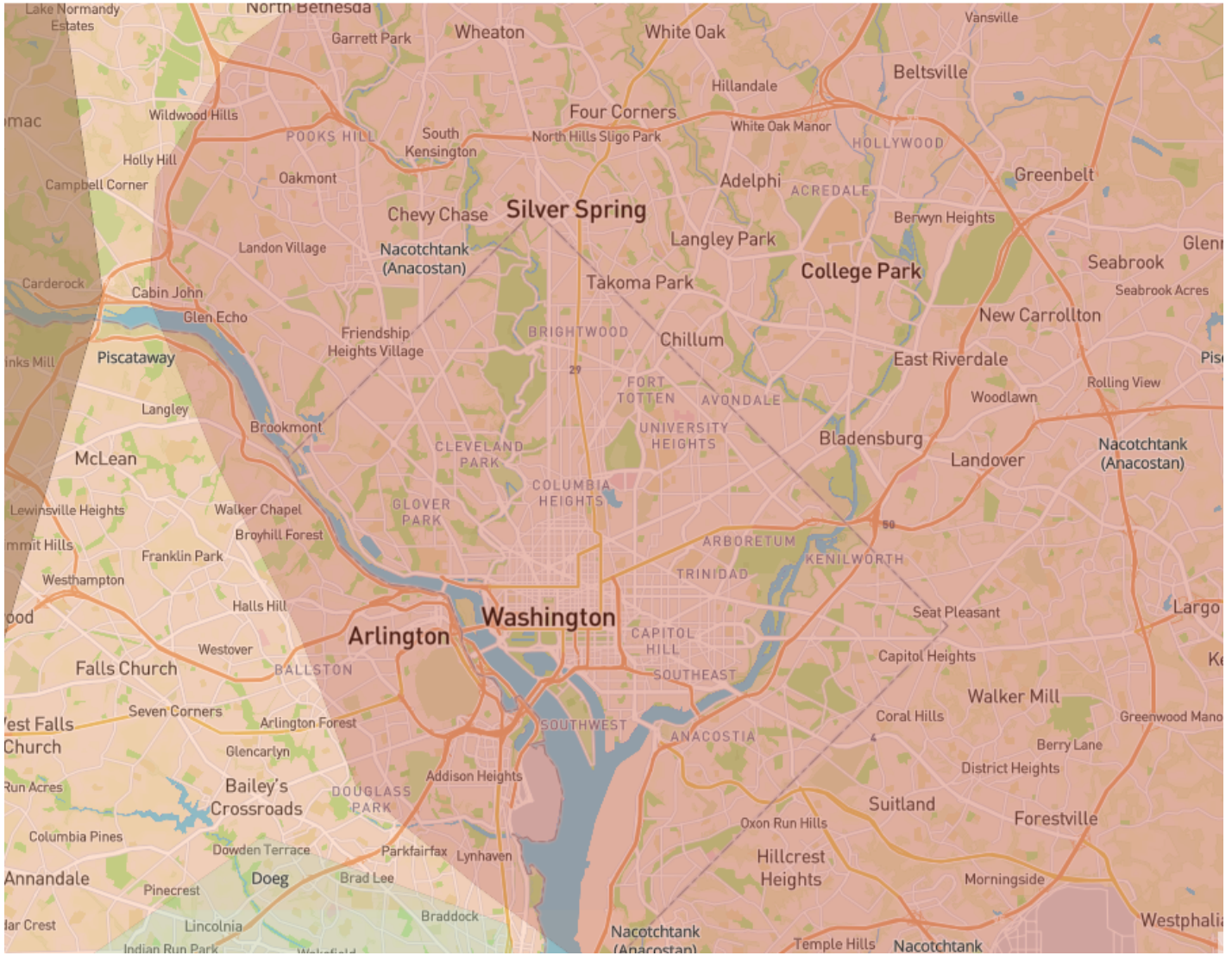
Data Science Education in the Biomedical Sciences

Online education for data science: Opportunities and challenges: AMIA 2022 LIEAF01: Panel

Daniel Chen

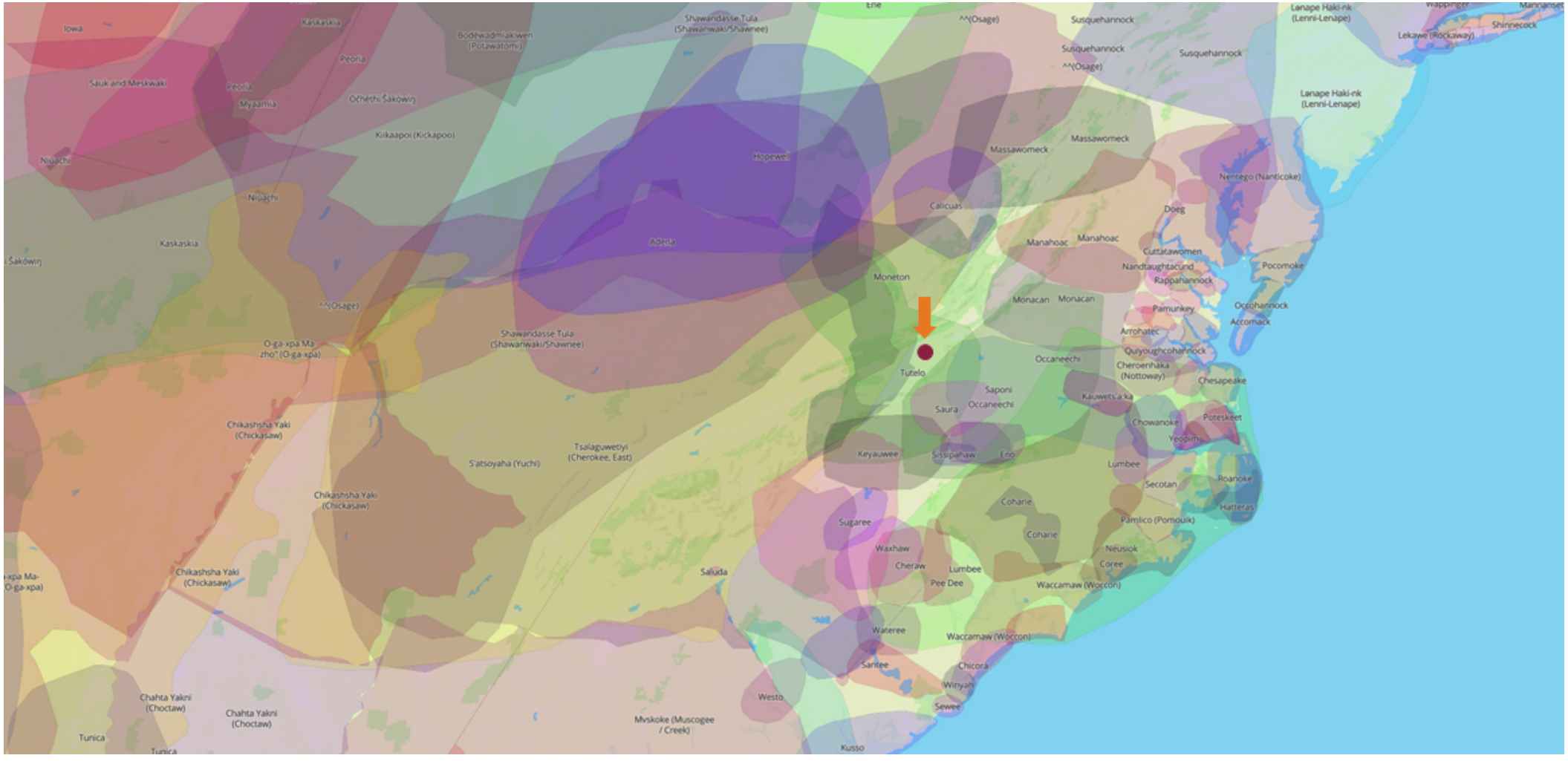
Monday, November 7, 2022

Washington DC



- https://native-land.ca/

Tutelo



- https://native-land.ca/

Daniel Chen, PhD, MPH

@chendaniely

- Postdoctoral Research and Teaching Fellow, UBC, MDS-Vancouver
- Data Science Educator, Posit, PBC (Posit Academy)
- The Carpentries
- Author, Pandas for Everyone

Data Science in the Biomedical Science

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

Informatics Interest Outpace Opportunities

ACCELERATING CHANGE IN MEDICAL EDUCATION

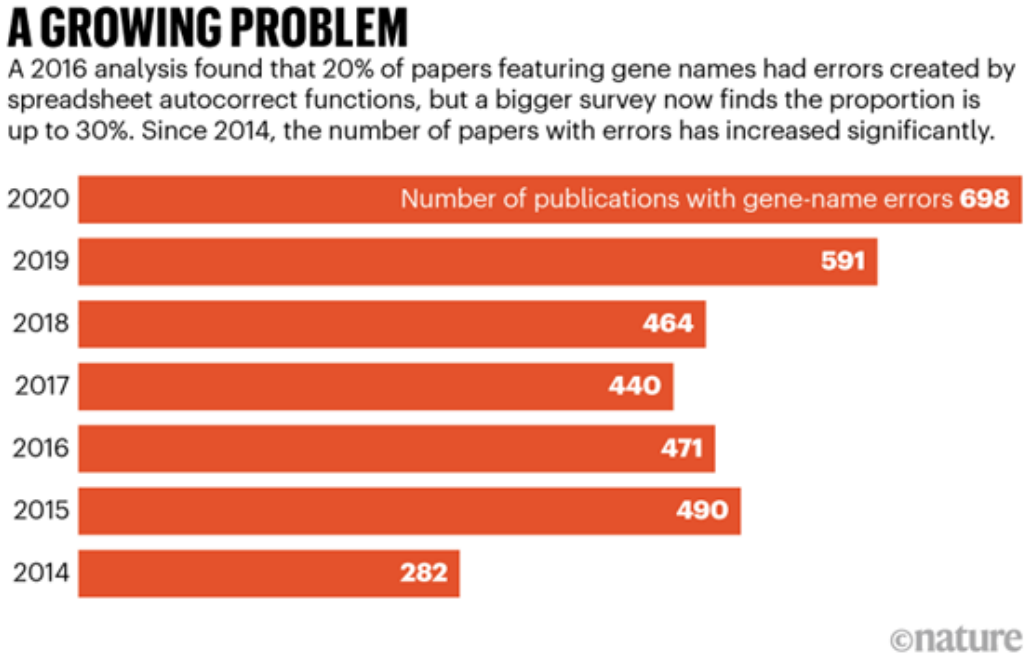
Student interest in informatics outpaces opportunities: Study

