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

**Preliminary Exam for Daniel Chen**

February 19, 2021

UNIVERSITY LIBRARIES
VIRGINIA TECH

# *Goal: Create and assess a data science curriculum*

Domain specific: biomedical sciences

1. **Who** are our learners
2. Are the learning materials **effective**
3. Tie in computer science **pedagogy**

# Current status of data science education

- Joint departments

**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 |

(Song and Zhu, 2016)

# *Current status of data science education*

- Joint departments

- Probability + Statistics
- Data Mining
- **Programming**

**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 |

(Song and Zhu, 2016)

# *Current status of data science education*

- Joint departments

- Probability + Statistics
- Data Mining
- **Programming**

- Exploratory Data Analysis
- **Database**
- Data Mining

**Table 3:** *Core courses in master's programmes (as of August 2014)*

| *Course* | *No. of universities offering the course* |
|---|---|
| Exploratory Data Analysis | 10 |
| Database | 10 |
| Data Mining | 9 |
| Data Visualization | 8 |
| Statistical Modelling | 8 |
| Machine Learning | 6 |
| Information Retrieval | 5 |
| Information and Social Network Analysis | 4 |
| Data Warehouse | 4 |
| Introduction to Data Science | 3 |
| Research Methods | 3 |
| Social Aspects of Data Science | 3 |
| Algorithms | 2 |
| Data Cleaning | 2 |
| Text Mining | 2 |
| Healthcare Analytics | 2 |

(Song and Zhu, 2016)

# *Most people are missing "data" classes*

- As **data science** education becomes a **commodity**
- Content is **not** an issue

- Learning platforms
- **Domain experts** to help navigate learners improve **data literacy**
- **Changing** open source **landscape**

**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 |

(Kross et al., 2020)

# *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 et al., 2020)

# *Why Domain Specificity?*

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

(Ambrose et al., 2010; Wilson, 2019)

# 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<br>•Connect NIH data systems | •Modernize data repository ecosystem<br>•Support storage and sharing of individual datasets<br>•Better integrate clinical and observational data into biomedical data science | •Support useful, generalizable, and accessible tools and workflows<br>•Broaden utility of and access to specialized tools<br>•Improve discovery and cataloging resources | •Enhance the NIH data-science workforce<br>•Expand the national research workforce<br>•Engage a broader community | •Develop policies for a FAIR data ecosystem<br>•Enhance stewardship |

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

(National Institutes of Health, 2020)

# *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

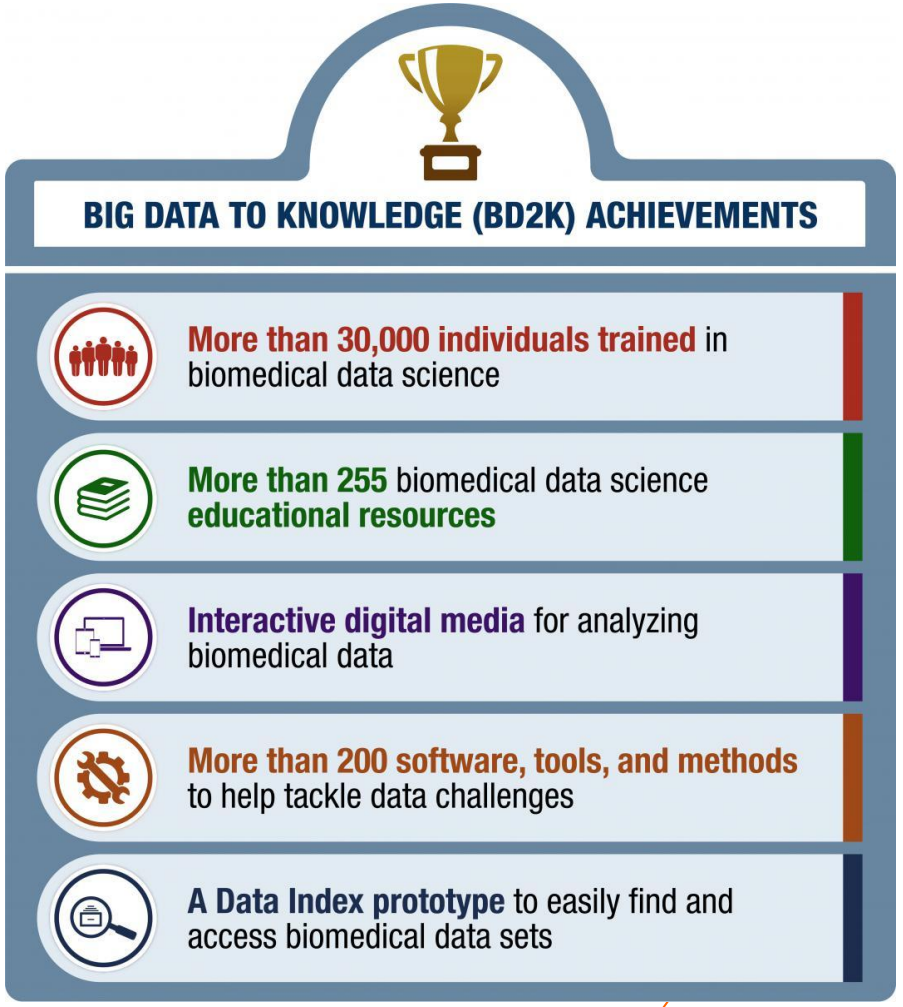
- **Develop a research workforce**



(National Institutes of Health, 2020)

# NIH The Big Data to Knowledge (BD2K)

- 2013 - 2018
- Narrow the gap in biomedical data science skills
- Train and educate workforce on analytical skills
- Main issues
  - Not maintained
  - No programming component
  - Not focused around data literacy




(Dunn and Bourne, 2017;
National Institutes of Health, 2013)

# *Computer Science Education*

- Adjacent to our goals
- ACM / IEEE curriculum recommendations
  - Computer engineering
  - Computer Science
  - Information System
  - Information Technology and Software Engineering

Difference

- “DataFrame” objects are not standard computer science data structures

# Data Carpentry

THE CARPENTRIES

*"Fundamental data skills needed to conduct research..."*

**Core Curriculum Materials**

Ecology curriculum
Genomics curriculum
Social Sciences curriculum
Geospatial data curriculum

**Under development / consideration**

Image Processing curriculum
Economics curriculum
Astronomy curriculum
Digital humanities curriculum
Other curricula

**Community Developed Lessons**

Carpentries Incubator
Carpentries Lab

**Data Science For Practicing Clinicians**

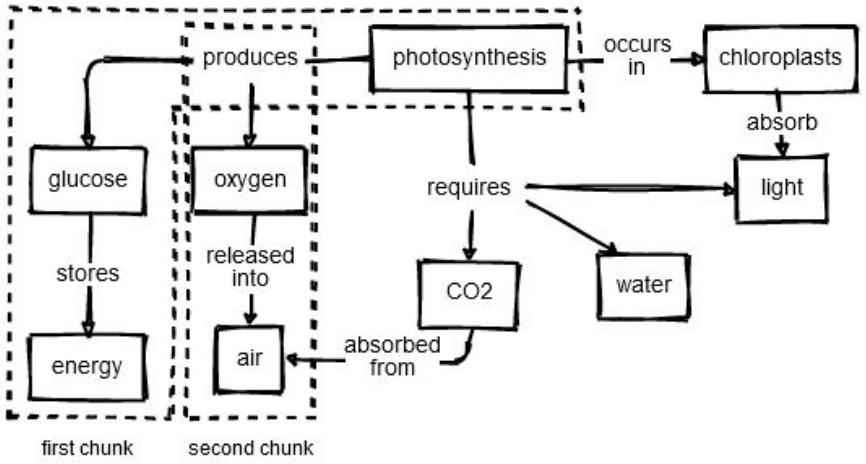
**There still is a need:**

1. Carpentries Instructor class at the Network of the National Library of Medicine (**NNLM**) March 2 - 5, 2021
2. Leverage the community to build a community to **maintain**

# *Who Are Our Learners?*

- AMA has a movement to change the medical curriculum (American Medical Association, 2021)
- What do we know about their prior knowledge?
  (Ambrose et al., 2010; Koch and Wilson, 2016; Wilson, 2019)

  - Concept maps (Wilson, 2019)




**Using concept maps in lesson design**

  - Dreyfus model of skill acquisition
    (Koch and Wilson, 2016; Dreyfus and Dreyfus, 1980)

# *How to Identify Our Learners?*

- Assessment tools
- Create learner personas

# *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



(K. Jordan, 2016, 2018; K. Jordan et al., 2018; K. L. Jordan, Marwick, Duckles, et al., 2017; K. L. Jordan, Marwick, Weaver, et al., 2017; K. L. Jordan and Michonneau, 2020)

# *Tie in Computer Science Pedagogy*

- The topics and ordering are not the same
- The teaching methodology applies
- Backwards-design testing of assessment question types
  - Faded examples (will add Parson's)
  - Prevents "empty page" syndrome
- Formative assessment question types
  - Not summative assessment

# *Hypothesis & Specific Aims*

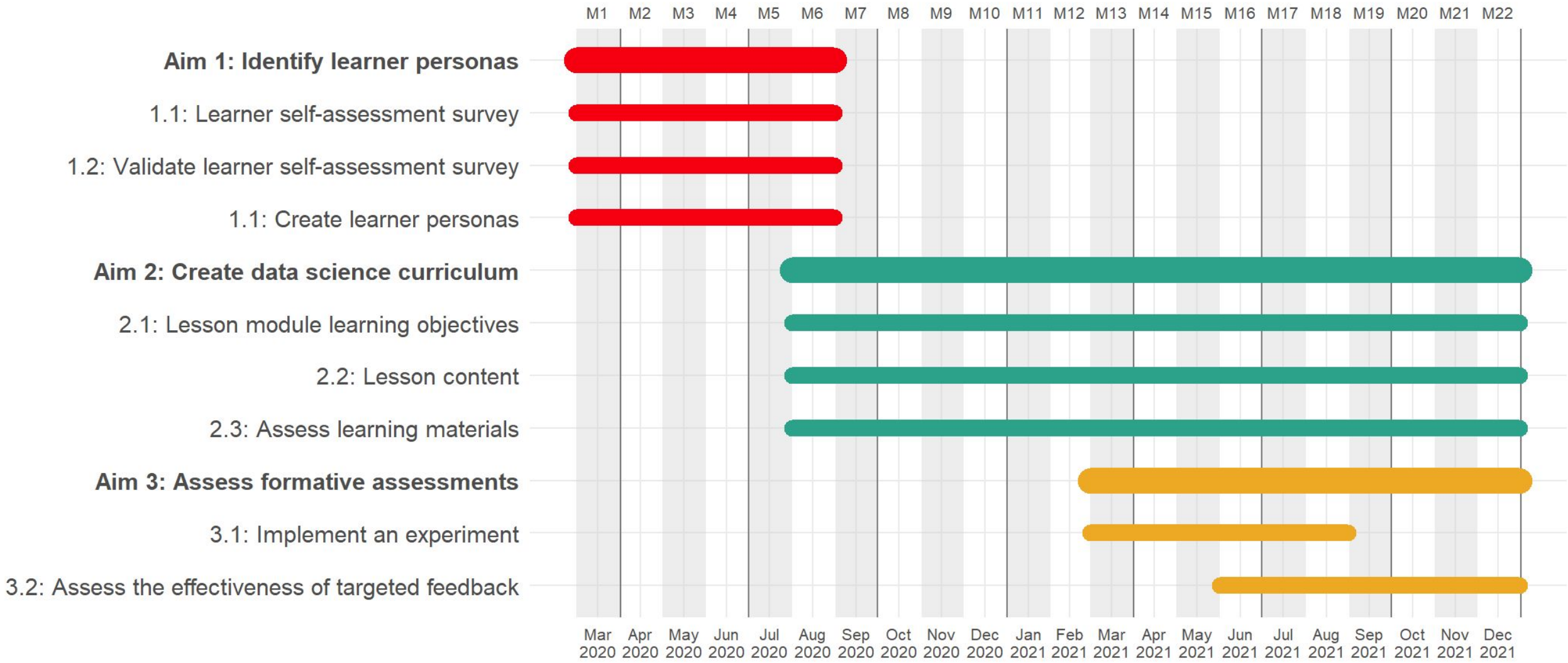
- ***Learning materials** around **tidy data principles** will help learners incorporate programming and data science skills from their **spreadsheet workflows.***
- NIH Data Management, Analytics, Tools + Workforce Development

**Aim 1:** Identify learner personas in the biomedical sciences by creating and validating learner self-assessment surveys.

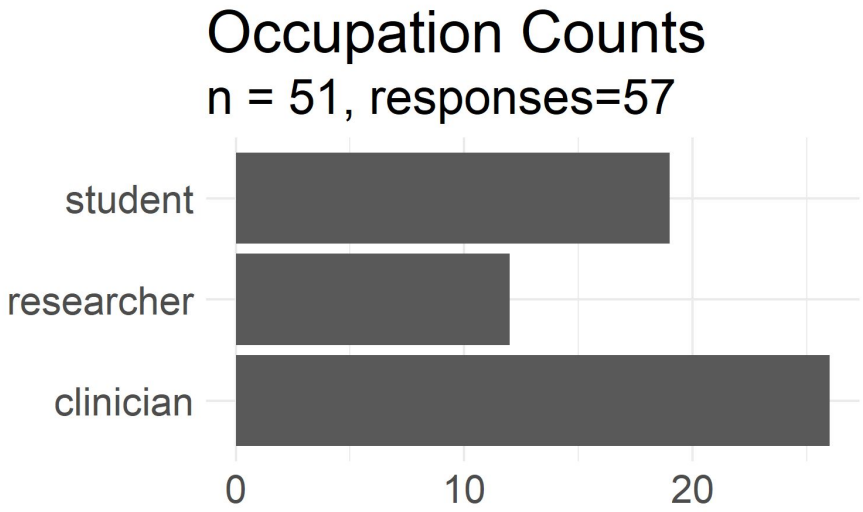
**Aim 2:** Create an effective data science for biomedical science curriculum based on best education and pedagogy practices.

