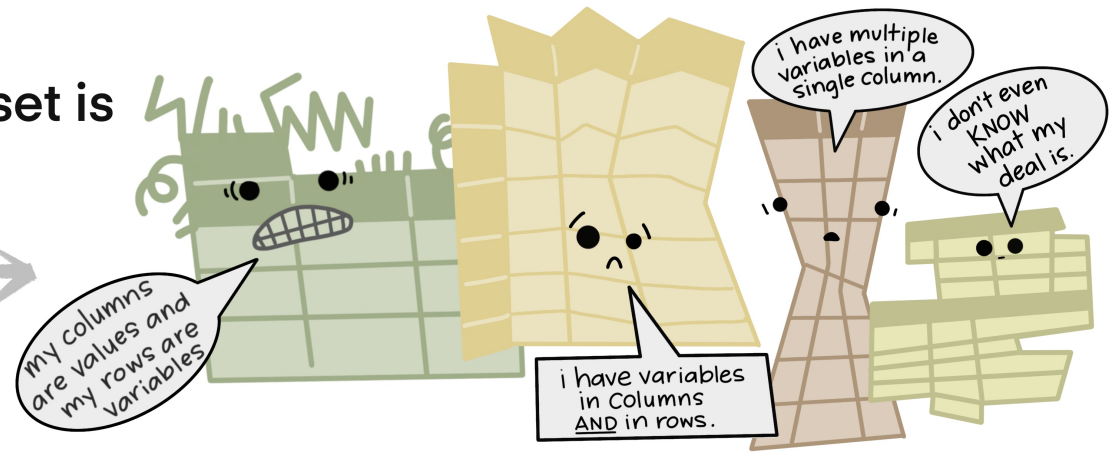
Data is messy in different ways

The standard structure of tidy data means that "tidy datasets are all alike..."



"...but every messy dataset is messy in its own way."

–HADLEY WICKHAM



- Allison Horst's Illustrations: https://github.com/allisonhorst/stats-illustrations

country	year	m014	m1524	m2534	m3544	m4554	m5564	m65	mu	f014
AD	2000	0	0	1	0	0	0	0	—	—
AE	2000	2	4	4	6	5	12	10	—	3
AF	2000	52	228	183	149	129	94	80	—	93
AG	2000	0	0	0	0	0	0	1	—	1
AL	2000	2	19	21	14	24	19	16	—	3
AM	2000	2	152	130	131	63	26	21	—	1
AN	2000	0	0	1	2	0	0	0	—	0
AO	2000	186	999	1003	912	482	312	194	—	247
AR	2000	97	278	594	402	419	368	330	—	121
AS	2000	—	—	—	—	1	1	—	—	—

Table 9: Original TB dataset. Corresponding to each 'm' column for males, there is also an 'f' column for females, `f1524`, `f2534` and so on. These are not shown to conserve space. Note the mixture of 0s and missing values (—). This is due to the data collection process and the distinction is important for this dataset.

Wickham, H. (2014). Tidy Data. Journal of Statistical Software, 59(1), 1–23. https://doi.org/10.18637/jss.v059.i10

country	year	column	cases
AD	2000	m014	0
AD	2000	m1524	0
AD	2000	m2534	1
AD	2000	m3544	0
AD	2000	m4554	0
AD	2000	m5564	0
AD	2000	m65	0
AE	2000	m014	2
AE	2000	m1524	4
AE	2000	m2534	4
AE	2000	m3544	6
AE	2000	m4554	5
AE	2000	m5564	12
AE	2000	m65	10
AE	2000	f014	3

(a) Molten data

country	year	sex	age	cases
AD	2000	m	0-14	0
AD	2000	m	15-24	0
AD	2000	m	25-34	1
AD	2000	m	35-44	0
AD	2000	m	45-54	0
AD	2000	m	55-64	0
AD	2000	m	65+	0
AE	2000	m	0-14	2
AE	2000	m	15-24	4
AE	2000	m	25-34	4
AE	2000	m	35-44	6
AE	2000	m	45-54	5
AE	2000	m	55-64	12
AE	2000	m	65+	10
AE	2000	f	0-14	3