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 interested in a clinical informatics related career
- Not aware of training opportunities
- Need to increase: quantity, quality, and publicity

- American Medical Association. Accelerating Change in Medical Education. American Medical Association. Accessed February 10, 2021. 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

Excel



Scientists rename human genes to stop Microsoft Excel from misreading them as dates

Sometimes it's easier to rewrite genetics than update Excel

By James Vincent | Aug 6, 2020, 8:44am EDT

- Lewis D. Autocorrect errors in Excel still creating genomics headache. Nature. Published online August 13, 2021. doi:10.1038/d41586-021-02211-4
- Vincent J. Scientists rename human genes to stop Microsoft Excel from misreading them as dates. The Verge. Published August 6, 2020. Accessed December 8, 2021. https://www.theverge.com/2020/8/6/21355674/human-genes-rename-microsoft-excel-misreading-dates

Consequences of Reproducibility Failures

Table. 1: Case studies that illustrate the significant real world consequences of reproducibility failures in data analyses.

Reproducibility error	Consequence	Source(s)
Limitations in Excel data formats	Loss of 16,000 COVID case records in the UK	(Kelion 2020)
Automatic formatting in Excel	Important genes disregarded in scientific studies	(Zeeberg et al. 2004; Ziemann, Eren, and El-Osta 2016)
Deletion of a cell caused rows to shift	Mix-up of which patient group received the treatment	(Wallensteen et al. 2018)
Using binary instead of explanatory labels	Mix-up of the intervention with the control group	(Aboumatar and Wise 2019)
Using the same notation for missing data and zero values	Paper retraction	(Whitehouse et al. 2021)
Incorrectly copying data in a spreadsheet	Delay in the opening of a hospital	(Picken 2020)

- Aboumatar, Hanan and Robert A. Wise (Oct. 2019). "Notice of Retraction. Aboumatar et al. Effect of a Program Combining Transitional Care and Long-Term Self-Management Support on Outcomes of Hospitalized Patients With Chronic Obstructive Pulmonary Disease: A Randomized Clinical Trial. JAMA. 2018;320(22):2335-2343." In: JAMA 322.14, pp. 1417–1418. issn: 0098-7484. doi: 10.1001/jama.2019.11954
- Kelion, Leo (Oct. 2020). "Excel: Why Using Microsoft's Tool Caused Covid-19 Results to Be Lost". en-GB. In: BBC News.
- Ostblom J, Timbers T. Opinionated practices for teaching reproducibility: motivation, guided instruction and practice. arXiv:210913656 [cs, stat]. Published online September 17, 2021. Accessed November 30, 2021. http://arxiv.org/abs/2109.13656
- Wallensteen, Lena et al. (2018). "Retraction notice to" Evaluation of behavioral problems after prenatal dexamethasone treatment in Swedish adolescents at risk of CAH"[Hormones and Behavior 85C (2016) 5-11]". In: Hormonesand behavior 103, p. 140.
- Whitehouse, Harvey et al. (July 2021). "Retraction Note: Complex Societies Precede Moralizing Gods throughout World History". en. In: Nature 595.7866, pp. 320–320. issn: 1476-4687. doi: 10.1038/s41586-021-03656-3.
- Zeeberg, Barry R et al. (2004). "Mistaken identifiers: gene name errors can be introduced inadvertently when using Excel in bioinformatics". In: BMC bioinformatics 5.1, pp. 1–6.
- Ziemann, Mark, Yotam Eren, and Assam El-Osta (2016). "Gene name errors are widespread in the scientific literature". In: Genome biology 17.1, pp. 1–3.

Backward Design Learning Materials

1. Identify your learners (learner persona)
2. Plan out your lesson content (concept maps)
3. Define overall goal (summative assessment)
4. Break down the goal (formative assessment)
5. Outline the course
6. Write a summary of the course

- Wilson G. Teaching Tech Together: How to Make Your Lessons Work and Build a Teaching Community around Them. Taylor & Francis; 2019. http://teachtogether.tech

Identification of Biomedical Data Science Learner Persons

Implications and Lessons Learned for Domain-Specific Data Science Curriculum

What are Personas?

- Come from product design
- Detailed description of an imaginary person
- Embodies assumptions of the user and product
- Cannot and should not represent every possible user

- Pruitt J, Adlin T. The Persona Lifecycle: Keeping People in Mind Throughout Product Design. 1st edition. Morgan Kaufmann; 2006.

Creating a "Wrong" Persona

- Still backed by data
- "Product"is still consistent
- Personas are a work in progress

- Pruitt J, Adlin T. The Persona Lifecycle: Keeping People in Mind Throughout Product Design. 1st edition. Morgan Kaufmann; 2006.

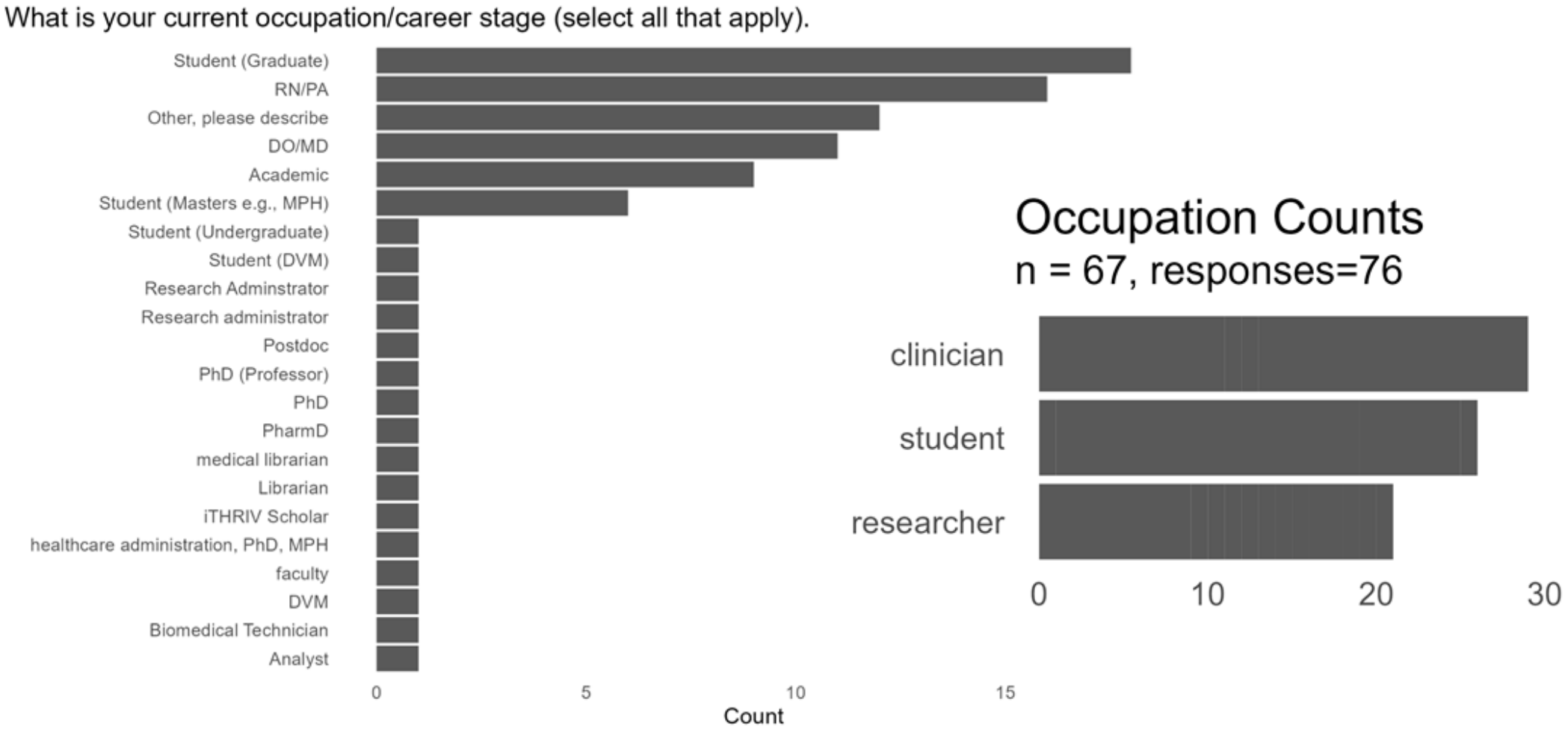
Creating Learner Personas

- Self-assessment survey (33 questions)
- Clustered to identify personas (23 Questions)
- 2 Waves (N=67): Summer 2020 (N=51) + Summer 2021