**Aim 3:** Assess the effectiveness of formative assessments in learning objectives.

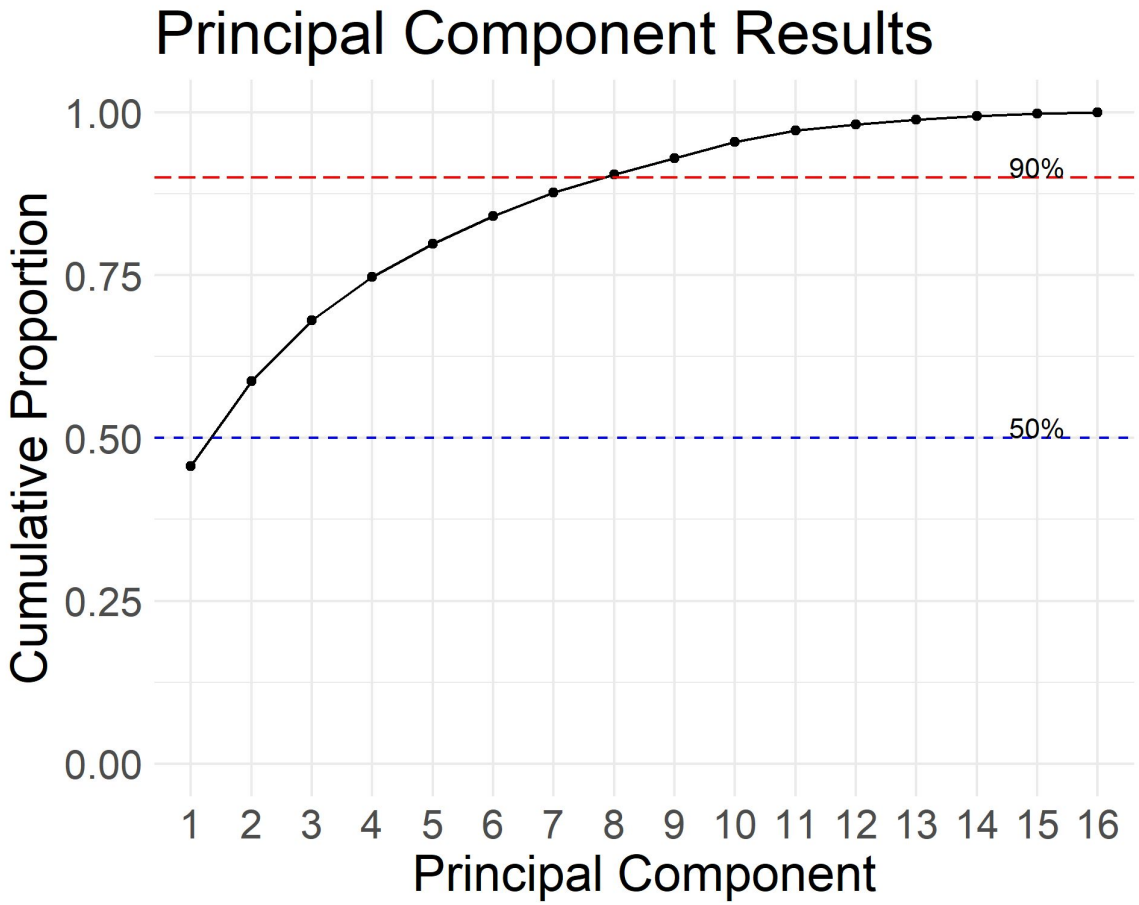
# Timeline


M1 M2 M3 M4 M5 M6 M7 M8 M9 M10 M11 M12 M13 M14 M15 M16 M17 M18 M19 M20 M21 M22
Aim 1: Identify learner personas
1.1: Learner self-assessment survey
1.2: Validate learner self-assessment survey
1.1: Create learner personas
Aim 2: Create data science curriculum
2.1: Lesson module learning objectives
2.2: Lesson content
2.3: Assess learning materials
Aim 3: Assess formative assessments
3.1: Implement an experiment
3.2: Assess the effectiveness of targeted feedback
Mar 2020 Apr 2020 May 2020 Jun 2020 Jul 2020 Aug 2020 Sep 2020 Oct 2020 Nov 2020 Dec 2020 Jan 2021 Feb 2021 Mar 2021 Apr 2021 May 2021 Jun 2021 Jul 2021 Aug 2021 Sep 2021 Oct 2021 Nov 2021 Dec 2021

# *Aim 1: Identify Learner Personas*




45 responses were used for the clustering due to missing responses



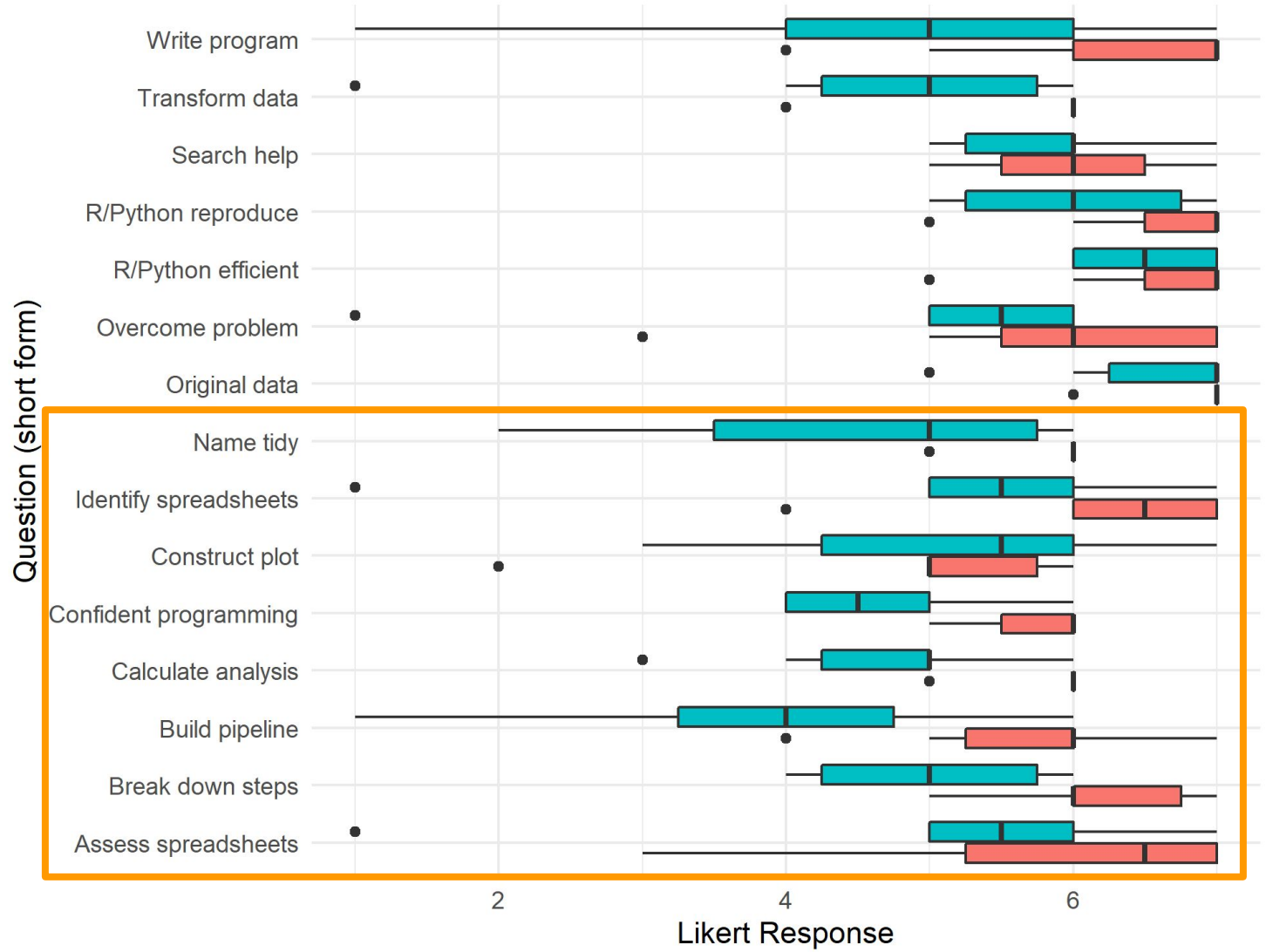

1. Alex Academic
2. Clare Clinician
3. Patricia Programmer
4. Samir Student

# *Aim 2: Create Data Science Curriculum*

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)

# Pre-Post Workshop Changes




Most participants have more confidence in their programming and data abilities.
N = 10 (pre = 5; post = 5)

Pre-workshop
Post-workshop

# *Wilcoxon Rank Sum Test*

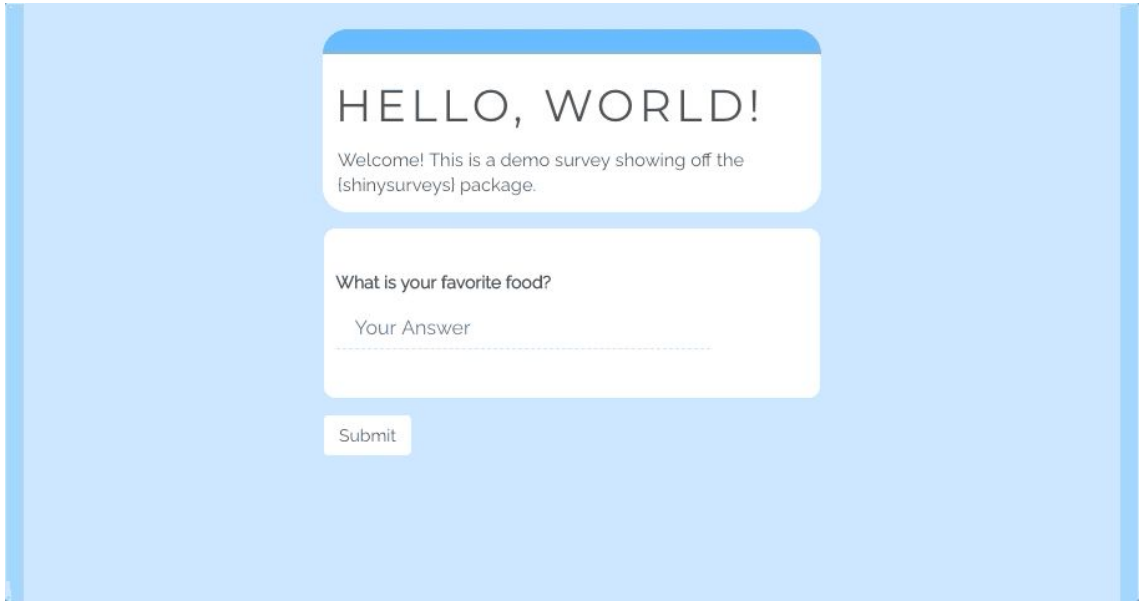
- Non-parametric alternative to paired t-test
- We can conclude that the median confidence of tasks before the workshop is significantly different from the median confidence of tasks after the workshop
- Two-sided
- Unpaired

| | Short question | V | p.value |
|---|---|---|---|
| 1 | Original data | 0 | 0.000102 |
| 2 | Write program | 2.5 | 0.000252 |
| 3 | Search help | 0 | 0.000130 |
| 4 | Overcome problem | 2.5 | 0.000271 |
| 5 | Confident programming | 0 | 0.000114 |
| 6 | R/Python reproduce | 0 | 0.000116 |
| 7 | R/Python efficient | 0 | 0.0000998 |
| 8 | Name tidy | 0 | 0.000116 |
| 9 | Transform data | 2.5 | 0.000232 |
| 10 | Identify spreadsheets | 2.5 | 0.000263 |
| 11 | Assess spreadsheets | 2.5 | 0.000266 |
| 12 | Break down steps | 0 | 0.000133 |
| 13 | Construct plot | 0 | 0.000133 |
| 14 | Build pipeline | 2.5 | 0.000268 |
| 15 | Calculate analysis | 0 | 0.000125 |

N = 10 (pre = 5; post = 5)
Dropped 1 observation that was paired

# *Aim 3: Assess Formative Assessments*

{shinysurveys}, {gradethis}, Data Science in a Box




Here is a number. You can do great things with it, like this:

- Take the square root of the log of the number `2`.

Then click Submit Answer.