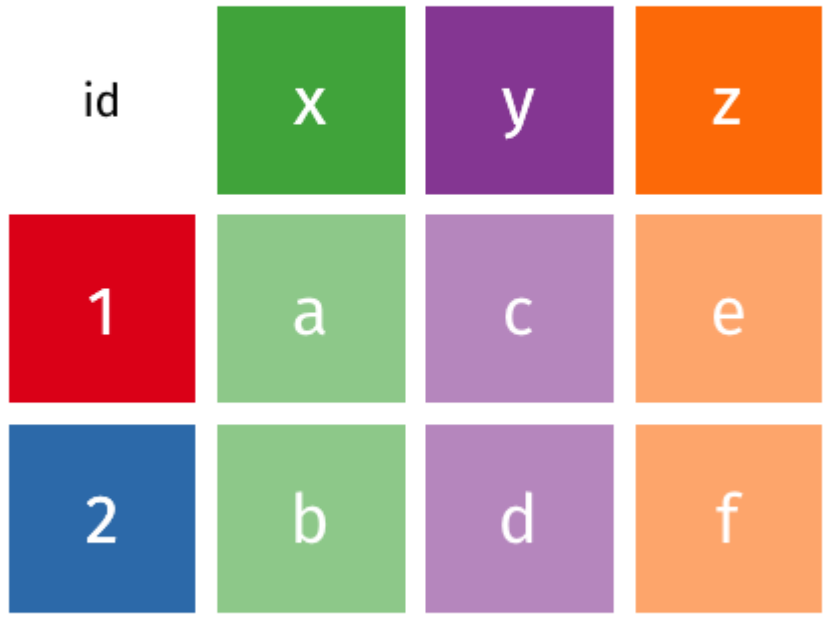
(b) Tidy data

Table 10: Tidying the TB dataset requires first melting, and then splitting the `column` column into two variables: `sex` and `age`.

Wickham, H. (2014). Tidy Data. Journal of Statistical Software, 59(1), 1–23. https://doi.org/10.18637/jss.v059.i10

A different view of data

wide



https://www.garrickadenbuie.com/project/tidyexplain/

Learning and Teaching Materials

ds4biomed Part 1 (6 Hours)

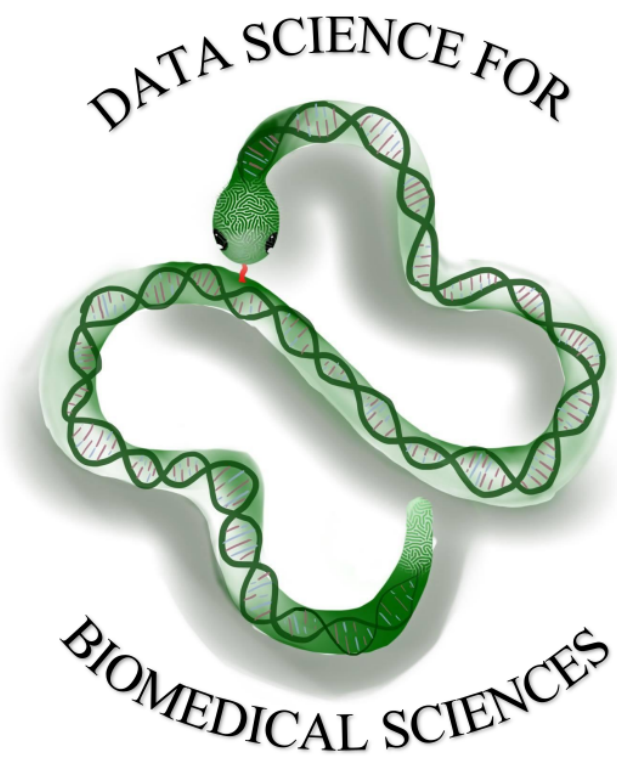
https://ds4biomed.tech/

1. Introduction
2. Spreadsheets
3. R + RStudio / Python + JupyterLab
4. Load Data
5. Descriptive Calculations

---

1. Clean Data (Tidy)
2. Visualization (Intro)
3. Analysis Intro (Logistic Regression)

ds4biomed Part 2 (6 Hours)



https://ds4biomed.tech/

1. 30-Day re-admittance
2. Working with multiple datasets
    - Joins
    - Databases