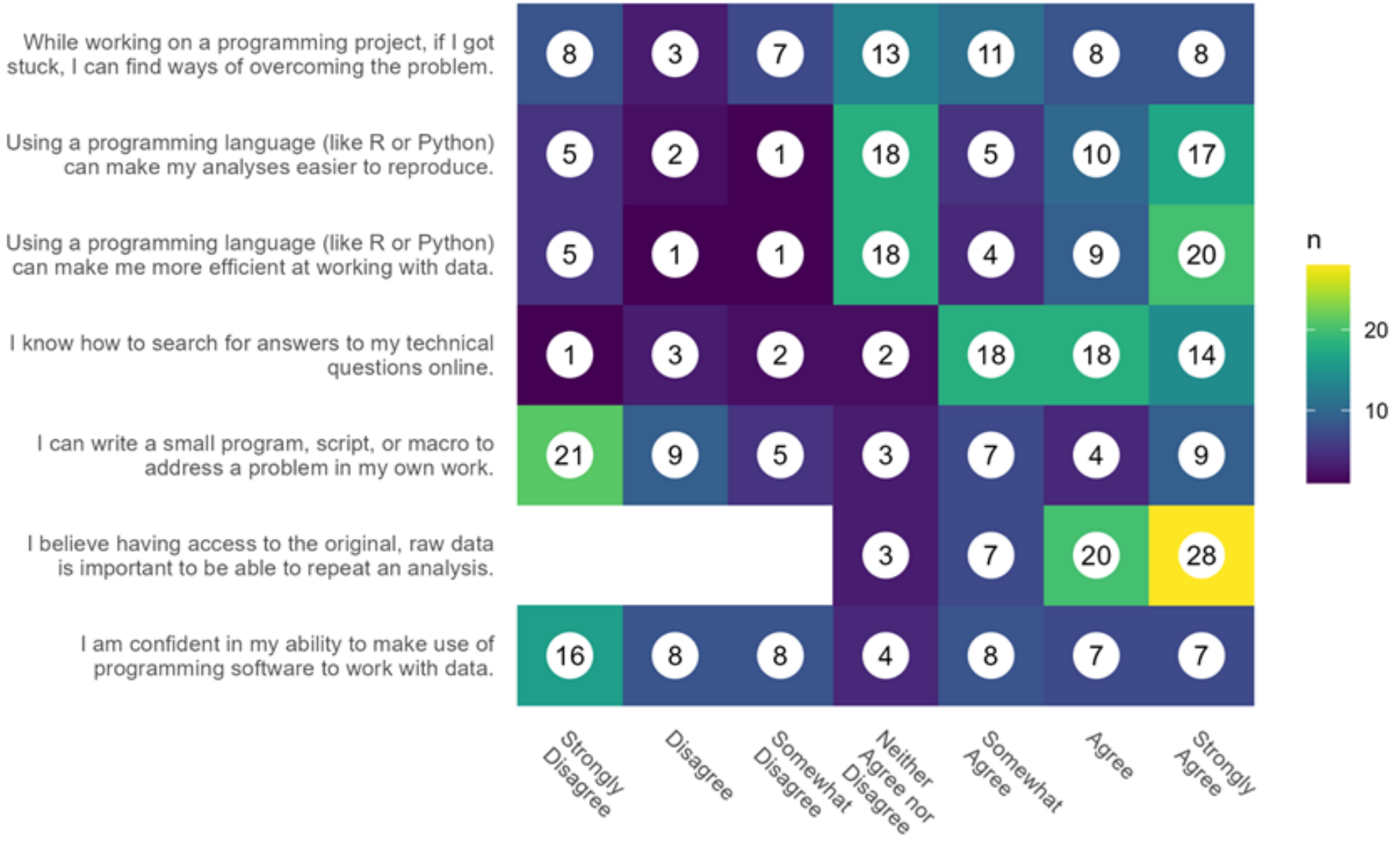
1. Demographics (6)
2. Programs Used in the Past (1)
3. *Programming Experience (6)
4. *Data Cleaning and Processing Experience (4)
5. *Project and Data Management (2)
6. *Statistics (4)
7. Workshop Framing and Motivation (3)
8. *Summary Likert (7)

- Ambrose SA, Bridges MW, DiPietro M, Lovett MC, Norman MK. How Learning Works: Seven Research-Based Principles for Smart Teaching. John Wiley & Sons; 2010.
- Jordan KL, Michonneau F. Analysis of The Carpentries Long-Term Surveys (April 2020). Zenodo; 2020. doi:10.5281/zenodo.3728205.
- Jordan K, Michonneau F, Weaver B. Analysis of Software and Data Carpentry's Pre- and Post-Workshop Surveys. Zenodo; 2018. doi:10.5281/zenodo.1325464.
- Wilson G. Teaching Tech Together: How to Make Your Lessons Work and Build a Teaching Community around Them. Taylor & Francis; 2019. http://teachtogether.tech

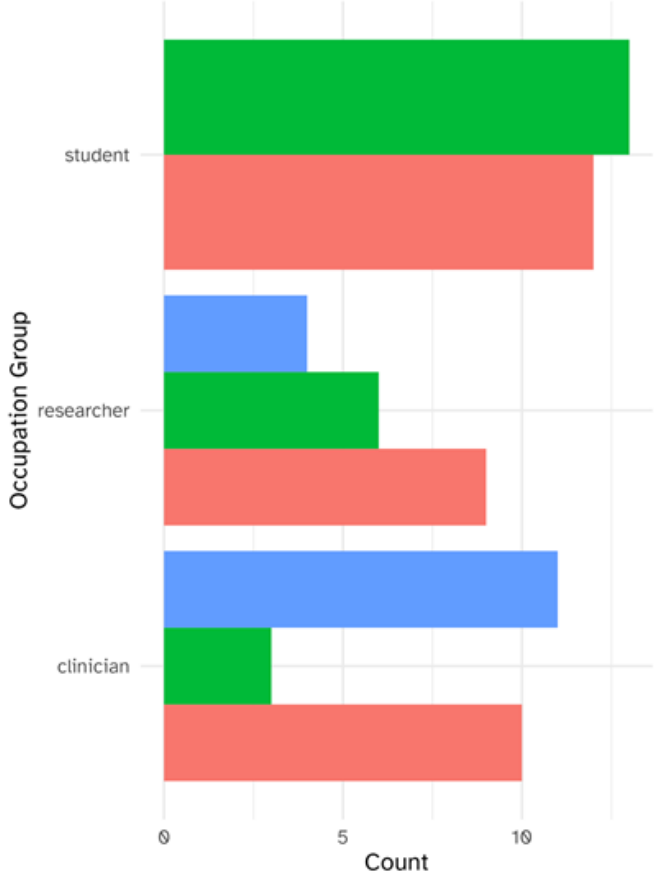
Ocupation


What is your current occupation/career stage (select all that apply).
Student (Graduate)
RN/PA
Other, please describe
DO/MD
Academic
Student (Masters e.g., MPH)
Student (Undergraduate)
Student (DVM)
Research Adminstrator
Research administrator
Postdoc
PhD (Professor)
PhD
PharmD
medical librarian
Librarian
iTHRIV Scholar
healthcare administration, PhD, MPH
faculty
DVM
Biomedical Technician
Analyst
0
5
10
15
Count
Occupation Counts
n = 67, responses=76
clinician
student
researcher
0
10
20
30

General Attitudes: Summary Likert (7)