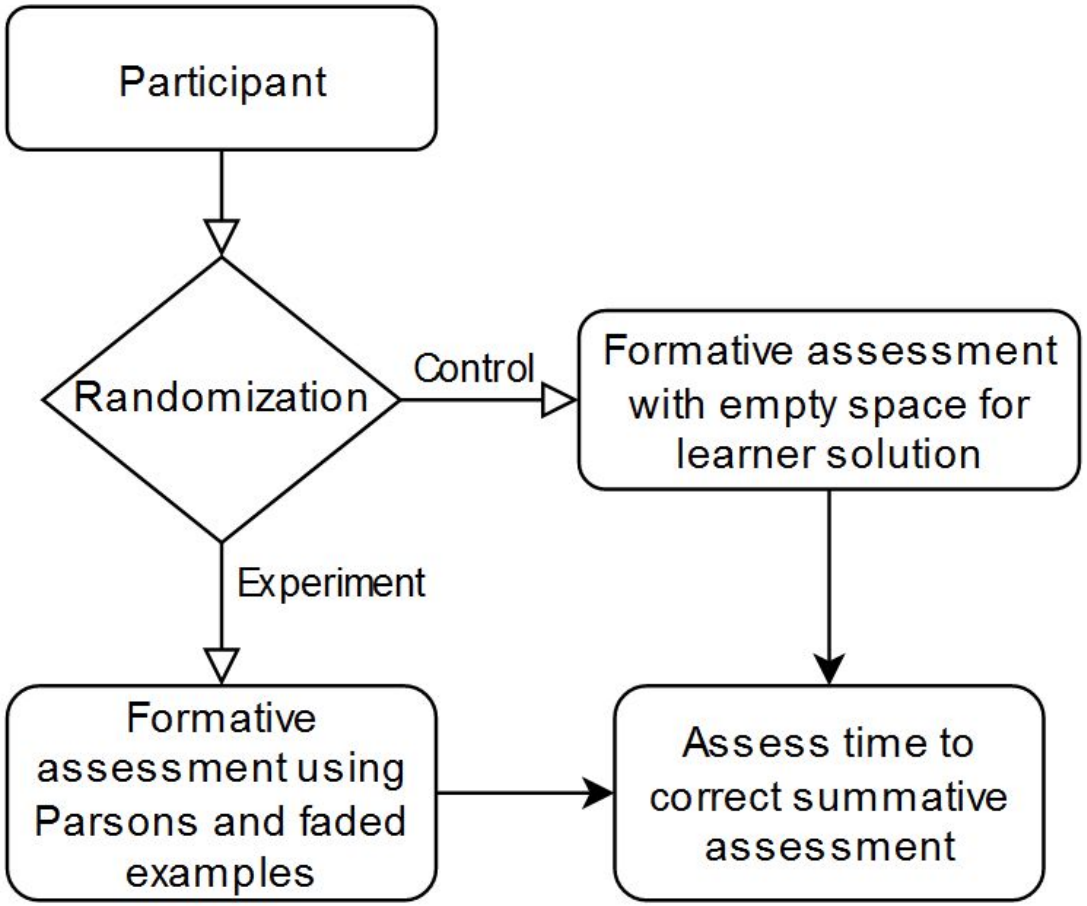
Code | Start Over | Hints | Run Code | Submit Answer

```
1 2
2
3
```

```
[1] 2
```

I expected a call to sqrt() where you wrote 2. Try it again; next time's the charm!

# Aim 3 Study Design




- Need a new IRB
- Can use the students from BEL lab
- Survey bias, but faster response rates

- 2 Groups,
  Continuous (time) + Binary (correct) measurements
  - t-test, ANOVA

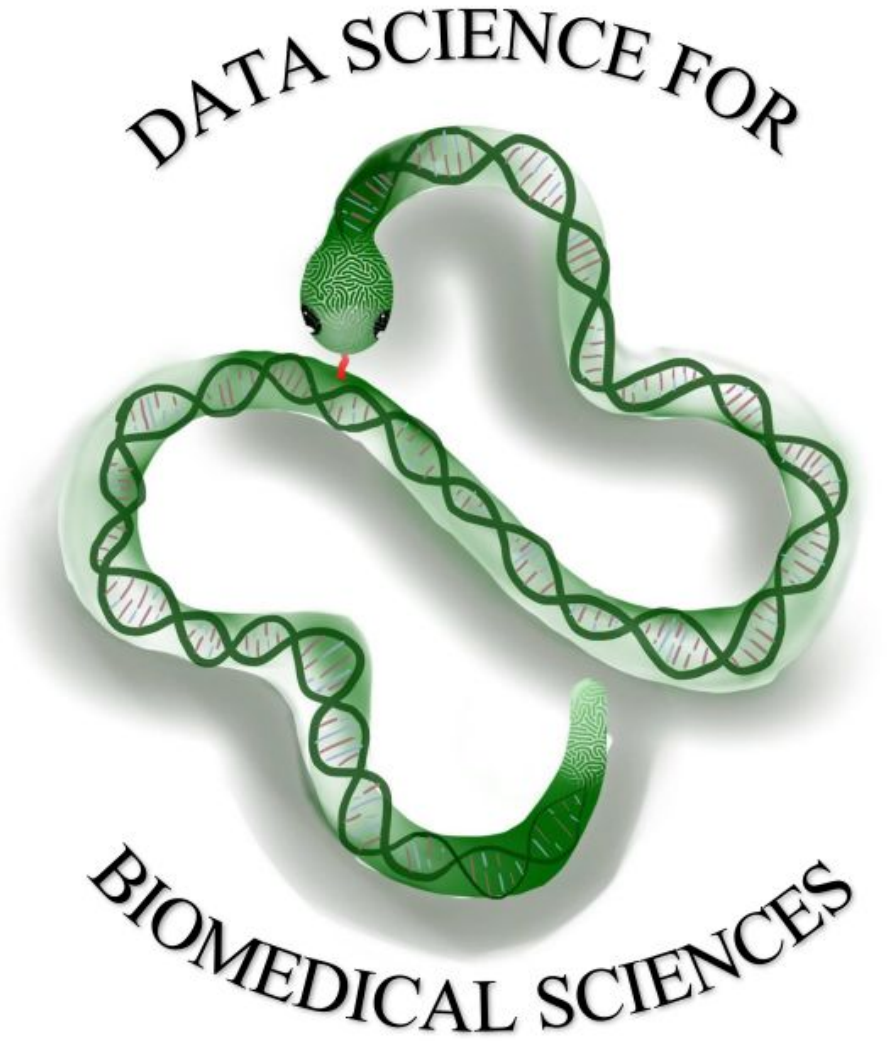
# *In conclusion…*

1. Identified and created learner personas and assessment survey  (Aim 1)
2. Working on building out more content (Aim 2)
   a. OMOP and HL7 FHIR data format examples

Next steps:

1. Build the experimental assessment platform (Aim 3)
2. Paper #1: Creation of personas a validated survey
3. More workshops + expand content

# *Questions?*



# References


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.

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

Anderson, L. W., Bloom, B. S., and others. (2001). *A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives*. Longman,.

Castro-Schilo, L. (2017, April 25). *Principal components or factor analysis?* JMP User Community. https://community.jmp.com/t5/JMP-Blog/Principal-components-or-factor-analysis/ba-p/38347

Chen, D. (2017). *Pandas for Everyone: Python Data Analysis* (1st edition). Addison-Wesley Professional.

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.

Dunn, M. C., and Bourne, P. E. (2017). Building the biomedical data science workforce. *PLoS Biology*, *15*(7), 1–9. http://login.ezproxy.lib.vt.edu/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=a9h&AN=124144283&scope=site

Feldon, D. F., Jeong, S., Peugh, J., Roksa, J., Maahs-Fladung, C., Shenoy, A., and Oliva, M. (2017). Null effects of boot camps and short-format training for PhD students in life sciences. *Proceedings of the National Academy of Sciences of the United States of America*, *114*(37), 9854–9858. https://doi.org/10.1073/pnas.1705783114

Gans, M., Tody, H., and Greg, W. (2020). *JavaScript for Data Science*. https://js4ds.org/

Harrison, M. (2016). *Learning the Pandas Library: Python Tools for Data Munging, Analysis, and Visual (9781533598240): Harrison, Matt, Prentiss, Michael: Books*. https://smile.amazon.com/Learning-Pandas-Library-Munging-Analysis/dp/153359824X/ref=sr_1_1?dchild=1&keywords=Learning+the+Pandas+Library&qid=1613551948&sr=8-1

Jordan, K. (2016). *Data Carpentry Assessment Report: Analysis of Post-Workshop Survey Results*. Zenodo. https://doi.org/10.5281/zenodo.165858

Jordan, K. (2018). *Analysis of The Carpentries Long-Term Impact Survey*. Zenodo. https://doi.org/10.5281/zenodo.1402200

Jordan, K. L., Marwick, B., Duckles, J., Zimmerman, N., and Becker, E. (2017). *Analysis of Software Carpentry's Post-Workshop Surveys*. Zenodo. https://doi.org/10.5281/zenodo.1043533

Jordan, K. L., Marwick, B., Weaver, B., Zimmerman, N., Williams, J., Teal, T., Becker, E., Duckles, J., Duckles, B., and Wickes, E. (2017). *Analysis of the Carpentries' Long-Term Feedback Survey*. Zenodo. https://doi.org/10.5281/zenodo.1039944

Jordan, K. L., and Michonneau, F. (2020). *Analysis of The Carpentries Long-Term Surveys (April 2020)*. Zenodo. https://doi.org/10.5281/zenodo.3728205

Jordan, K., Michonneau, F., and Weaver, B. (2018). *Analysis of Software and Data Carpentry's Pre- and Post-Workshop Surveys*. Zenodo. https://doi.org/10.5281/zenodo.1325464

# References


Koch, C., and Wilson, G. (2016). *Software carpentry: Instructor Training*. https://doi.org/10.5281/zenodo.57571

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

Lee, G., Bacon, S., Bush, I., Fortunato, L., Gavaghan, D., Lestang, T., Morton, C., Robinson, M., Rocca-Serra, P., Sansone, S.-A., and Webb, H. (2021). Barely sufficient practices in scientific computing. *Patterns*, *2*(2). https://doi.org/10.1016/j.patter.2021.100206

McKinney. (2017). *Python for Data Analysis* (2nd ed.). https://www.oreilly.com/library/view/python-for-data/9781449323592/

National Institutes of Health. (2013, June 15). *Big Data to Knowledge*. https://commonfund.nih.gov/bd2k

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

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

UCLA: Statistical Consulting Group. (2021, February 17). *A Practical Introduction to Factor Analysis*. https://stats.idre.ucla.edu/spss/seminars/introduction-to-factor-analysis/

Wickham, H., and Garrett, G. (2017). *R for Data Science*. https://r4ds.had.co.nz/

Wilson, G. (2019). *Teaching tech together: How to make your lessons work and build a teaching community around them*. CRC Press.

Wilson, G., Aruliah, D. A., Brown, C. T., Hong, N. P. C., Davis, M., Guy, R. T., Haddock, S. H., Huff, K. D., Mitchell, I. M., Plumbley, M. D., and others. (2014). Best practices for scientific computing. *PLoS Biology*, *12*(1), e1001745.

Wilson, G., Bryan, J., Cranston, K., Kitzes, J., Nederbragt, L., and Teal, T. K. (2017). Good enough practices in scientific computing. *PLoS Computational Biology*, *13*(6), e1005510.

Word, K. R. (2017, December 11). *When Do Workshops Work? A Response to the "Null Effects" paper from Feldon et al.* Software Carpentry. http://software-carpentry.org//blog/2017/12/response-null-effects.html

# *Response Rates*

Self-assessment (persona): 51 respondents, 45 non-missing

3 Workshops:

| | Registrants | Attendees Day 1 | Attendees Day 2 | Pre workshop | Post workshop | Long-term |
|---|---|---|---|---|---|---|
| 1 | 27 | 20 | 11 | 7 | 7 | N/A |
| 2 | 5 (?) | 5 | N/A | 0 | 0 | N/A |
| 3 | 45 | 19 | 16 | 12 | 7 | N/A |
| Total | 77 | 44 | 27 | 19 | 14 | ~ 0.5 |
| % | | -43% Registrants | -37% Day 1 | -57% Day 1<br>-75% Registrants | -26% Pre<br>-68% Day 1<br>-82% Registrants | -98% Post |

# *Survey Response Rates*

Carpentries

| | Pre workshop | Post workshop | Long-term |
|---|---|---|---|
| dc | 1,259 | 862 | |
| swc | 14,154 | 6,458 | |
| Total | 15,413 | 7320 | 162 |

# *Listservs*

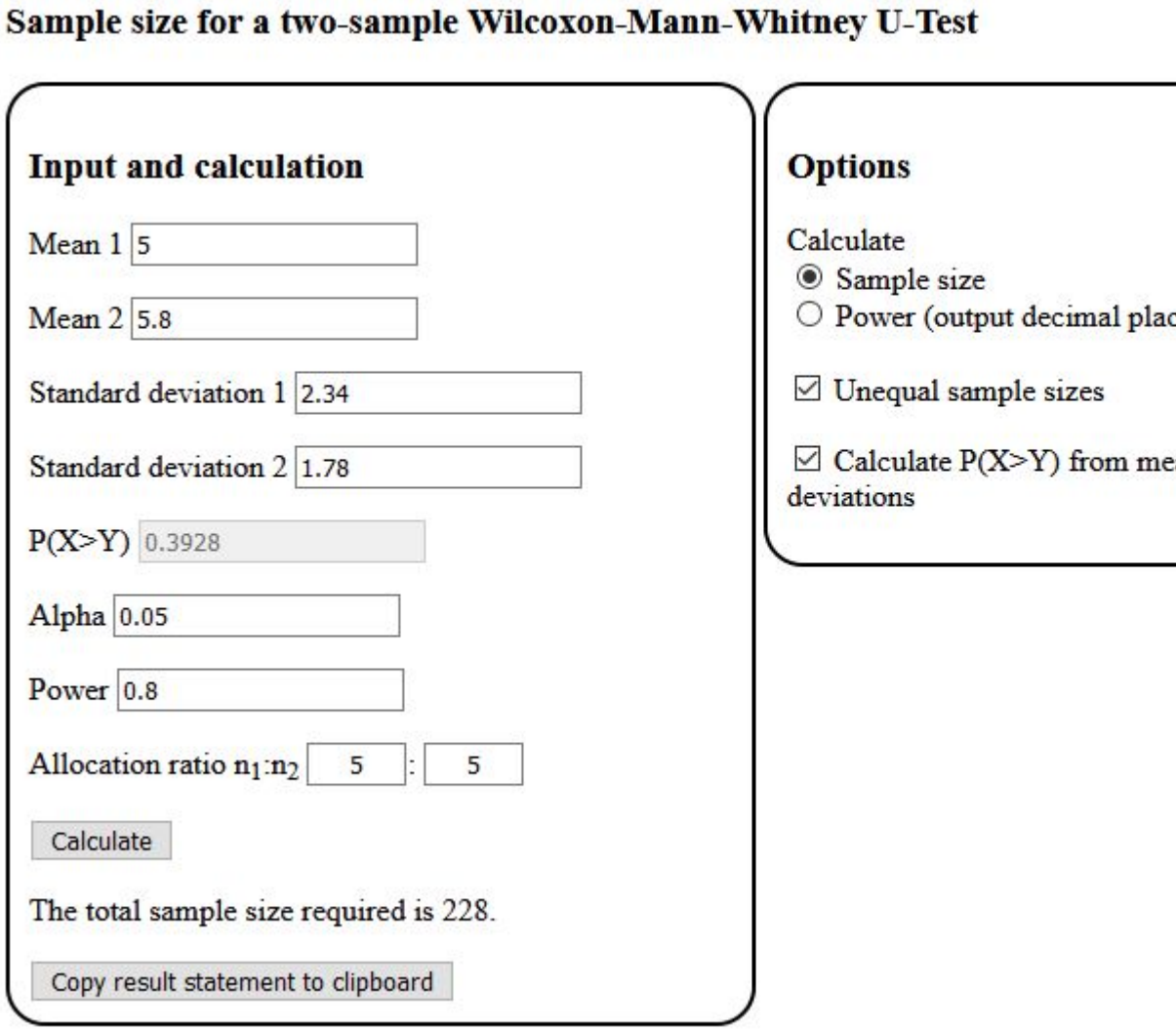
1. iTHRIV (Taryn Luoma, MHA)
2. GBCB
3. IGEP (Dennie Munson)
4. VetMed (Andrea Green)
5. VT Carpentries workshops (Nathaniel Porter, PhD)
6. VT Roanoke Center (David Conners)