---

1. APIs + Census data
2. Functions
3. Survival Analysis
4. Machine Learning Basics

Example: Load a dataset

Python

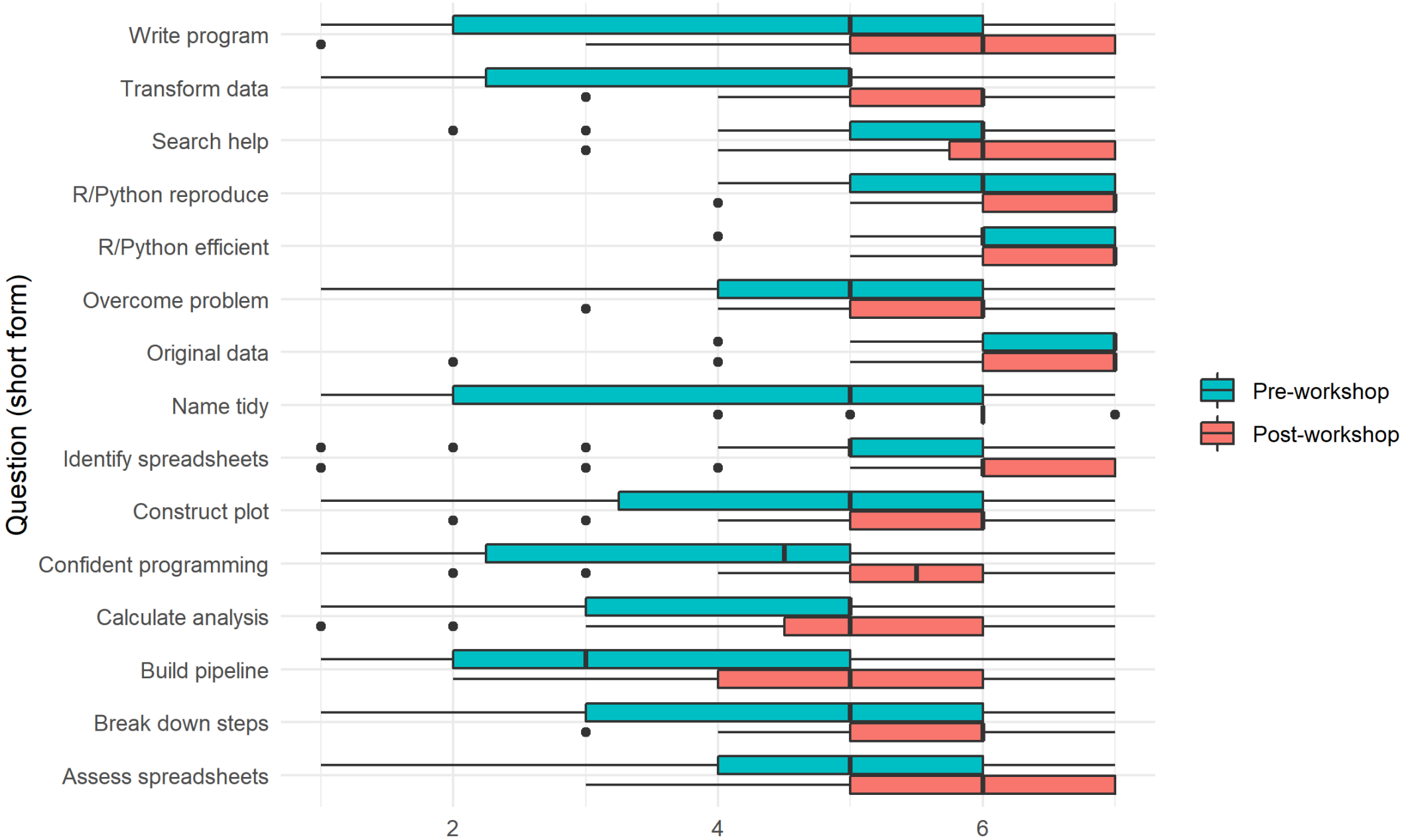
```python
# load a library
# library alias
import pandas as pd

# use a library function
# know about paths
# variable assignment
# function arguments
dat = pd.read_excel("./data/cmv.xlsx")
```

R

```r
# load library
library(tidyverse)
library(readxl)

# use a library function
# know about paths
# variable assignment
# function arguments
dat <- read_excel("./data/cmv.xlsx")
```


Question (short form)
Write program
Transform data
Search help
R/Python reproduce
R/Python efficient
Overcome problem
Original data
Name tidy
Identify spreadsheets
Construct plot
Confident programming
Calculate analysis
Build pipeline
Break down steps
Assess spreadsheets
2
4
6
Pre-workshop
Post-workshop

How does this help my practice?

- You can explore your own patient data
- Can work on curating your own data

- Potentially faster research-question cycle

- Continuing education

Get Started

Create Your Own Learner Personas

If you do end up teaching a domain specific group (e.g., biomedical sciences)

1. Identify who your learners are
2. Figure out what they need and want to know
3. Plan a guided learning tract

- Use the surveys I've compiled.

https://github.com/chendaniely/dissertation-irb/tree/master/irb-20-537-data_science_workshops

What's Next?

- Survey Validation (Factor Analysis)
- Learner pre/post workshop "confidence"
- Long-term survey for confidence + retention (summative assessment)
- Different types of formative assessment questions

Organizing Committee

STEPHAN KADAUKE — CHAIR
Children's Hospital of Philadelphia

MICHAEL KANE
Yale University School of Public Health

MARA ALEXEEV
Boston Children's Hospital

DANIELLA MARK
Procogia

JOSEPH RICKERT
RStudio

Program Committee

PETER HIGGINS — CHAIR
University of Michigan

BETH ATKINSON
Mayo Clinic

BOB ENGLE
Biogen

STEVE SCHWAGER
Cornell

ROMI ADMANIT
Biogen

DENISE ESSERMAN
Yale University School of Public Health

Resources and Communities

- R4DS Community: Slack: r4ds.io/join
- R-Ladies: https://rladies.org/
- Py-Ladies: https://pyladies.com/
- R/Medicine: Twitter: https://twitter.com/r_medicine
- OHDSI: https://ohdsi.org/
- Tidy Tuesday: https://github.com/rfordatascience/tidytuesday
- Big Book of R: https://www.bigbookofr.com/

Thanks!

Slides: https://speakerdeck.com/chendaniely/a-pedagogical-approach-to-create-and-assess-domain-specific-data-science-learning-materials-in-the-biomedical-and-health-sciences

Repo: https://github.com/chendaniely/2021-09-30-changemeded-ds4biomed

Prelims: https://chendaniely.github.io/dissertation-prelim