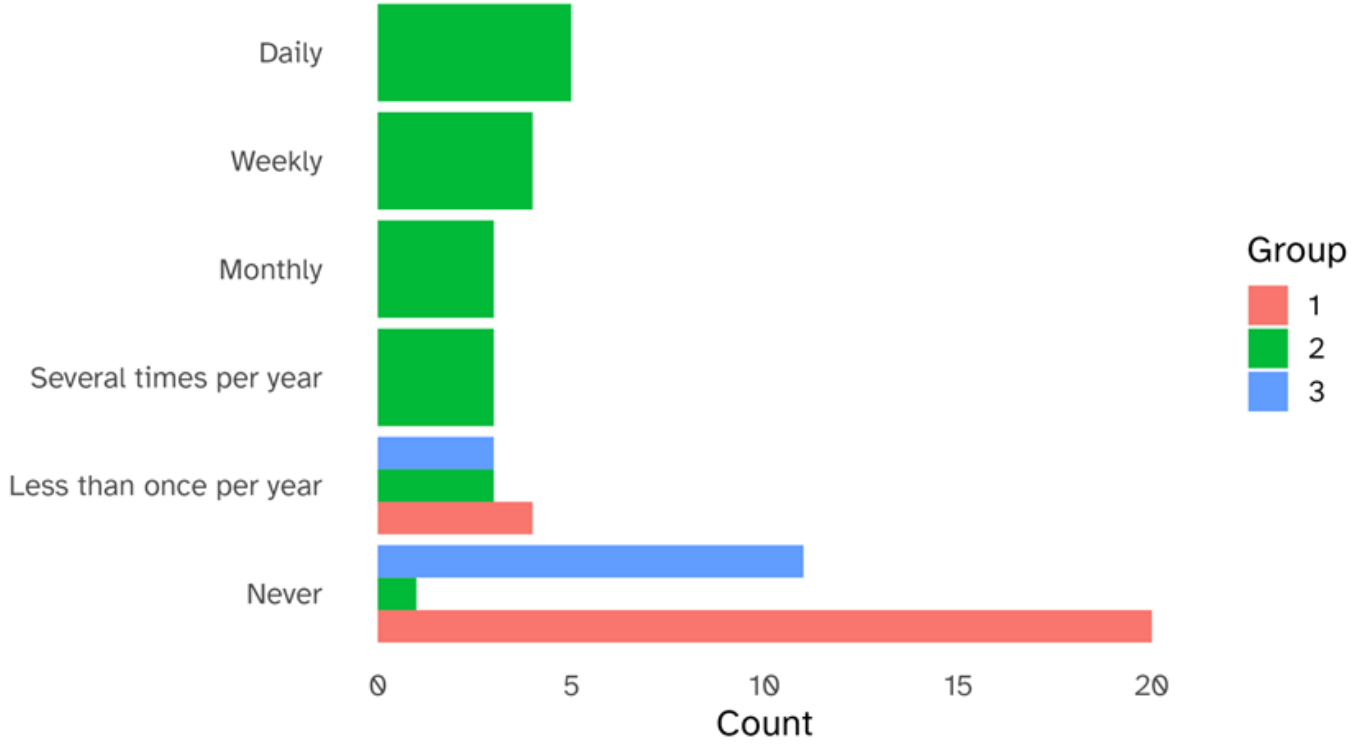


Identifying Personas: Programming Experience

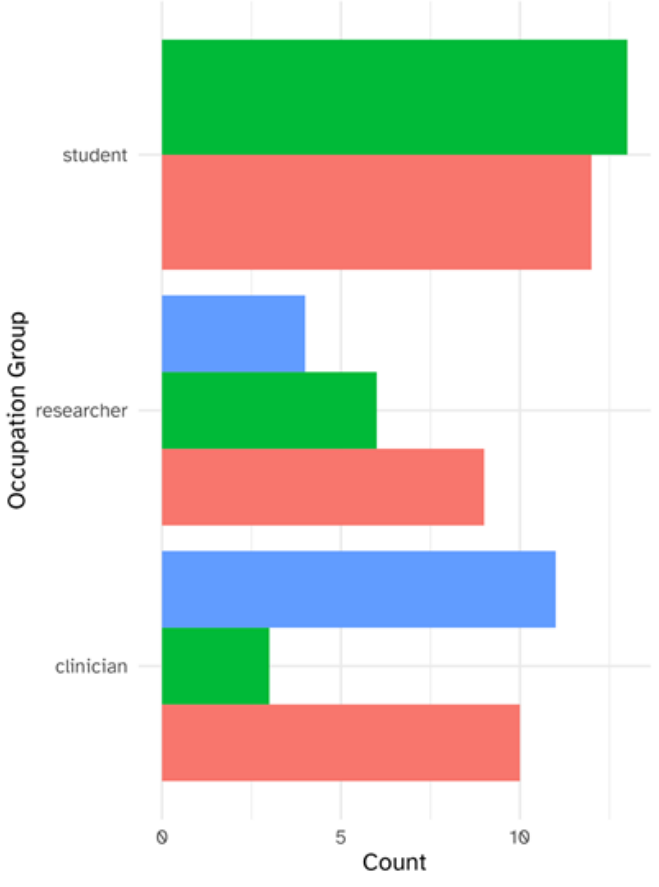


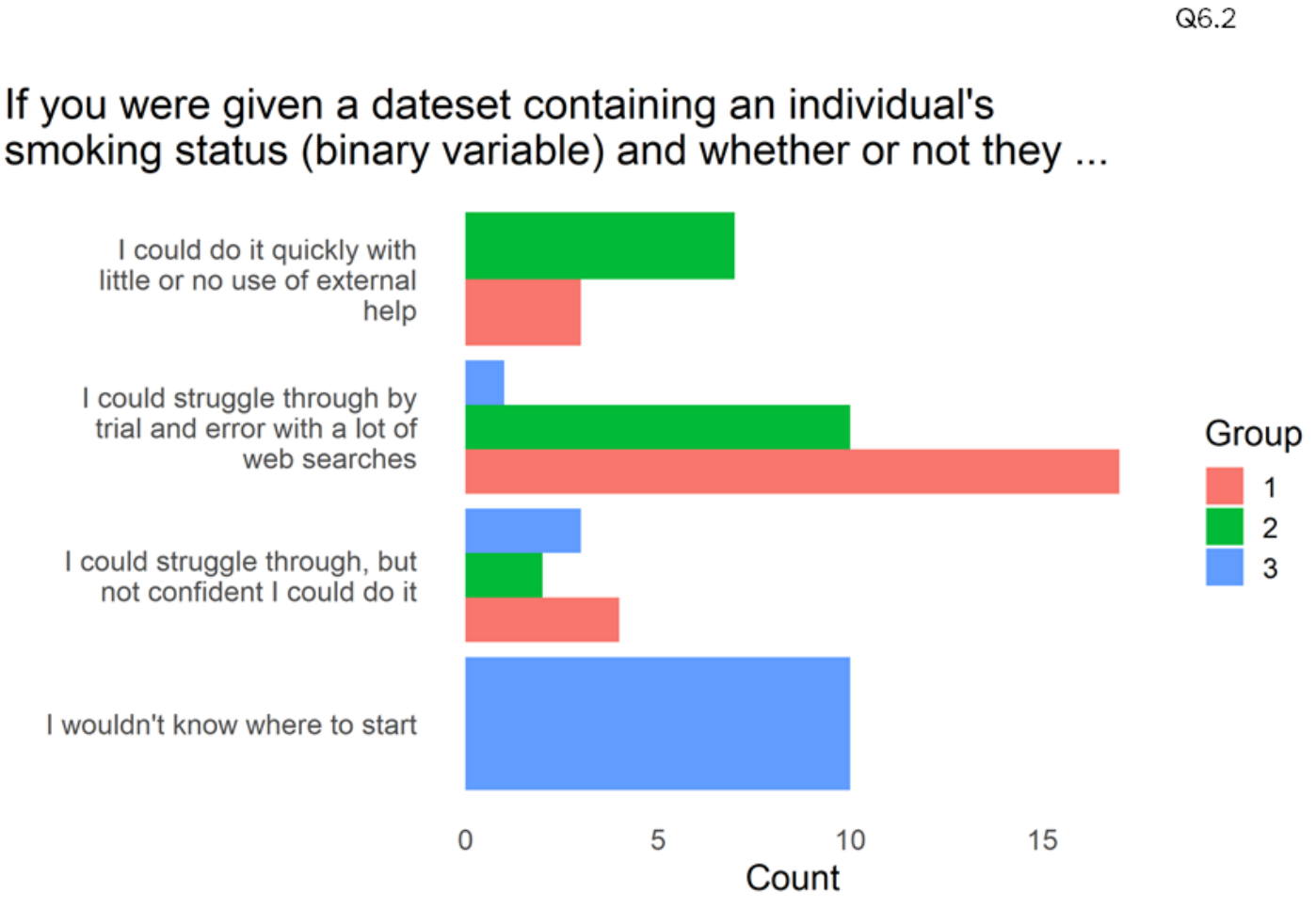
Q3.4

How often do you currently use programming languages (R, Python, etc.)?