- No explicit email sent to the for VCOM listserve
- Sample population does **not** contain 8 iTHRIV Scholars

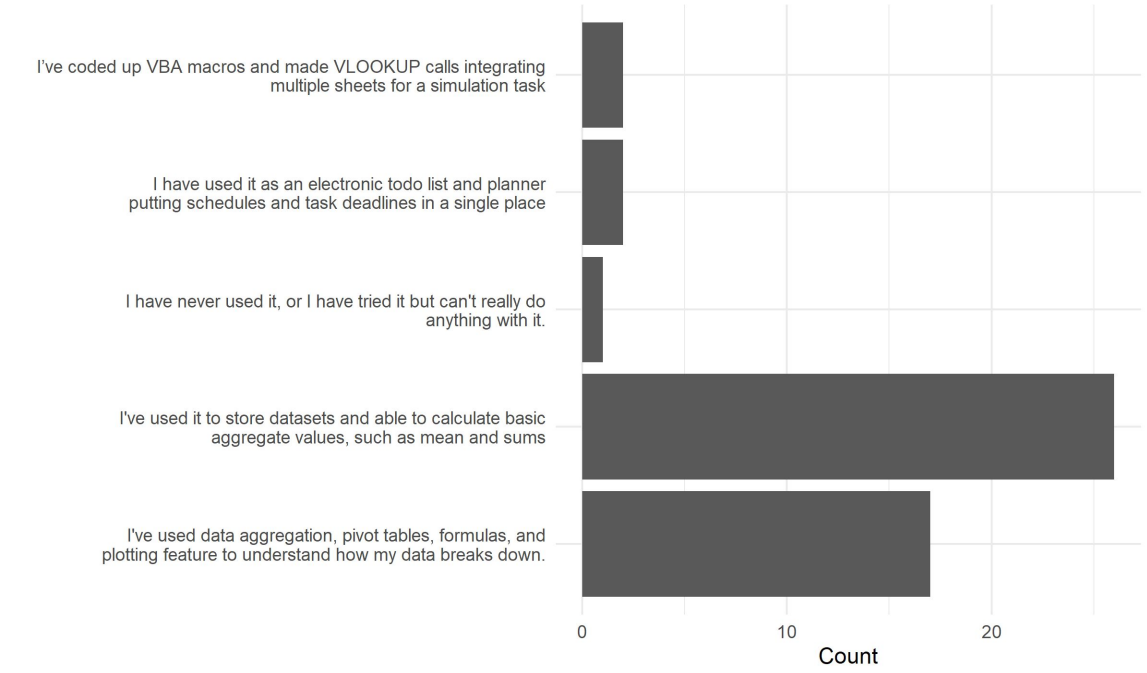
# Power Calculation

https://homepage.univie.ac.at/robin.ristl/samplesize.php?test=wilcoxon

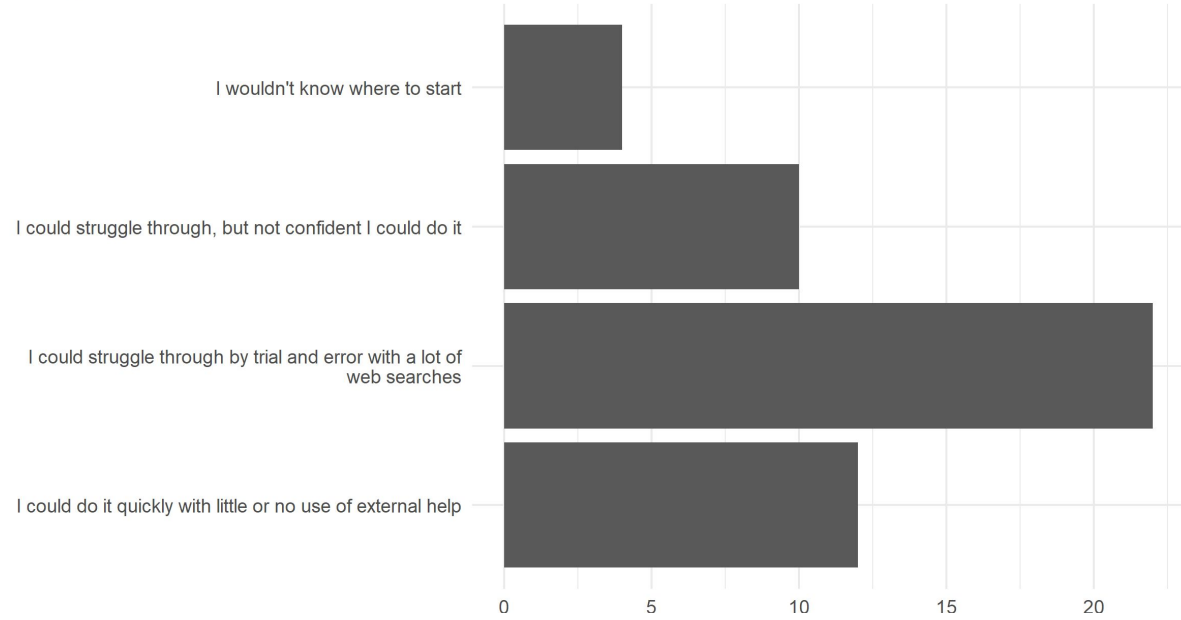
```
> ms
# A tibble: 15 x 3
   variable               mean    sd
   <chr>                 <dbl> <dbl>
 1 Original data           6.6 0.699
 2 Write program           5.6 2.01
 3 Search help             6   0.816
 4 Overcome problem        5.3 1.95
 5 Confident programming   5.3 0.949
 6 R/Python reproduce      6.4 0.843
 7 R/Python efficient      6.7 0.483
 8 Name tidy               5.1 1.45
 9 Transform data          5   1.63
10 Identify spreadsheets   5.6 1.90
11 Assess spreadsheets     5.4 2.01
12 Break down steps        5.7 0.949
13 Construct plot          5.1 1.52
14 Build pipeline          4.8 1.81
15 Calculate analysis      5.2 1.03
> responses[responses$pre_post == 0, "Assess spreadsheets", drop = TRUE] %>% mean()
[1] 5
> responses[responses$pre_post == 1, "Assess spreadsheets", drop = TRUE] %>% mean()
[1] 5.8
> responses[responses$pre_post == 0, "Assess spreadsheets", drop = TRUE] %>% sd()
[1] 2.345208
> responses[responses$pre_post == 1, "Assess spreadsheets", drop = TRUE] %>% sd()
[1] 1.788854
```



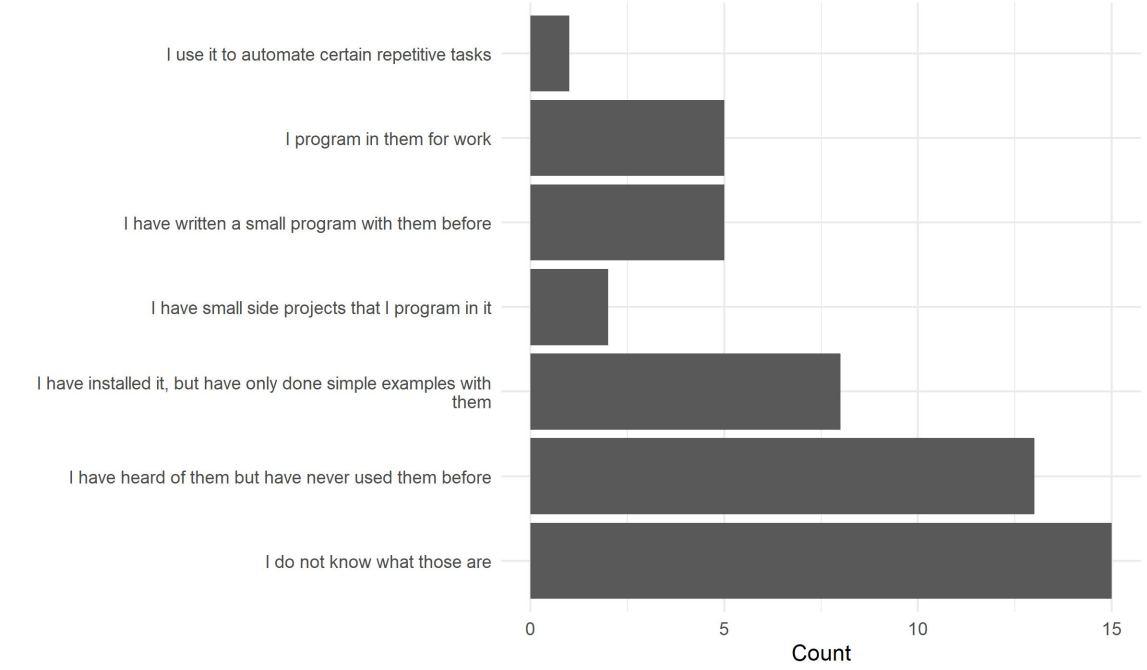
## How familiar are you with Microsoft Excel?



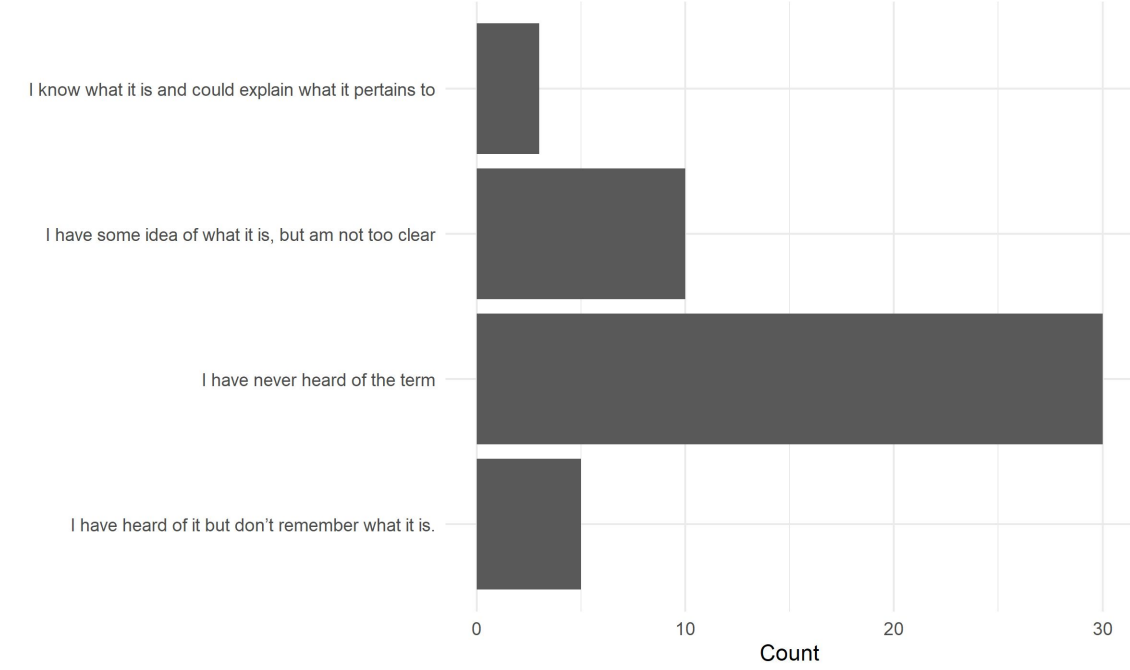

## If you were given a dataset (e.g., Excel file, CSV file) and asked to do some preliminary analysis on it, which of these best describe how easily you can accomplish the task?

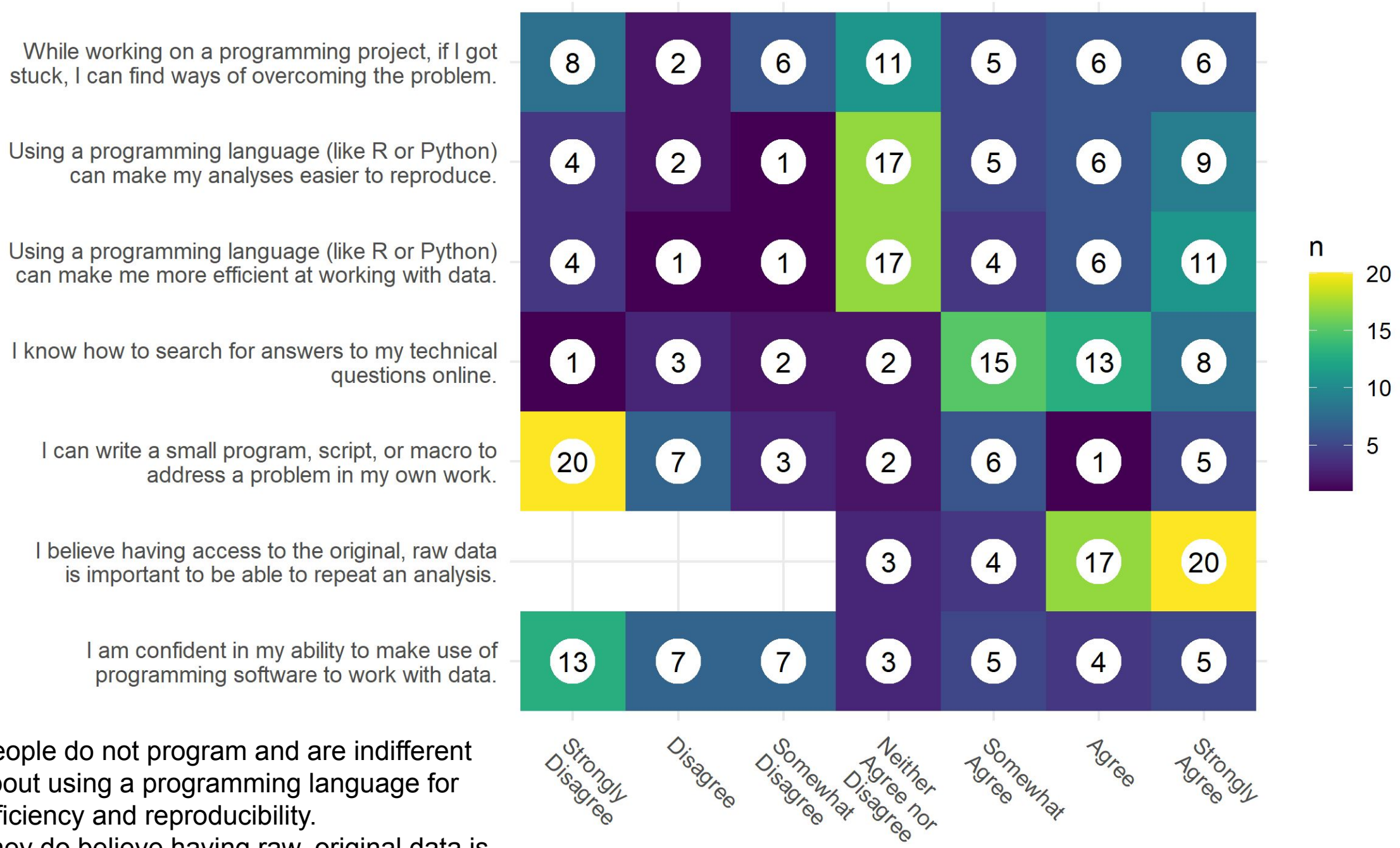



## How familiar are you with interactive programming languages like Python or R?



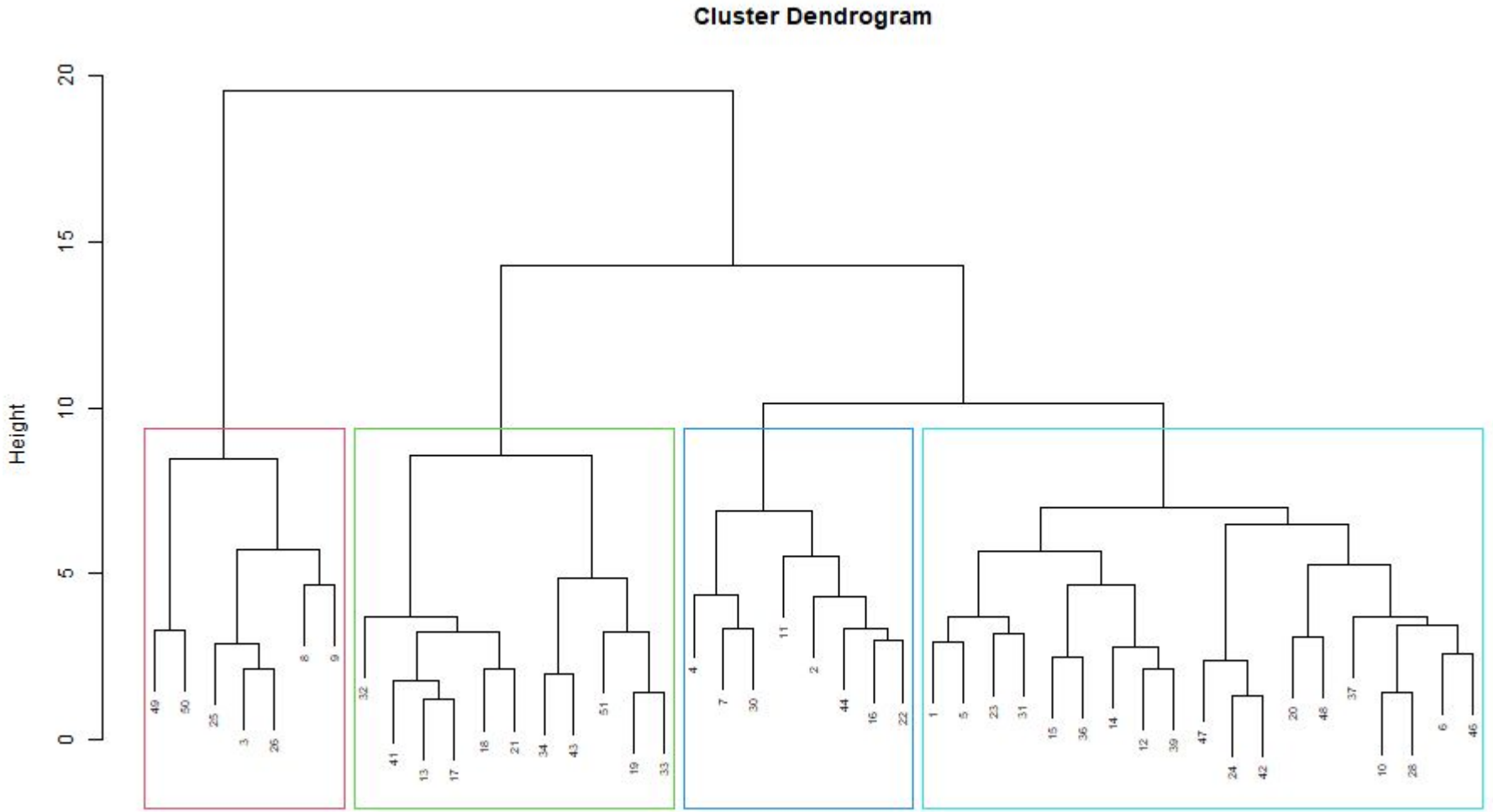

## Are you familiar with the term "tidy data"?



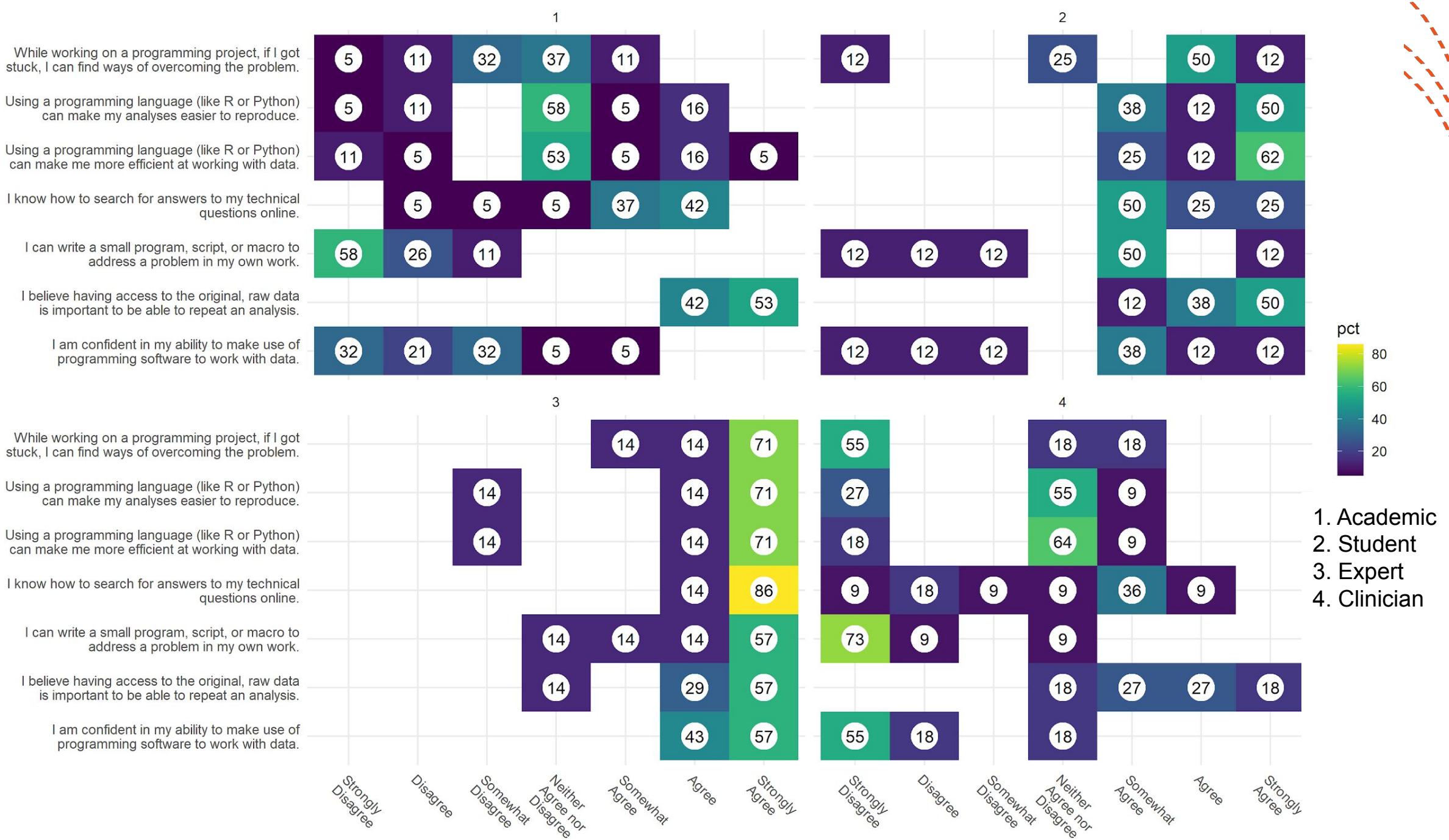



People do not program and are indifferent about using a programming language for efficiency and reproducibility.
They do believe having raw, original data is important for reproducibility.