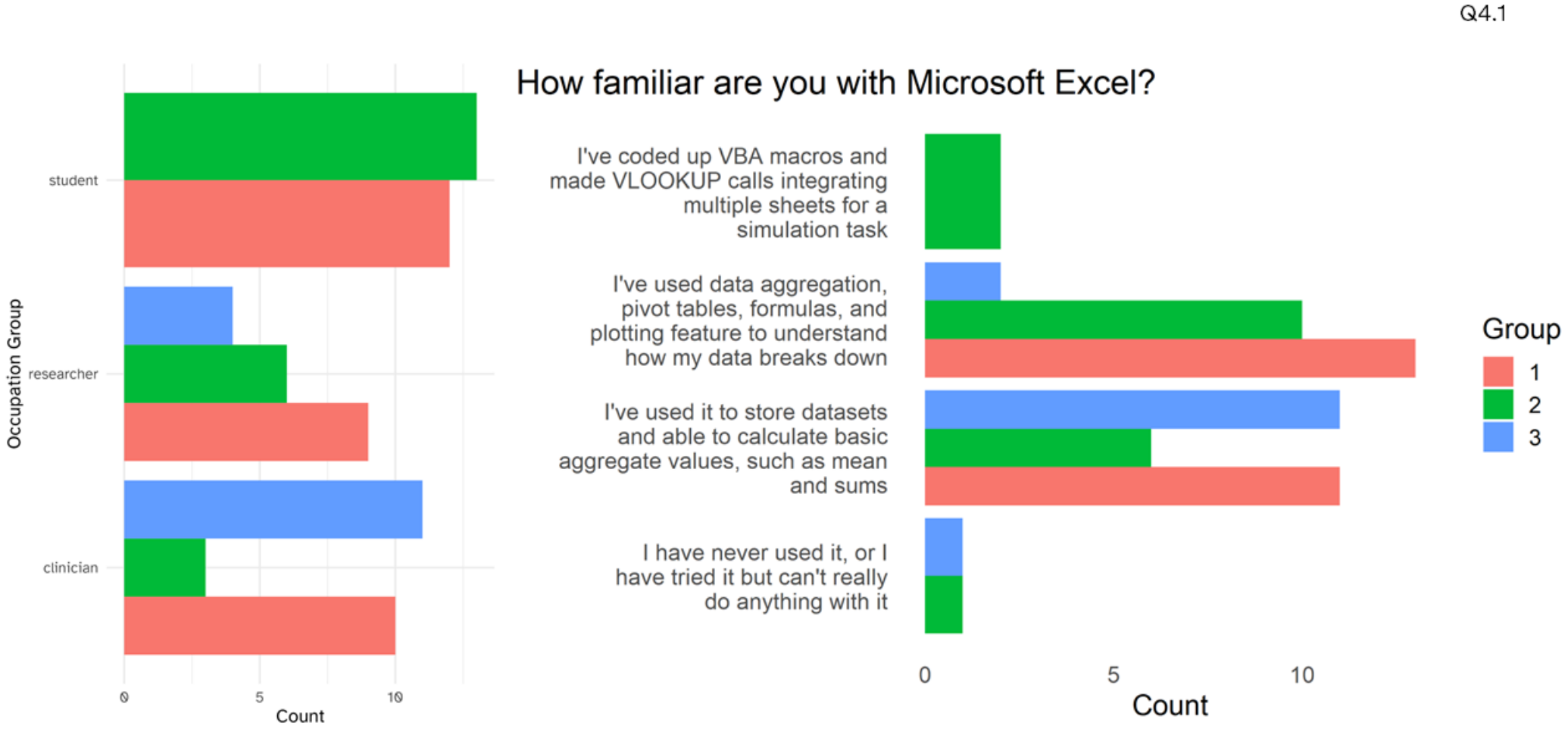


Identifying Personas: Statistics

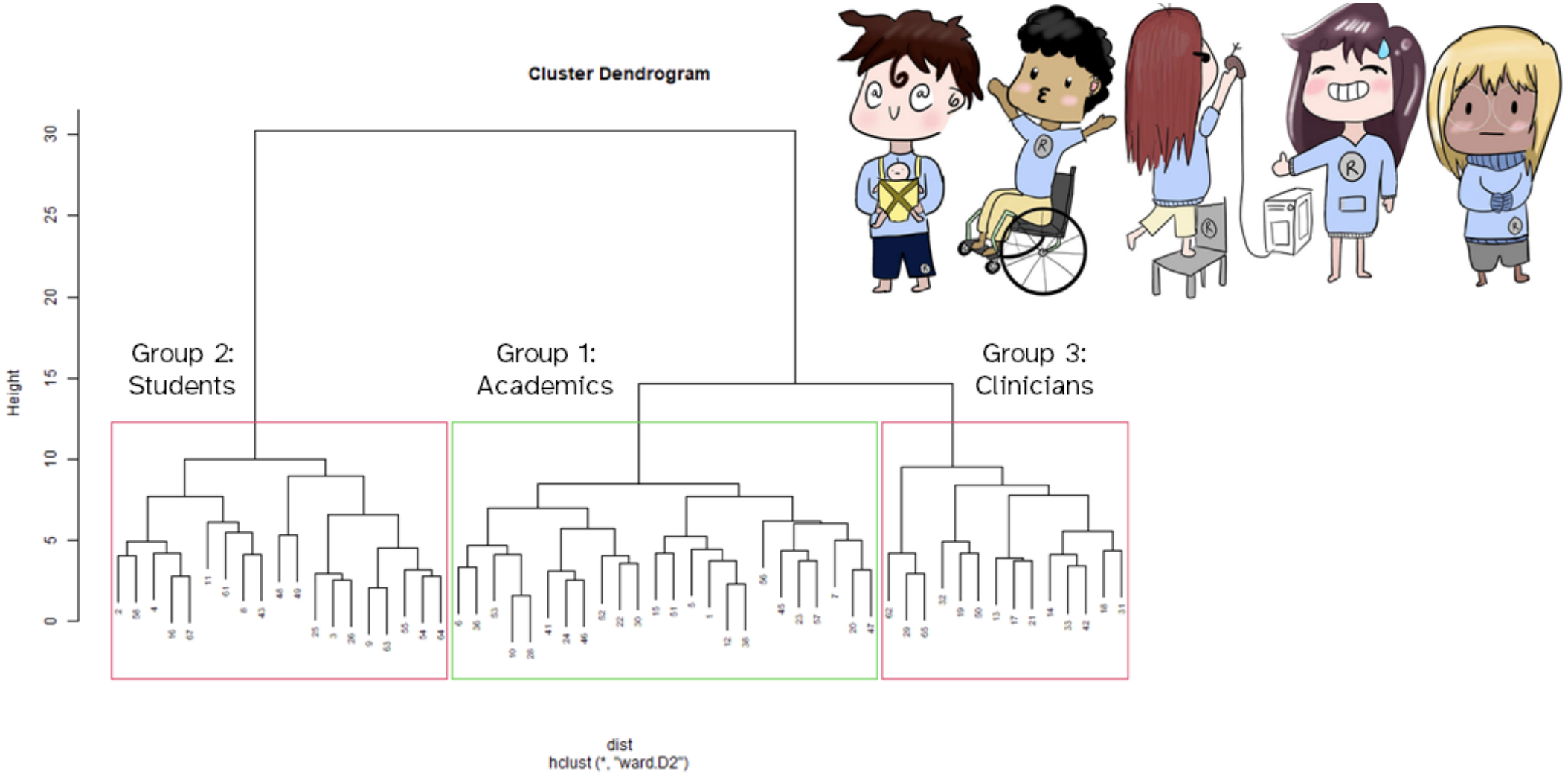




Identifying Personas: Excel



Hierarchical Clustering for Personas



Overall Persona Differences

1. Ash Academic
2. Samir Student
3. Clare Clinician

Group	Programming	Statistics	Data Programming
1	Low	Medium	Medium
2	High	High	High
3	Low	Low	Low

- `stats::hclust()` for clustering: https://github.com/chendaniely/dissertation-analysis/blob/master/analysis/030-persona/03-pca_clustering.Rmd#L191
- `stats:cutree()` for cutting the tree: https://github.com/chendaniely/dissertation-analysis/blob/master/analysis/030-persona/03-pca_clustering.Rmd#L222

Primary Target User

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.

- RStudio. Learner Personas. Published 2019. https://rstudio-education.github.io/learner-personas/

Assessing the Efficacy of Domain-Specific Data Science Curriculum in the Biomedical Sciences

How Learner Personas Can Guide Educational Needs in the Short-Term and Long-Term

Backward Design

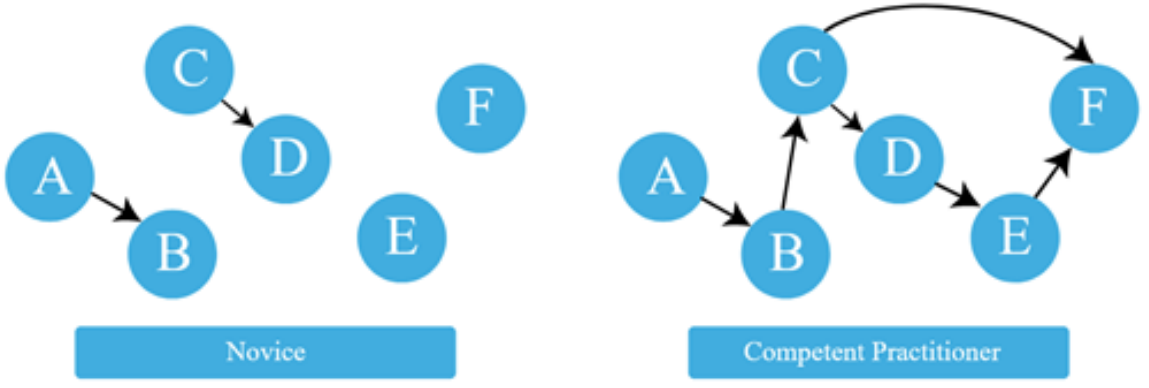
1. Identify your learners (learner persona)
2. Plan out your lesson content (concept maps)
3. Define overall goal (summative assessment)
4. Break down the goal (formative assessment)
5. Outline the the course
6. Write a summary of the course

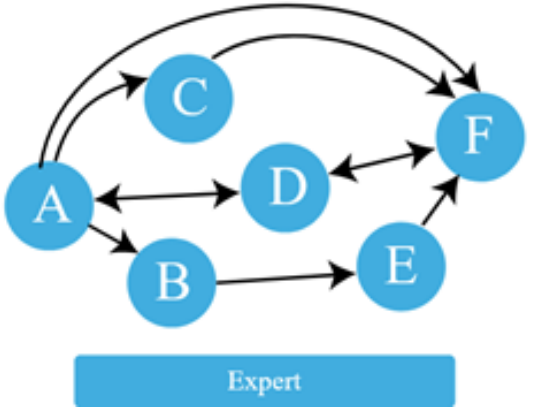
- Wilson G. Teaching Tech Together: How to Make Your Lessons Work and Build a Teaching Community around Them. Taylor & Francis; 2019. http://teachtogether.tech

Creating the Learning Materials

Managing Prior Knowledge

- Concept maps: graphic of a mental model
- Learner's prior knowledge can help or hinder learning





- Ambrose SA, Bridges MW, DiPietro M, Lovett MC, Norman MK. How Learning Works: Seven Research-Based Principles for Smart Teaching. John Wiley & Sons; 2010.
- Wilson G. Teaching Tech Together: How to Make Your Lessons Work and Build a Teaching Community around Them. Taylor & Francis; 2019. http://teachtogether.tech

Summative Assessment

ID	age	prior.radiation	aKIRs	cmv	donor_negative	donor_positive
<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<chr>	<chr>
1	61	0	1	1	recipient_positive	NA
2	62	1	5	0	recipient_negative	NA
3	63	0	3	0	NA	recipient_positive
4	33	1	2	0	recipient_positive	NA
5	54	0	6	0	NA	recipient_positive
6	55	0	2	1	NA	recipient_positive
7	67	0	1	0	NA	recipient_positive
8	51	0	2	0	NA	recipient_positive
9	44	1	2	1	NA	recipient_positive
10	59	0	4	0	recipient_negative	NA