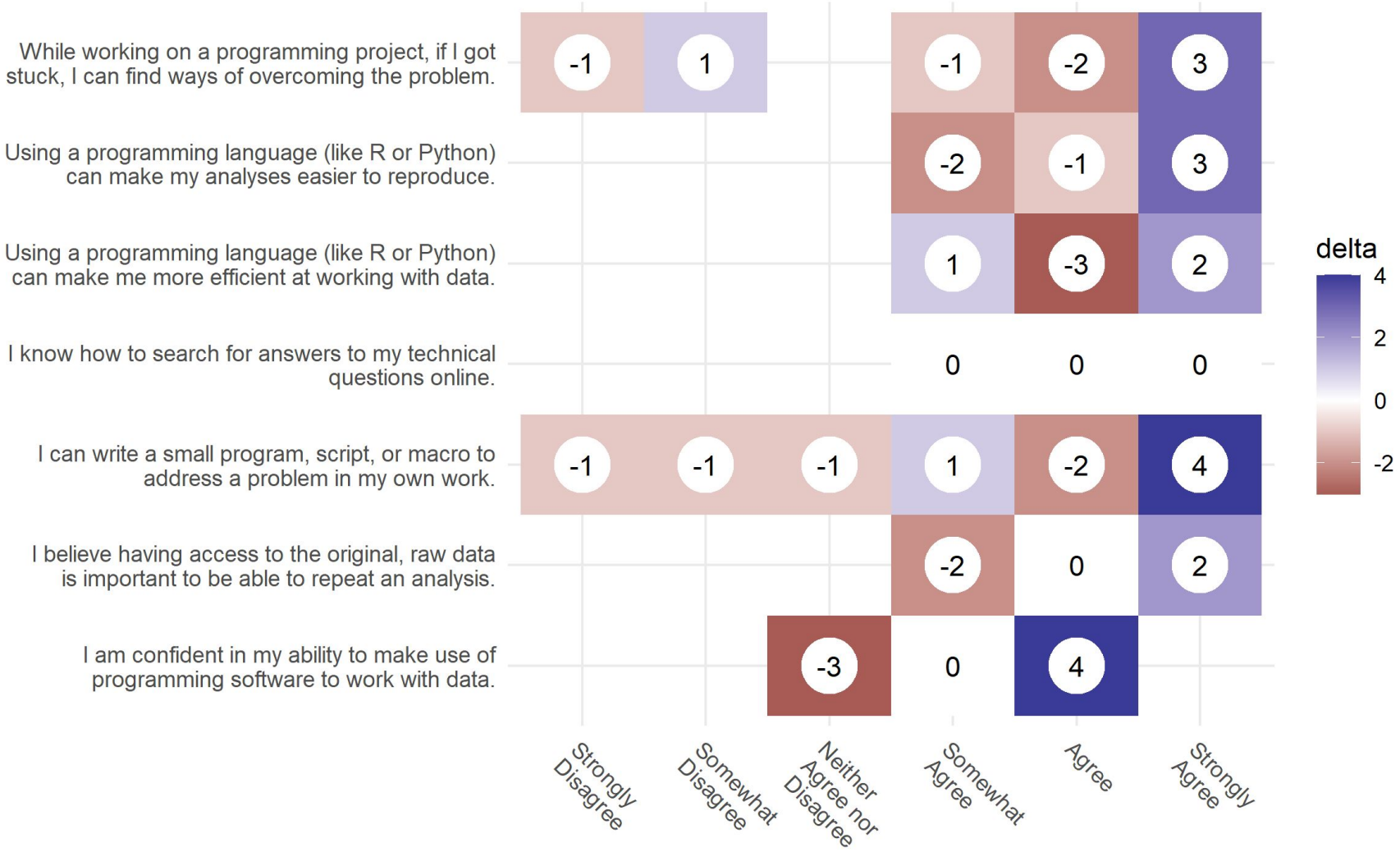
# Clustering Dendrogram




Experts Clinician Academic Student

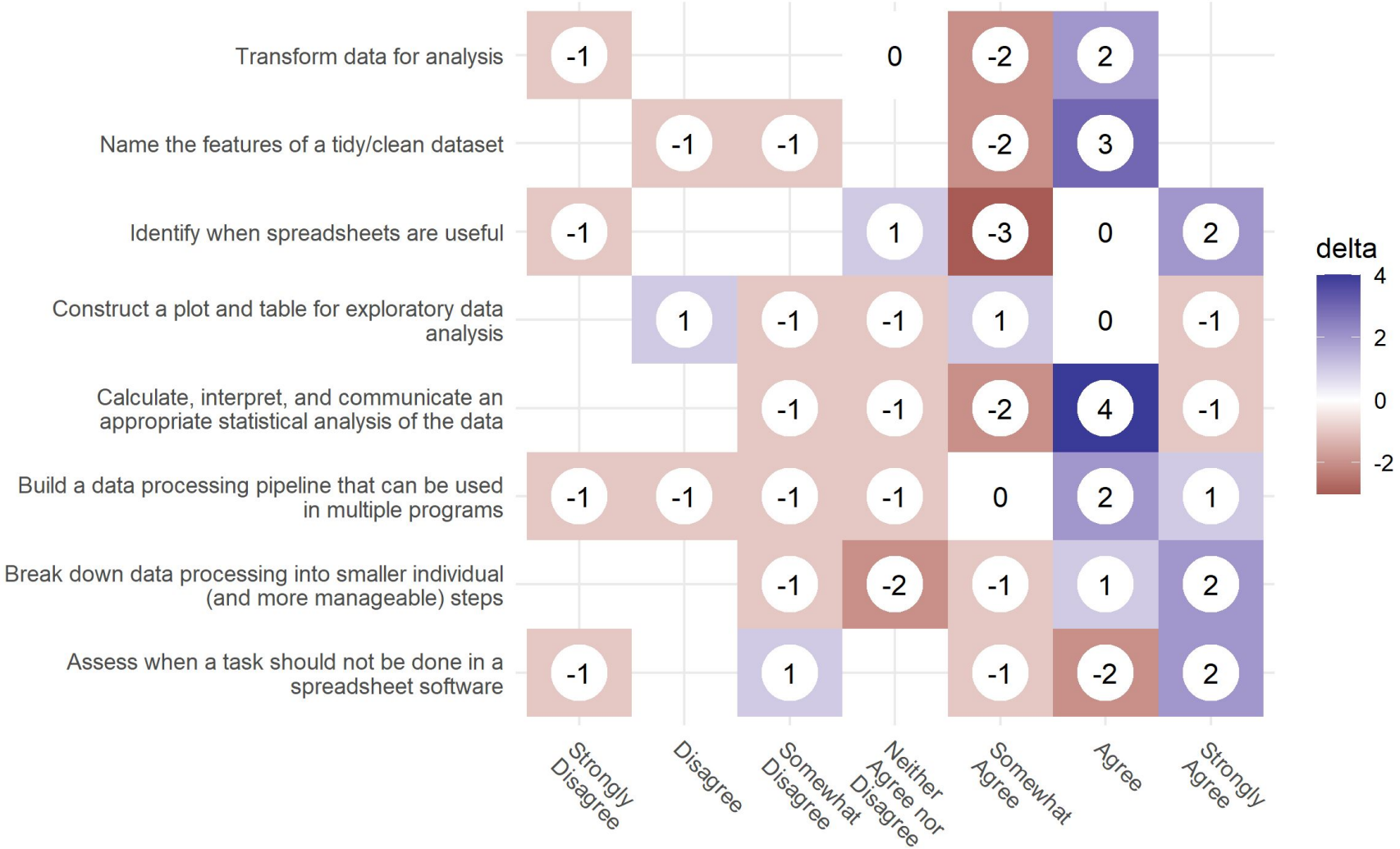

1
2
3
4
While working on a programming project, if I got stuck, I can find ways of overcoming the problem.
Using a programming language (like R or Python) can make my analyses easier to reproduce.
Using a programming language (like R or Python) can make me more efficient at working with data.
I know how to search for answers to my technical questions online.
I can write a small program, script, or macro to address a problem in my own work.
I believe having access to the original, raw data is important to be able to repeat an analysis.
I am confident in my ability to make use of programming software to work with data.
Strongly Disagree
Disagree
Somewhat Disagree
Neither Agree nor Disagree
Somewhat Agree
Agree
Strongly Agree
pct
80
60
40
20
1. Academic
2. Student
3. Expert
4. Clinician

# Changes from the pre-workshop and post-workshop responses



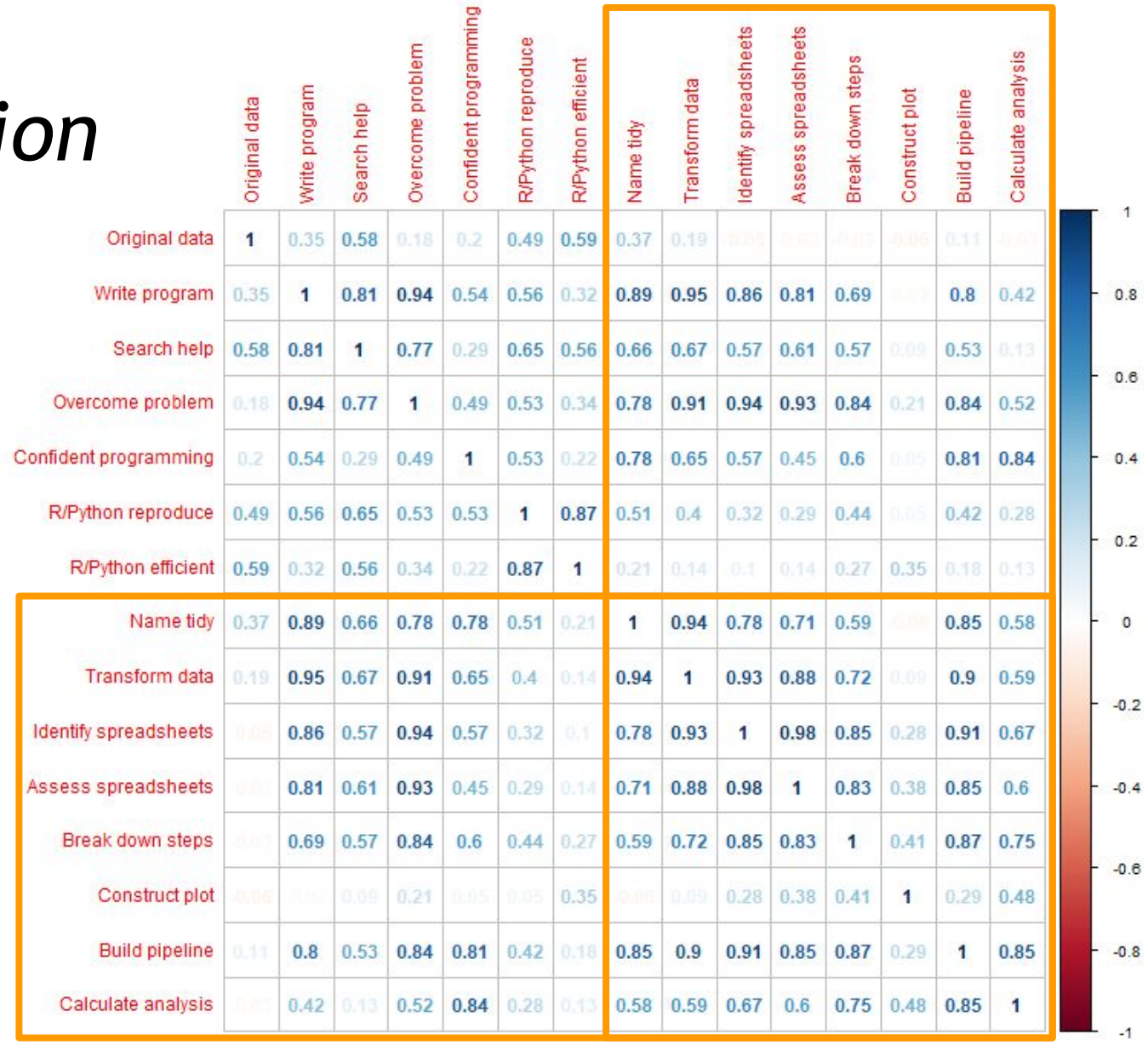

Most participants have more confidence in their programming and data abilities

# *Learning objective pre-post changes*




Most participants have more confidence in meeting the learning objectives

# *Pre+Post Correlation*



| | Original data | Write program | Search help | Overcome problem | Confident programming | R/Python reproduce | R/Python efficient | Name tidy | Transform data | Identify spreadsheets | Assess spreadsheets | Break down steps | Construct plot | Build pipeline | Calculate analysis |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Original data | 1 | 0.35 | 0.58 | 0.18 | 0.2 | 0.49 | 0.59 | 0.37 | 0.19 | 0.05 | 0.01 | 0.03 | -0.09 | 0.11 | -0.01 |
| Write program | 0.35 | 1 | 0.81 | 0.94 | 0.54 | 0.56 | 0.32 | 0.89 | 0.95 | 0.86 | 0.81 | 0.69 | 0.07 | 0.8 | 0.42 |
| Search help | 0.58 | 0.81 | 1 | 0.77 | 0.29 | 0.65 | 0.56 | 0.66 | 0.67 | 0.57 | 0.61 | 0.57 | 0.09 | 0.53 | 0.13 |
| Overcome problem | 0.18 | 0.94 | 0.77 | 1 | 0.49 | 0.53 | 0.34 | 0.78 | 0.91 | 0.94 | 0.93 | 0.84 | 0.21 | 0.84 | 0.52 |
| Confident programming | 0.2 | 0.54 | 0.29 | 0.49 | 1 | 0.53 | 0.22 | 0.78 | 0.65 | 0.57 | 0.45 | 0.6 | 0.05 | 0.81 | 0.84 |
| R/Python reproduce | 0.49 | 0.56 | 0.65 | 0.53 | 0.53 | 1 | 0.87 | 0.51 | 0.4 | 0.32 | 0.29 | 0.44 | 0.05 | 0.42 | 0.28 |
| R/Python efficient | 0.59 | 0.32 | 0.56 | 0.34 | 0.22 | 0.87 | 1 | 0.21 | 0.14 | 0.1 | 0.14 | 0.27 | 0.35 | 0.18 | 0.13 |
| Name tidy | 0.37 | 0.89 | 0.66 | 0.78 | 0.78 | 0.51 | 0.21 | 1 | 0.94 | 0.78 | 0.71 | 0.59 | 0.08 | 0.85 | 0.58 |
| Transform data | 0.19 | 0.95 | 0.67 | 0.91 | 0.65 | 0.4 | 0.14 | 0.94 | 1 | 0.93 | 0.88 | 0.72 | 0.09 | 0.9 | 0.59 |
| Identify spreadsheets | 0.05 | 0.86 | 0.57 | 0.94 | 0.57 | 0.32 | 0.1 | 0.78 | 0.93 | 1 | 0.98 | 0.85 | 0.28 | 0.91 | 0.67 |
| Assess spreadsheets | 0.01 | 0.81 | 0.61 | 0.93 | 0.45 | 0.29 | 0.14 | 0.71 | 0.88 | 0.98 | 1 | 0.83 | 0.38 | 0.85 | 0.6 |
| Break down steps | 0.03 | 0.69 | 0.57 | 0.84 | 0.6 | 0.44 | 0.27 | 0.59 | 0.72 | 0.85 | 0.83 | 1 | 0.41 | 0.87 | 0.75 |
| Construct plot | -0.09 | 0.07 | 0.09 | 0.21 | 0.05 | 0.05 | 0.35 | 0.08 | 0.09 | 0.28 | 0.38 | 0.41 | 1 | 0.29 | 0.48 |
| Build pipeline | 0.11 | 0.8 | 0.53 | 0.84 | 0.81 | 0.42 | 0.18 | 0.85 | 0.9 | 0.91 | 0.85 | 0.87 | 0.29 | 1 | 0.85 |
| Calculate analysis | -0.01 | 0.42 | 0.13 | 0.52 | 0.84 | 0.28 | 0.13 | 0.58 | 0.59 | 0.67 | 0.6 | 0.75 | 0.48 | 0.85 | 1 |

# *Identifying the Core Concepts: Scientific Computing*

## Best Practices

1. Write programs for people, not computers
2. Let the computer to the work
3. Make incremental changes
4. Don't repeat yourself (or others)
5. Plan for mistakes
6. Optimize software only after it works correctly
7. Document design and and purpose, not mechanics
8. Collaborate

## Good enough

1. **Data management**
2. Software
3. Collaboration
4. **Project organization**
5. Keeping track of changes
6. Manuscripts

## Barely sufficient

1. Availability of software for others to use
2. Document setup, use, and expectations of code
3. Version control
4. Test, mostly at the unit level
5. Support or maintenance of the software clarified

(Lee et al., 2021; Wilson et al., 2014, 2017)

# *Workshop Efficacy*

They don't work…

- Spacing instruction over time is better for learning

They do work…

- long-format courses are often **impractical**, but should be used over short-course if practical
- Limited instructor **capacity**
- Learner **time commitment**
- Bring people **together**
- Easily **adaptable** to needs

How to make workshops work:

- Streamline content
- Teach strategically
- Meet learners where they are
- Normalize error and demonstrate recovery
- Explicitly address motivation and self-efficacy
- Build community

(Feldon et al., 2017; Word, 2017)

# *Existing Data Science Book Topics*

**R for Data Science**

1. Explore Introduction
2. Data visualisation
3. Workflow: basics
4. Data transformation
5. Workflow: scripts
6. Exploratory Data Analysis
7. Workflow: projects
8. Wrangle Introduction
9. Tibbles
10. Data import
11. **Tidy data**

**Data Science for JavaScript**

1. Introduction
2. Basic Features
3. Callbacks
4. Objects and Classes
5. HTML and CSS
6. Manipulating Pages
7. Dynamic Pages
8. Visualizing Data
9. Promises
10. Interactive Sites
11. **Managing Data**

**Python for Data Analysis**

1. Preliminaries
2. Introductory Examples
3. IPython: An Interactive Computing and Development Environment
4. NumPy Basics: Arrays and Vectorized Computation
5. Getting Started with pandas
6. Data Loading, Storage, and File Formats
7. **Data Wrangling: Clean, Transform, Merge, Reshape**

**Pandas for Everyone**

1. Pandas DataFrame Basics
2. Pandas Data Structures
3. Introduction to Plotting
4. Data Assembly
5. Missing Data
6. **Tidy Data**

**Learning the Pandas Library**

1. Introduction
2. Installation
3. Data Structures
4. Series
5. Series CRUD
6. Series Indexing
7. Series Methods
8. Series Plotting
9. Another Series Example
10. DataFrames
11. Data Frame Example
12. Data Frame Methods
13. Data Frame Statistics
14. **Grouping, Pivoting, and Reshaping**
15. Dealing With Missing Data
16. Joining Data Frames
17. Avalanche Analysis and Plotting
18. Summary

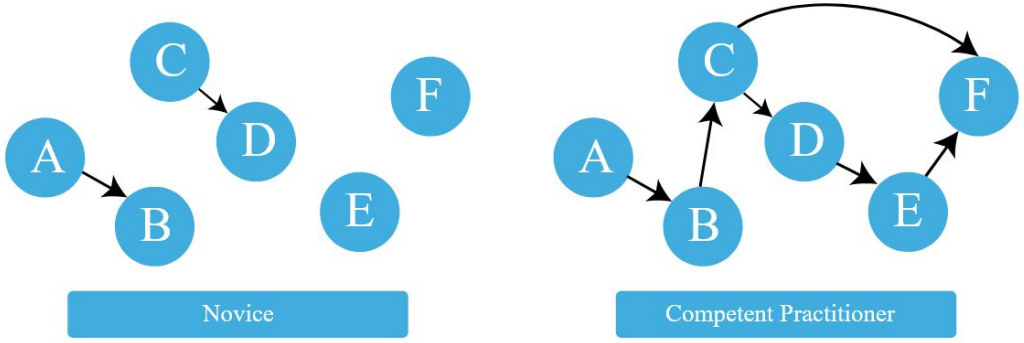
**ds4biomed**

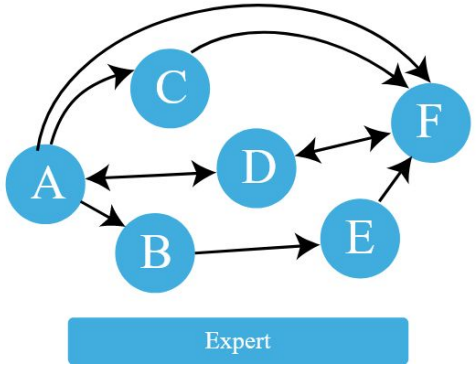
1. Introduction
2. Spreadsheets
3. R + RStudio
4. Load Data
5. Descriptive Calculations
6. **Clean Data (Tidy)**

(Chen, 2017; Gans et al., 2020; Harrison, 2016; McKinney, 2017; Wickham and Garrett, 2017)

# *Managing Prior Knowledge*

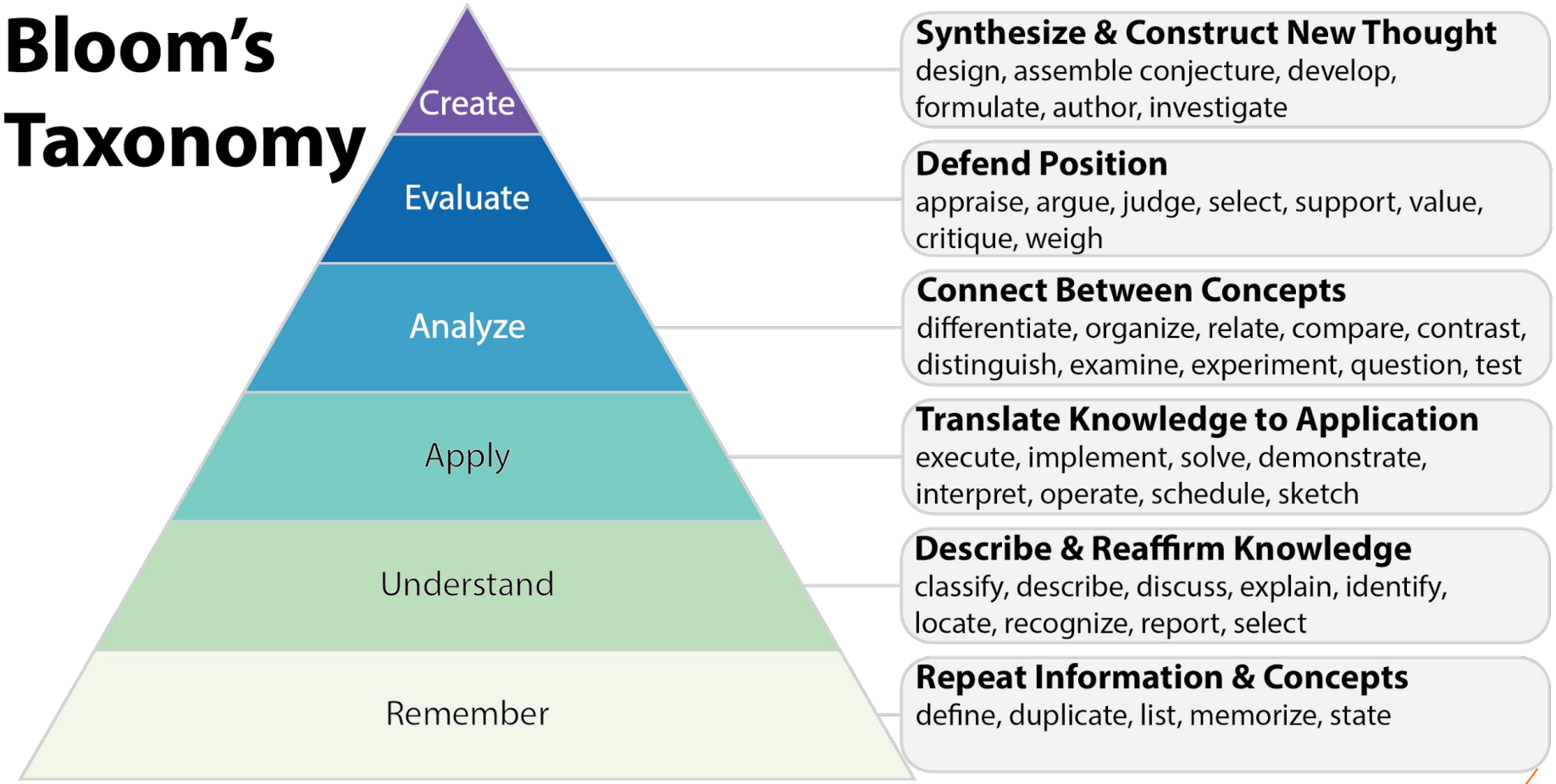
- Learner's prior knowledge can help or hinder learning
- Concept maps: graphic of a mental model




(Ambrose et al., 2010; Koch and Wilson, 2016)

# *Bloom's Taxonomy*




(Anderson et al., 2001)

# Dreyfus Model of Skill Acqusition

| Skill Level / Mental Function | NOVICE | COMPETENT | PROFICIENT | EXPERT | MASTER |
|---|---|---|---|---|---|
| Recollection | Non-situational | Situational | Situational | Situational | Situational |
| Recognition | Decomposed | Decomposed | Holistic | Holistic | Holistic |
| Decision | Analytical | Analytical | Analytical | Intuitive | Intuitive |
| Awareness | Monitoring | Monitoring | Monitoring | Monitoring | Absorbed |

(Dreyfus and Dreyfus, 1980)

# *Computer Science Assessment Studies*

- Novices: never a blank page
  - Existing code to modify gives structure + more realistic
- Canterbury QuestionBank

| | |
|---|---|
| Question | Suppose you try to perform a binary search on a 5-element array sorted in the reverse order of what the binary search algorithm expects. How many of the items in this array will be found if they are searched for? |
| A | 5 |
| B | 0 |
| *C* | 1 |
| D | 2 |
| E | 3 |
| Explanation | Only the middle element will be found. The remaining elements will not be contained in the subranges that we narrow our search to. |

# *Auto-graders*

- Many technical hurdles
- Also need to distinguish learner satisfaction with learner outcomes
- Might help identify bugs, but may discourage critical thinking



(Wilson, 2019)

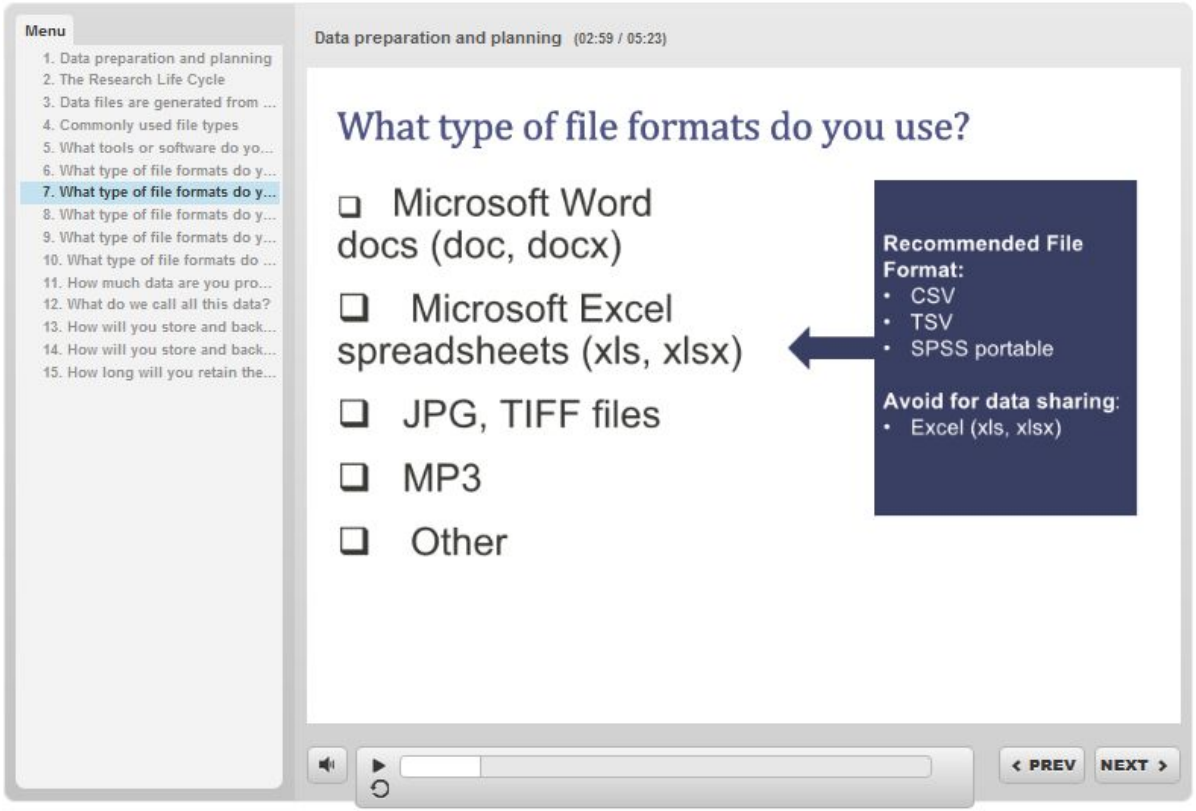
# *BD2K Open Educational Resources for Biomedical Big Data (R25)*