ID	age	prior.radiation	aKIRs	cmv	donor_status	recipient_status
<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<chr>	<chr>
1	61	0	1	1	donor_negative	recipient_positive
2	62	1	5	0	donor_negative	recipient_negative
3	63	0	3	0	donor_positive	recipient_positive
4	33	1	2	0	donor_negative	recipient_positive
5	54	0	6	0	donor_positive	recipient_positive
6	55	0	2	1	donor_positive	recipient_positive
7	67	0	1	0	donor_positive	recipient_positive
8	51	0	2	0	donor_positive	recipient_positive
9	44	1	2	1	donor_positive	recipient_positive
10	59	0	4	0	donor_negative	recipient_negative

```r
library(tidyverse)
library(readxl)
library(writexl)

# load the excel sheet into R
cmv <- read_excel("data/cmv.xlsx")

# filter data
cmv_subset <- cmv %>%
  filter(age > 65)

# save out subset data
write_xlsx(cmv_subset, "data/cmv_subset.xlsx")

# tidy data
cmv_tidy <- cmv %>%
  pivot_longer(donor_negative:donor_positive,
               names_to = "donor_status",
               values_to = "recipient_status")

# plot data
ggplot(cmv_tidy, aes(x = age)) +
  geom_histogram()

# fit a model
mod <- glm(cmv ~ age + prior.radiation + donor_status,
           data = cmv_tidy,
           family = "binomial")
summary(mod)
```

R + Python

```python
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
import statsmodels.formula.api as smf

# load the excel sheet into Python
cmv = pd.read_excel("data/cmv.xlsx")

# filter data
cmv_subset = cmv.loc[cmv["age"] > 65]

# save out subset data
cmv_subset.to_excel("data/cmv_subset.xlsx")

# tidy data
cmv_tidy = (cmv
            .melt(id_vars=["ID", "age", "prior.radiation", "aKIRs", "cmv"],
                  var_name="donor_status",
                  value_name="recipient_status")
)

# plot data
sns.histplot(data=cmv_tidy, x="age")
plt.show()

# fit a model
cmv_tidy = cmv_tidy.rename(columns={"prior.radiation": "prior_radiation"})

mod = (smf
       .logit(formula="cmv ~ age + prior_radiation + donor_status",
              data=cmv_tidy)
       .fit()
)
mod.summary()
```

```r
library(tidyverse)
library(readxl)
library(writexl)

# load the excel sheet into R
cmv ← read_excel("data/cmv.xlsx")

# filter data
cmv_subset ← cmv %>%
  filter(age > 65)

# save out subset data
write_xlsx(cmv_subset, "data/cmv_subset.xlsx")

# tidy data
cmv_tidy ← cmv %>%
  pivot_longer(donor_negative:donor_positive,
               names_to = "donor_status",
               values_to = "recipient_status")

# plot data
ggplot(cmv_tidy, aes(x = age)) +
  geom_histogram()

# fit a model
mod ← glm(cmv ~ age + prior.radiation + donor_status,
          data = cmv_tidy,
          family = "binomial")
summary(mod)
```

Are the Materials Effective?

- Create the materials
- Test retest design
  - Pre, post, and long-term survey
- Workshop not classroom setting
- Assessment needs to be more flexible
  - Questions need to be broken down for learners
- Ask about confidence not objective assessment

- Jordan K. Data Carpentry Assessment Report: Analysis of Post-Workshop Survey Results. Zenodo; 2016. doi:10.5281/zenodo.165858
- Jordan K. Analysis of The Carpentries Long-Term Impact Survey. Zenodo; 2018. doi:10.5281/zenodo.1402200
- Jordan KL, Marwick B, Duckles J, Zimmerman N, Becker E. Analysis of Software Carpentry's Post-Workshop Surveys. Zenodo; 2017. doi:10.5281/zenodo.1043533
- Jordan KL, Marwick B, Weaver B, et al. Analysis of the Carpentries' Long-Term Feedback Survey. Zenodo; 2017. doi:10.5281/zenodo.1039944
- Jordan KL, Michonneau F. Analysis of The Carpentries Long-Term Surveys (April 2020). Zenodo; 2020. doi:10.5281/zenodo.3728205
- Jordan K, Michonneau F, Weaver B. Analysis of Software and Data Carpentry's Pre- and Post-Workshop Surveys. Zenodo; 2018. doi:10.5281/zenodo.1325464

Learning Objectives

- Name the features of a tidy/clean dataset
- Transform data for analysis
- Identify when spreadsheets are useful
- Assess when a task should not be done in a spreadsheet software
- Break down data processing into smaller individual (and more manageable) steps
- Construct a plot and table for exploratory data analysis
- Calculate, interpret, and communicate an appropriate statistical analysis of the data

```r
library(tidyverse)
library(readxl)
library(writexl)

# load the excel sheet into R
cmv <- read_excel("data/cmv.xlsx")

# filter data
cmv_subset <- cmv %>%
  filter(age > 65)

# save out subset data
write_xlsx(cmv_subset, "data/cmv_subset.xlsx")

# tidy data
cmv_tidy <- cmv %>%
  pivot_longer(donor_negative:donor_positive,
               names_to = "donor_status",
               values_to = "recipient_status")

# plot data
ggplot(cmv_tidy, aes(x = age)) +
  geom_histogram()

# fit a model
mod <- glm(cmv ~ age + prior.radiation + donor_status,
           data = cmv_tidy,
           family = "binomial")
summary(mod)
```

Create Data Science Learning Materials

https://ds4biomed.tech/

1. Introduction
2. Spreadsheets
3. R + RStudio
4. Load Data
5. Descriptive Calculations
6. Clean Data (Tidy)
7. Visualization (Intro)
8. Analysis Intro (Logistic)

ds4biomed: https://ds4biomed.tech/

Part I

1. Introduction
2. Spreadsheets
3. R + RStudio
4. Load Data
5. Descriptive Calculations
6. Clean Data (Tidy)
7. Visualization (Intro)
8. Analysis Intro (Logistic)

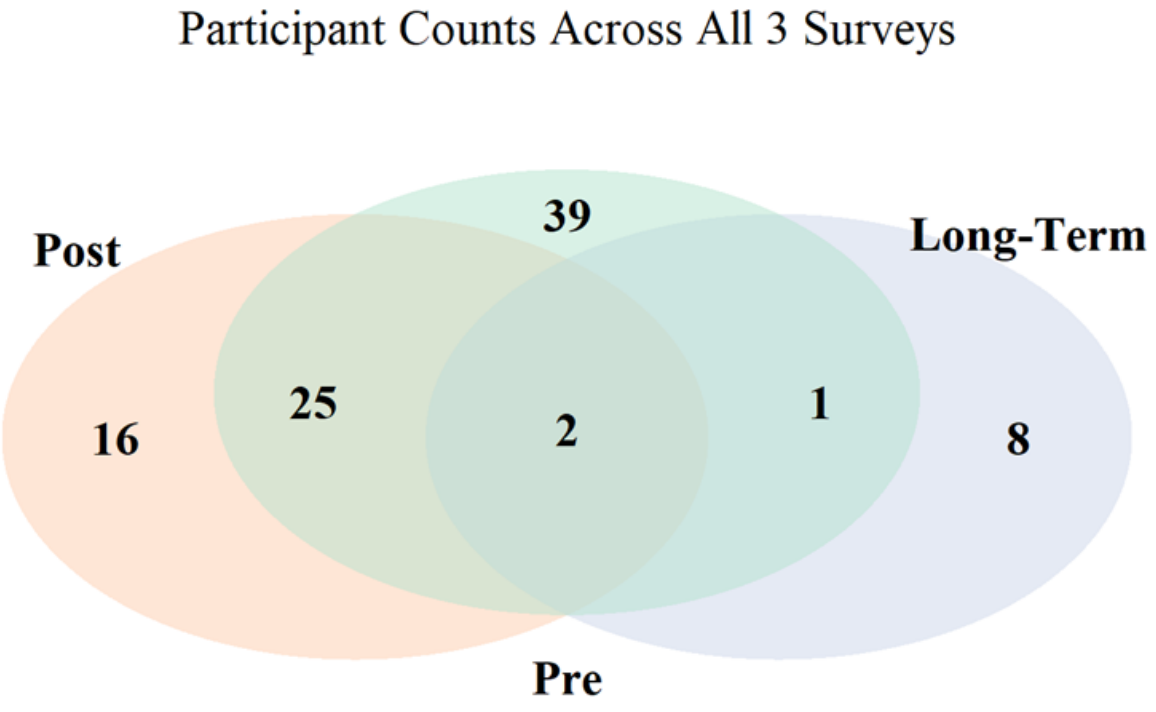
Part II

1. 30-Day Re-admittance
2. Working with multiple datasets
3. APIs
4. Functions
5. Survival Analysis
6. Machine Learning Intro

Assessing Workshop Effacy

Workshop Attendees

- 8 Workshops
  - 200 Attendees across 2 days
- 91 Responses
  - 67 Pre-workshop
  - 43 Post-workshop
  - 11 Long-term