Oregon Health & Science University (OHSU) Educational Materials

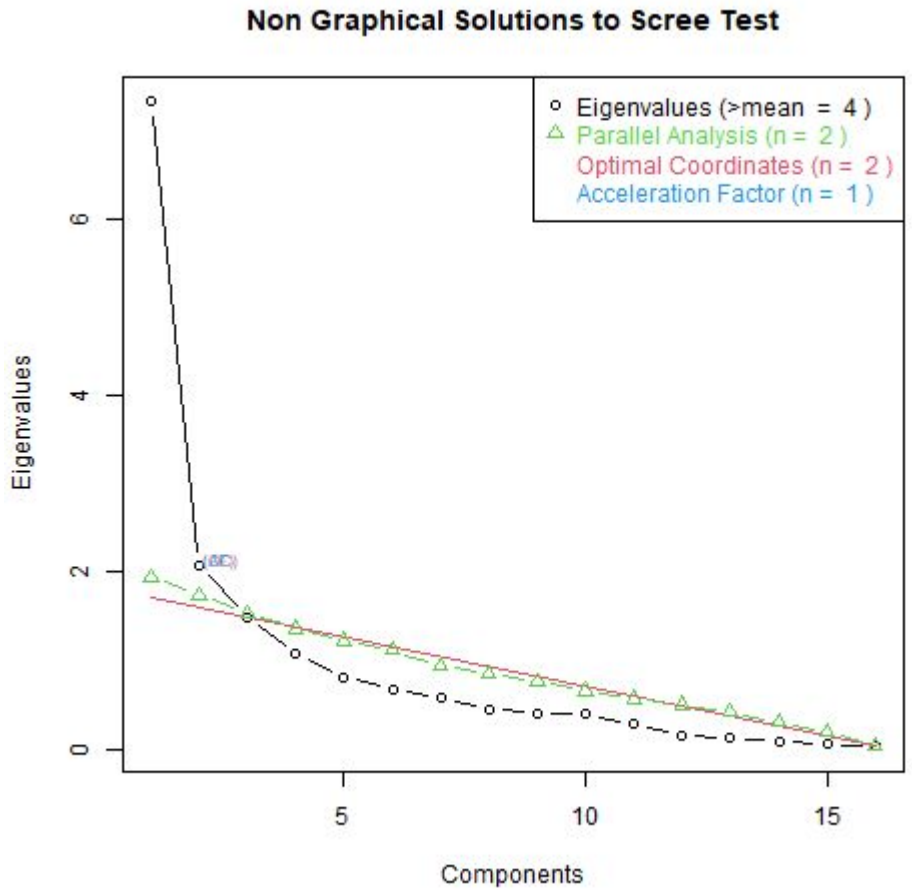
OHSU Big Data to Knowledge (BD2K)

Open Educational Resources (OERs) Project

https://dmice.ohsu.edu/bd2k/topics.html



# EFA: 2 Optimal Loadings




```
## Loadings:
##       Factor1 Factor2
## Q3.1   0.88
## Q3.3   0.97
## Q3.4   1.04
## Q3.5   0.96
## Q4.3   0.57
## Q5.2   0.68
## Q3.6           0.65
## Q6.1           0.95
## Q6.2           1.05
## Q6.3           0.93
## Q3.7
## Q4.1           0.43
## Q4.2   0.32    0.42
## Q4.4
## Q5.1          -0.37
## Q6.4
```

```
## Loadings:
##       Factor1 Factor2 Factor3
## Q3.1   0.82
## Q3.3   0.92
## Q3.4   0.90
## Q3.5   0.89
## Q4.3   0.51
## Q5.2   0.67             0.32
## Q6.1           0.83     0.32
## Q6.2           0.90     0.32
## Q6.3           0.84
## Q4.1                    0.51
## Q4.4                    0.78
## Q6.4   0.33             0.62
## Q3.6           0.46     0.40
## Q3.7
## Q4.2   0.42    0.39     0.41
## Q5.1                   -0.43
```

# EFA Results

3 factors with promax rotation.

Each factor coincided with a survey question blocks

- Programming
  - Q3.3: How familiar are you with interactive programming languages like Python or R?
- Statistics
  - Q6.2: If you were given a dateset containing an individual's smoking status (binary variable) and whether or not they have hypertension (binary variable), would you know how to conduct a statistical analysis to see if smoking has an increased relative risk or odds of hypertension? Any type of model will suffice.
- Data
  - Q4.4: Do you know what "long" and "wide" data are?

# EFA: Selected questions

## 6.2: Logistic regression

1. I wouldn't know where to start
2. I could struggle through, but not confident I could do it
3. I could struggle through by trial and error with a lot of web searches
4. I could do it quickly with little or no use of external help

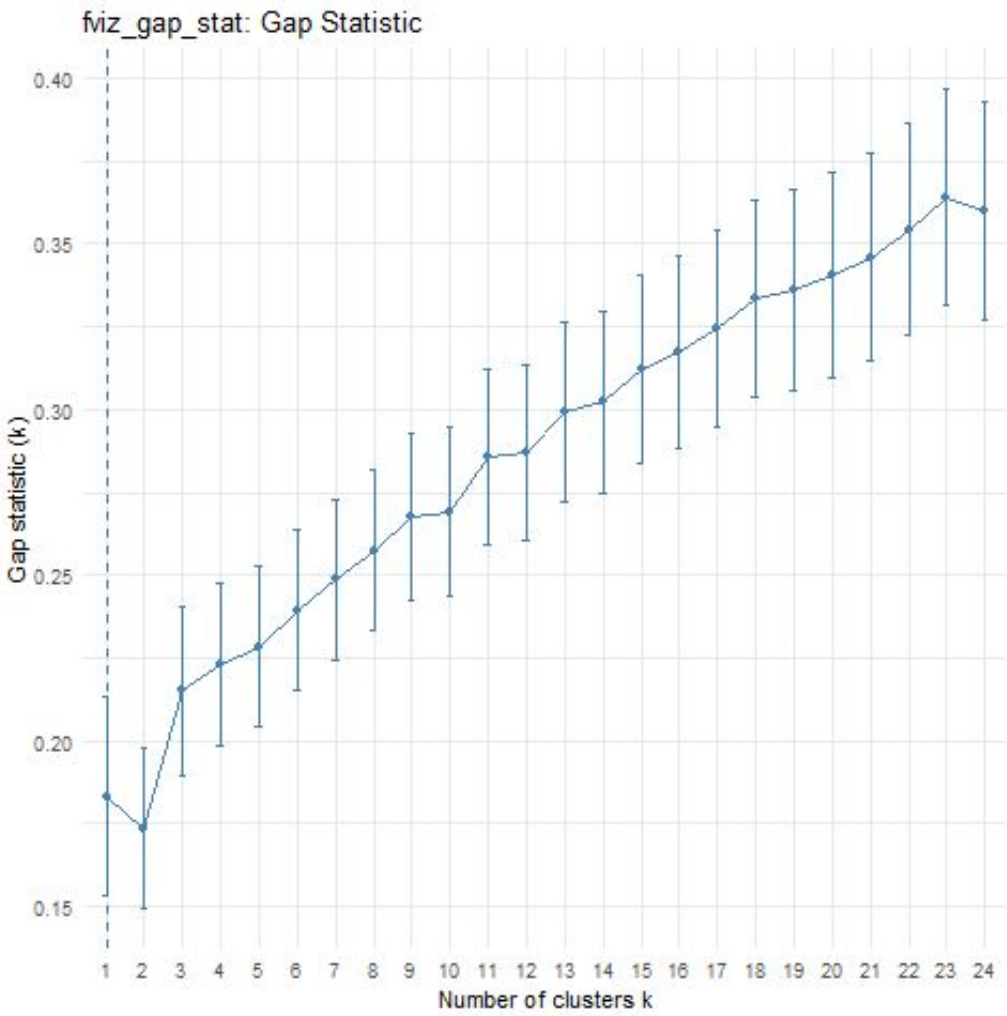
## 3.3: Python/R

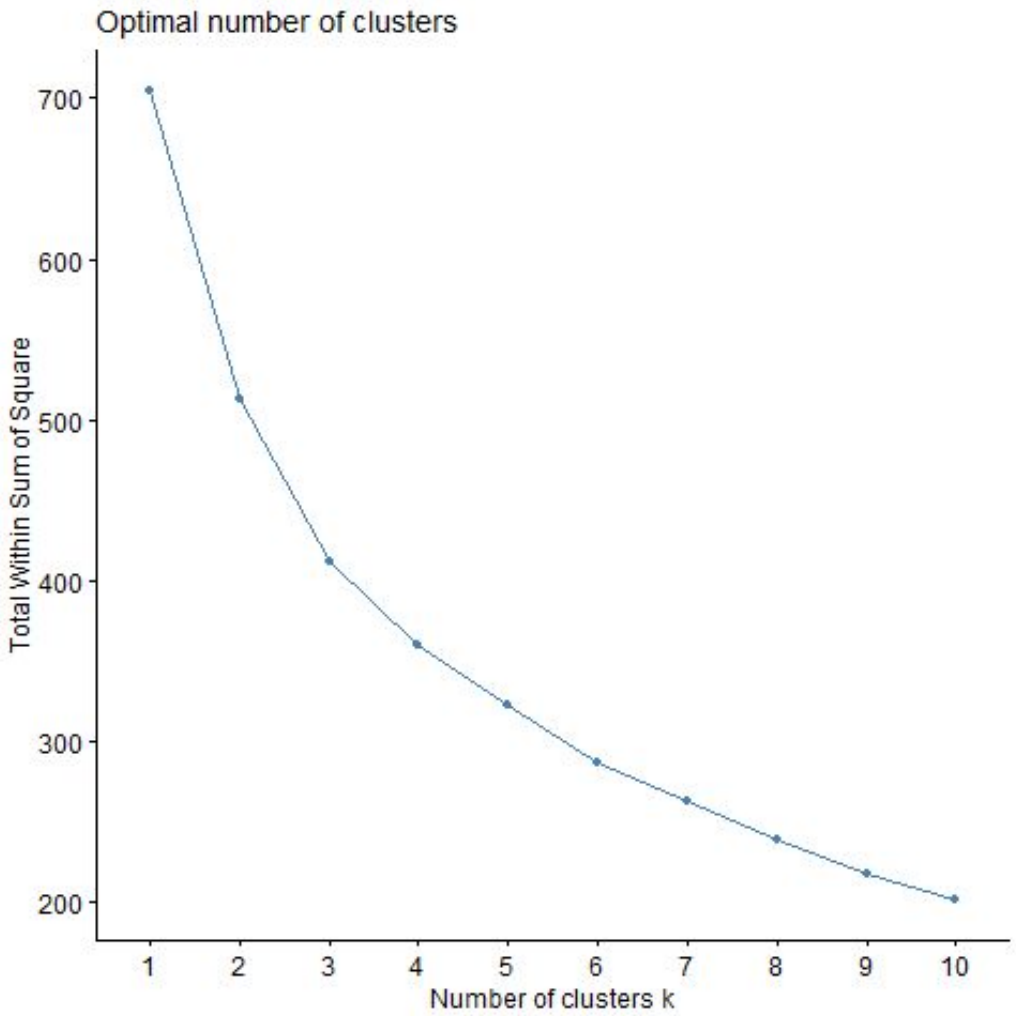
1. I do not know what those are
2. I have heard of them but have never used them before
3. I have installed it, but have only done simple examples with them
4. I have written a small program with them before
5. I use it to automate certain repetitive tasks
6. I have small side projects that I program in it
7. I program in them for work

## 4.4: Long and wide

1. I have never heard of the term
2. I have heard of it but don't remember what it is.
3. I have some idea of what it is, but am not too clear
4. I know what it is and could explain what it pertains to

# *Clustering: 2 Optimal Groups*



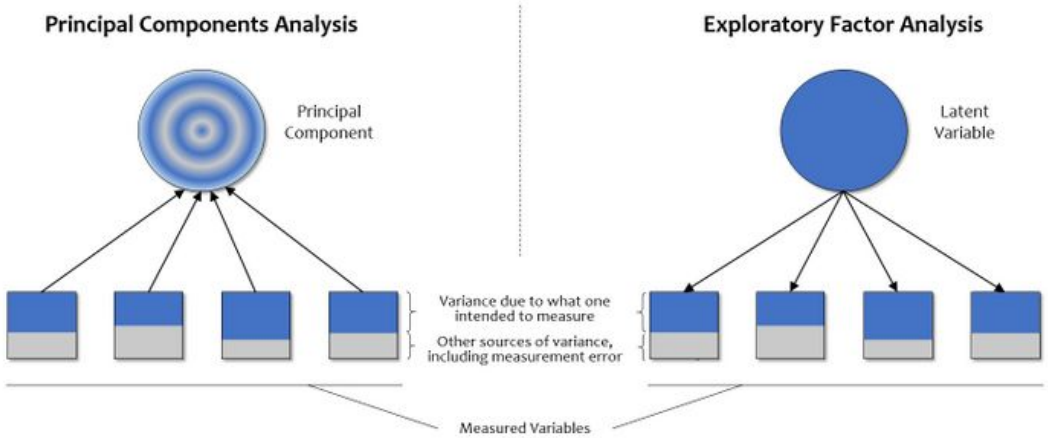
# PCA vs EFA

## PCA

- dimension reduction
- eigenvalue decomposition on full correlation matrix
- components are a mix of what the variables intend to measure

## EFA

- latent variable identification
- eigenvalue decomposition on reduced correlation matrix (squared multiple correlations)
- SMC estimates variance that the underlying factors explain (communality)




(Castro-Schilo, 2017)

# *EFA vs CFA*

**EFA**

- Explore underlying structure
- Our survey has not been explored before in this manner

**CFA**

- Verify factor structure

(UCLA: Statistical Consulting Group, 2021)