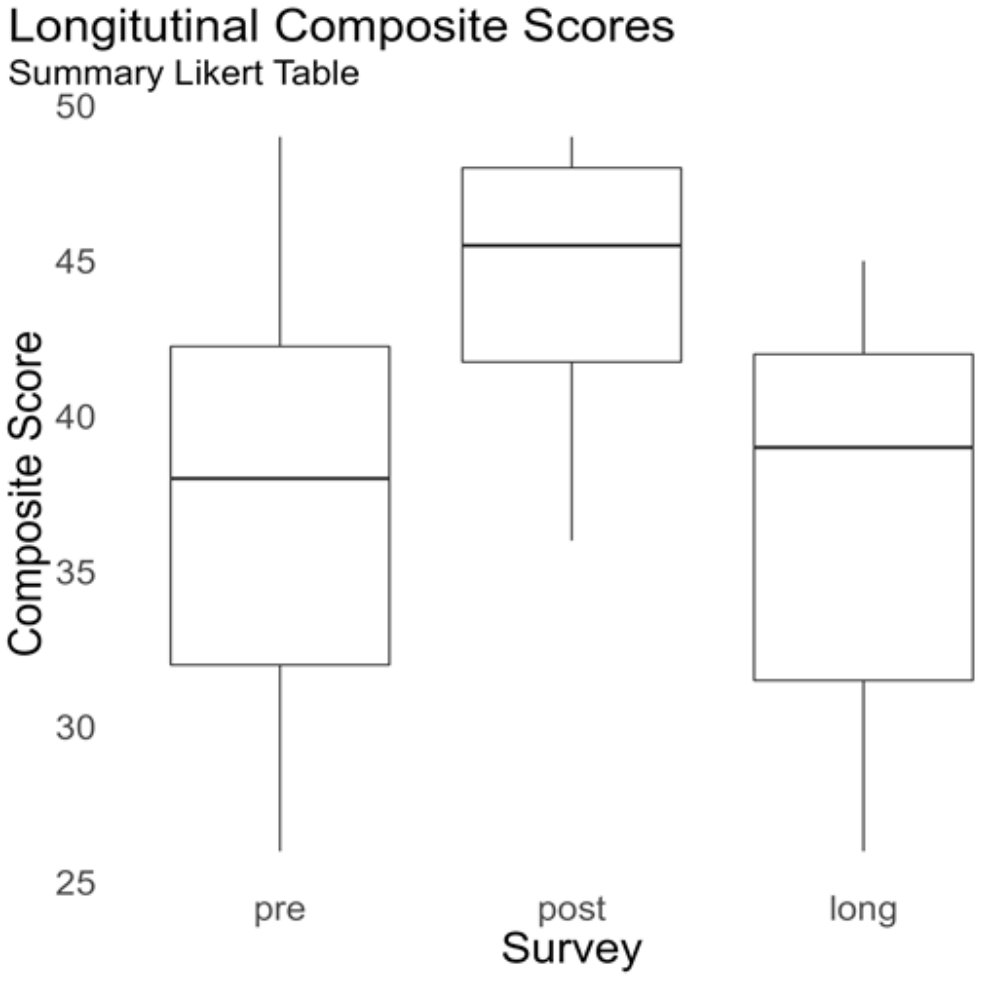


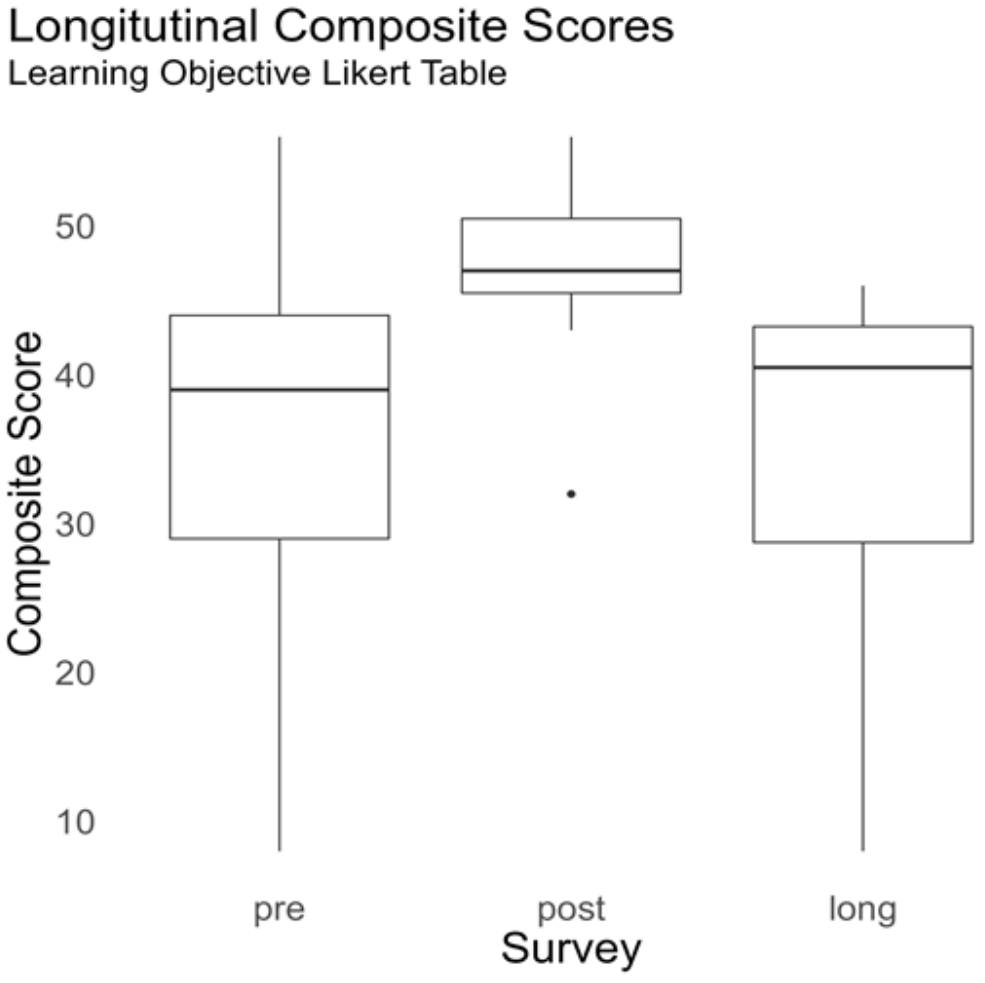
Learning Objectives

- Name the features of a tidy/clean dataset
- Transform data for analysis
- Identify when spreadsheets are useful
- Assess when a task should not be done in a spreadsheet software
- Break down data processing into smaller individual (and more manageable) steps
- Construct a plot and table for exploratory data analysis
- Calculate, interpret, and communicate an appropriate statistical analysis of the data

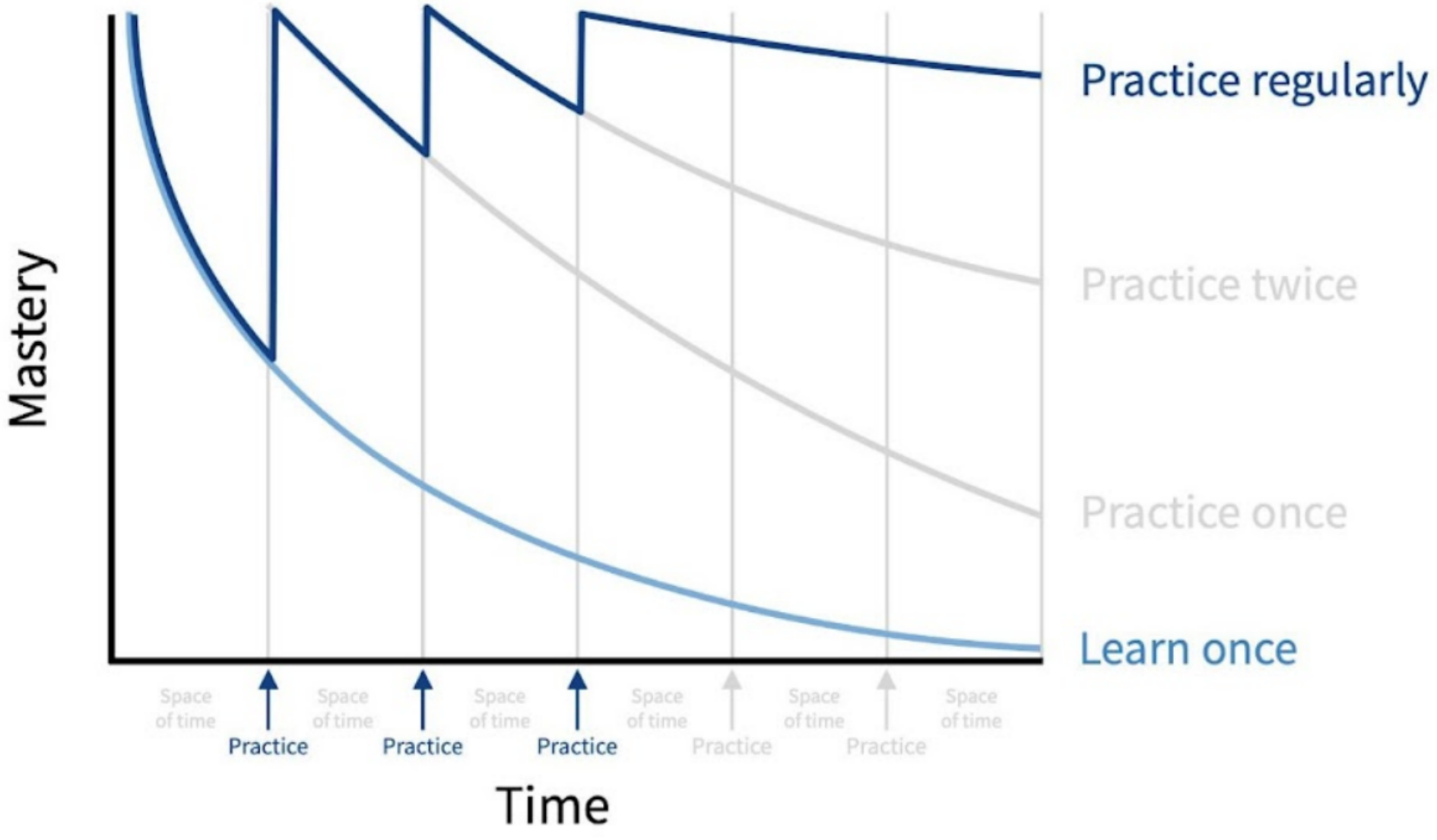
```r
library(tidyverse)
library(readxl)
library(writexl)

# load the excel sheet into R
cmv <- read_excel("data/cmv.xlsx")

# filter data
cmv_subset <- cmv %>%
  filter(age > 65)

# save out subset data
write_xlsx(cmv_subset, "data/cmv_subset.xlsx")

# tidy data
cmv_tidy <- cmv %>%
  pivot_longer(donor_negative:donor_positive,
               names_to = "donor_status",
               values_to = "recipient_status")

# plot data
ggplot(cmv_tidy, aes(x = age)) +
  geom_histogram()

# fit a model
mod <- glm(cmv ~ age + prior.radiation + donor_status,
           data = cmv_tidy,
           family = "binomial")
summary(mod)
```

Pre-Post-Long Composite

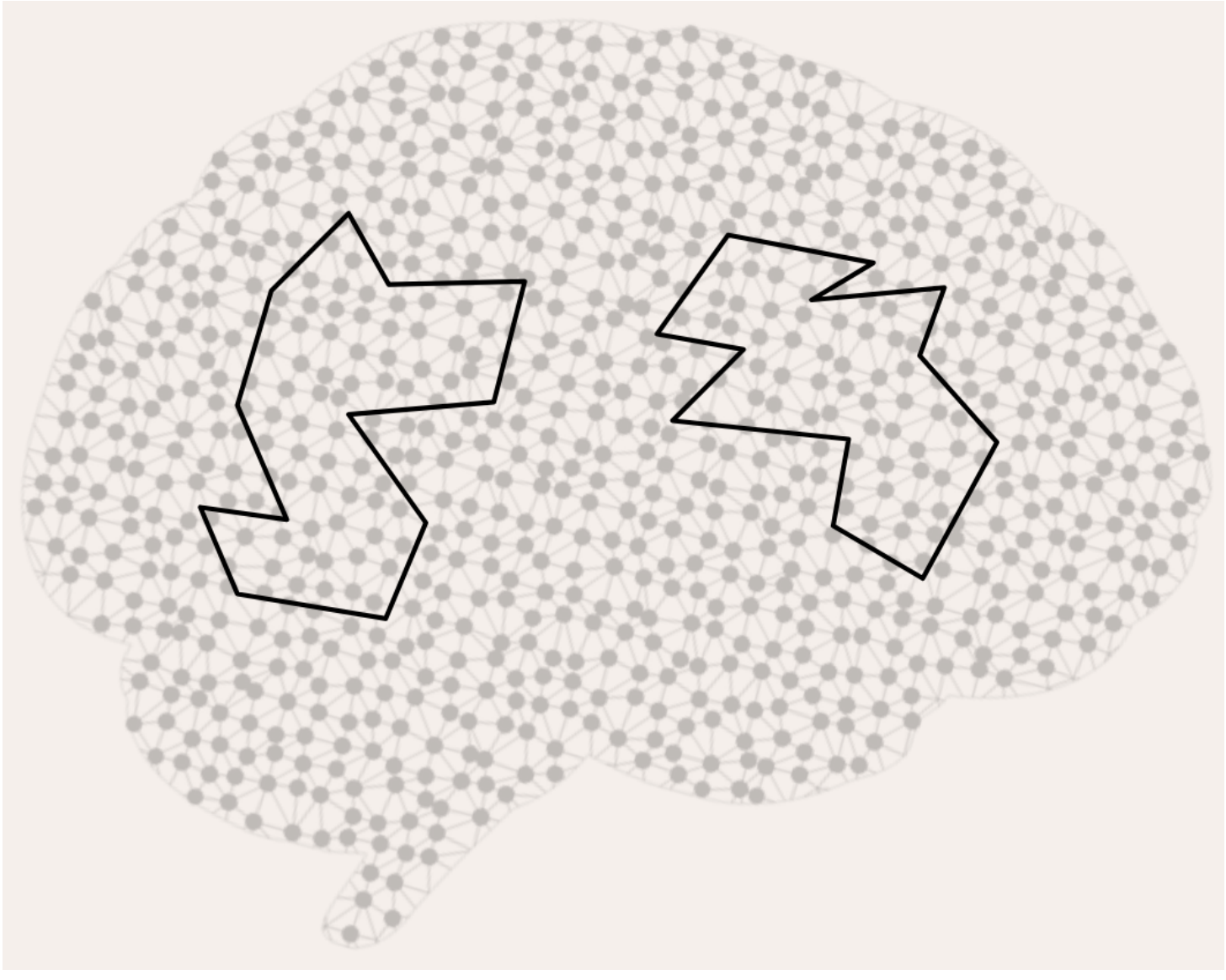




The Forgetting Curve

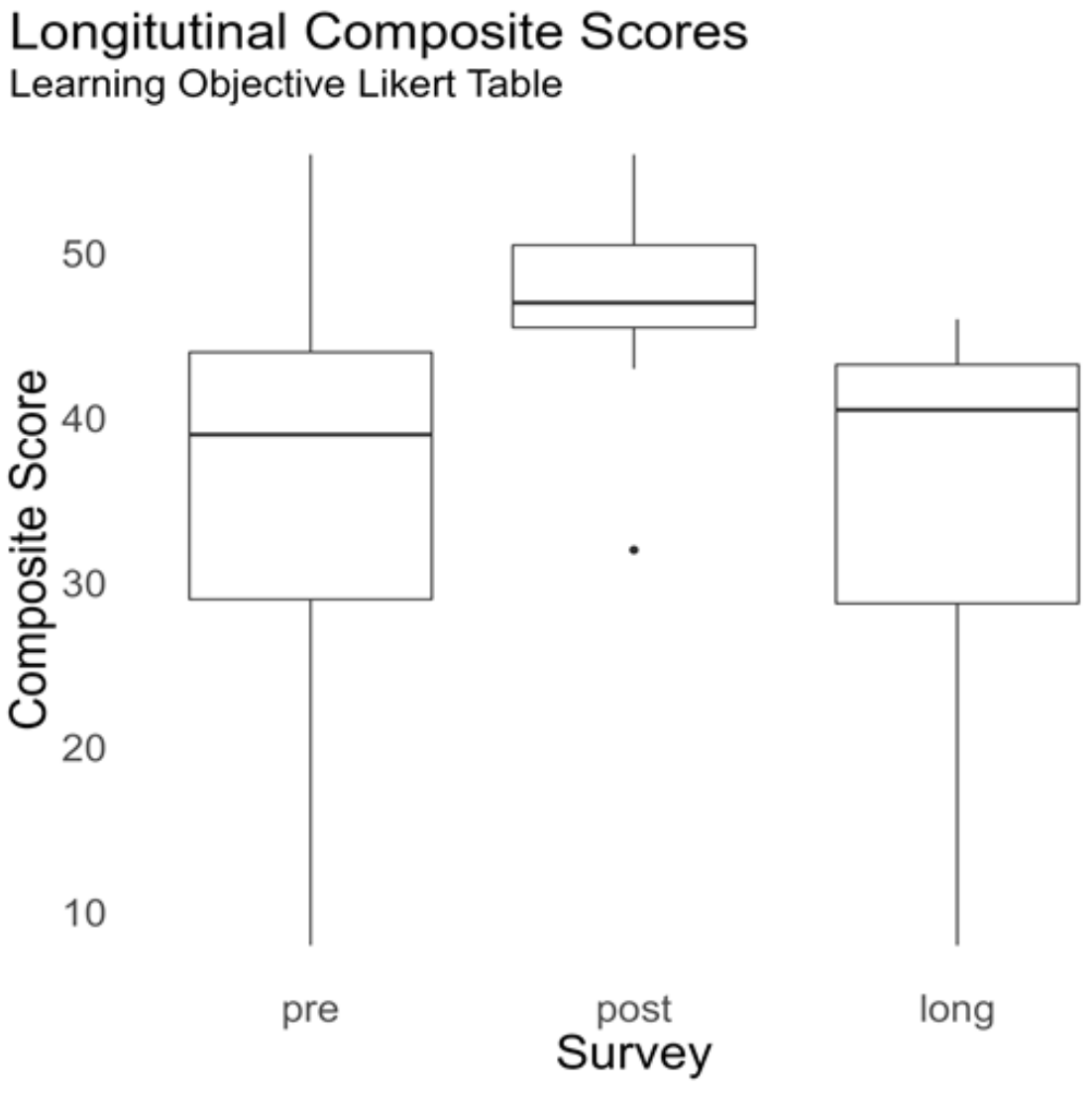


“Neurons that fire together wire together”



Overall Conclusions

- Objective way of backward design lesson development
- Domain-specific workshops seem beneficial to meet learning objectives
- Data science have different set of programming skills
- Long-term learning is more important
- Formative + summative assessments in long-term learning
- “10,000 hour rule”, “deliberate practice”, “forgetting curve”



- Malcolm Gladwell: 10,000 Hour Rule
- László and Klara Polgár: deliberate practice
- Hermann Ebbinghaus: forgetting curve

Communities (of Practice)

- The Carpentires
- r/medicine (slack), r/pharma
- Tidy Tuesday*
- R-Ladies: https://rladies.org/
- Py-Ladies: https://pyladies.com/
- R4DS Community (slack): r4ds.io/join
- Nursing & Data Science Collaboratory (slack)
- OHDSI (MS Teams)
- Observational Health Data Sciences and Informatics

- Shrestha N, Barik T, Parnin C. Remote, but Connected: How #TidyTuesday Provides an Online Community of Practice for Data Scientists. Proc ACM Hum-Comput Interact. 2021;5(CSCW1):52:1-52:31. doi:10.1145/3449126

Teaching Tech Together: The Rules

1. **Be kind: all else is details.**
2. Remember that you are not your learners…
3. …that most people would rather fail than change…
4. …and that ninety percent of magic consists of knowing one extra thing.
5. Never teach alone.
6. Never hesitate to sacrifice truth for clarity.
7. Make every mistake a lesson.
8. Remember that no lesson survives first contact with learners…
9. …that every lesson is too short for the teacher and too long for the learner…
10. …and that nobody will be more excited about the lesson than you are.

- Wilson G. Teaching Tech Together: How to Make Your Lessons Work and Build a Teaching Community around Them. Taylor & Francis; 2019. http://teachtogether.tech