Education and Pedagogy of Domain-Specific Learning Materials Using Learning Personas

JSM 2021: Classroom Teaching and Pedagogy Session

Daniel Chen, MPH

Anne Brown, PhD

2021-08-11

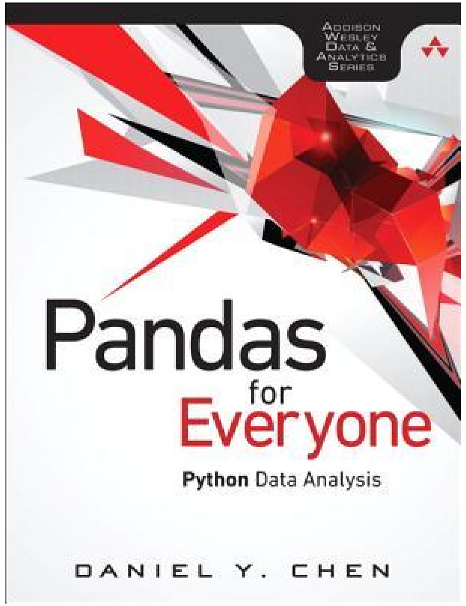
Committee Chair: Anne Brown, PhD

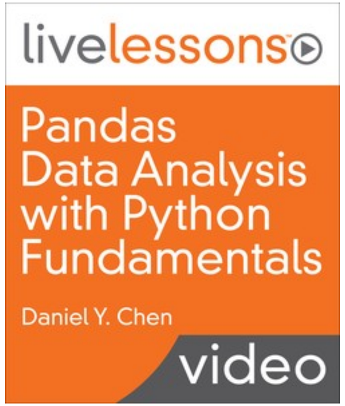
- Assistant Professor
- Biochemistry and Data Science
- Molecular Modeling & Drug Design
- Applied Data Science & Education
- Bevan Brown Lab + DataBridge

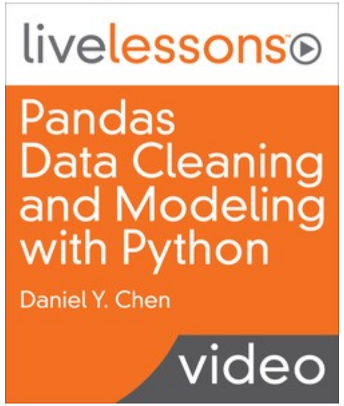
Hello!

- 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
- Workshop Instructor
- R + Python!

Educational Materials







DATA SCIENCE FOR BIOMEDICAL SCIENCES

ds4biomed.tech

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
- **Domain experts** can help learners improve **data literacy**

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?

- **Democratization** of data science education enables **more domain specific learning materials**

- You learn better when things are more relevant

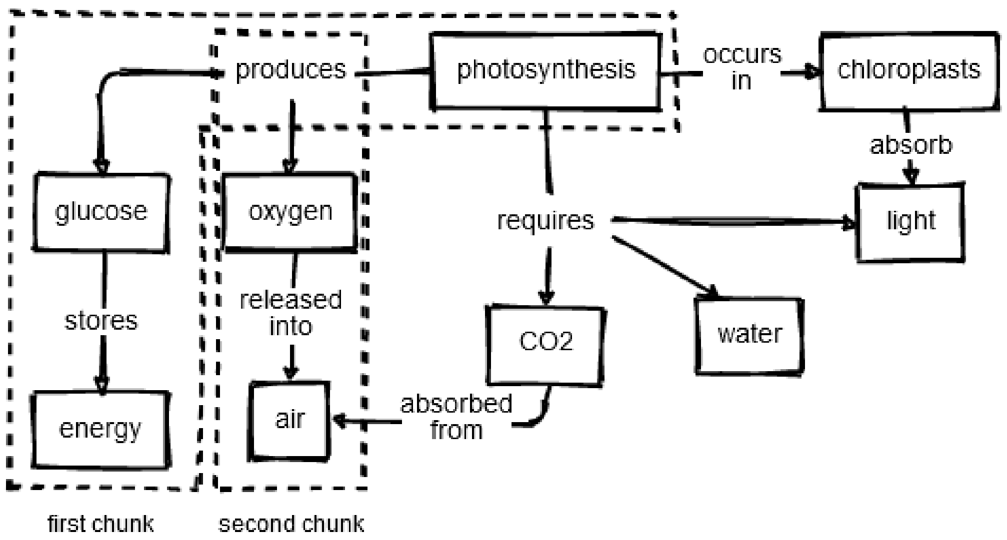
- Internal factors for motivation
- Create feedback loops for learning
- Self-directed learners

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

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

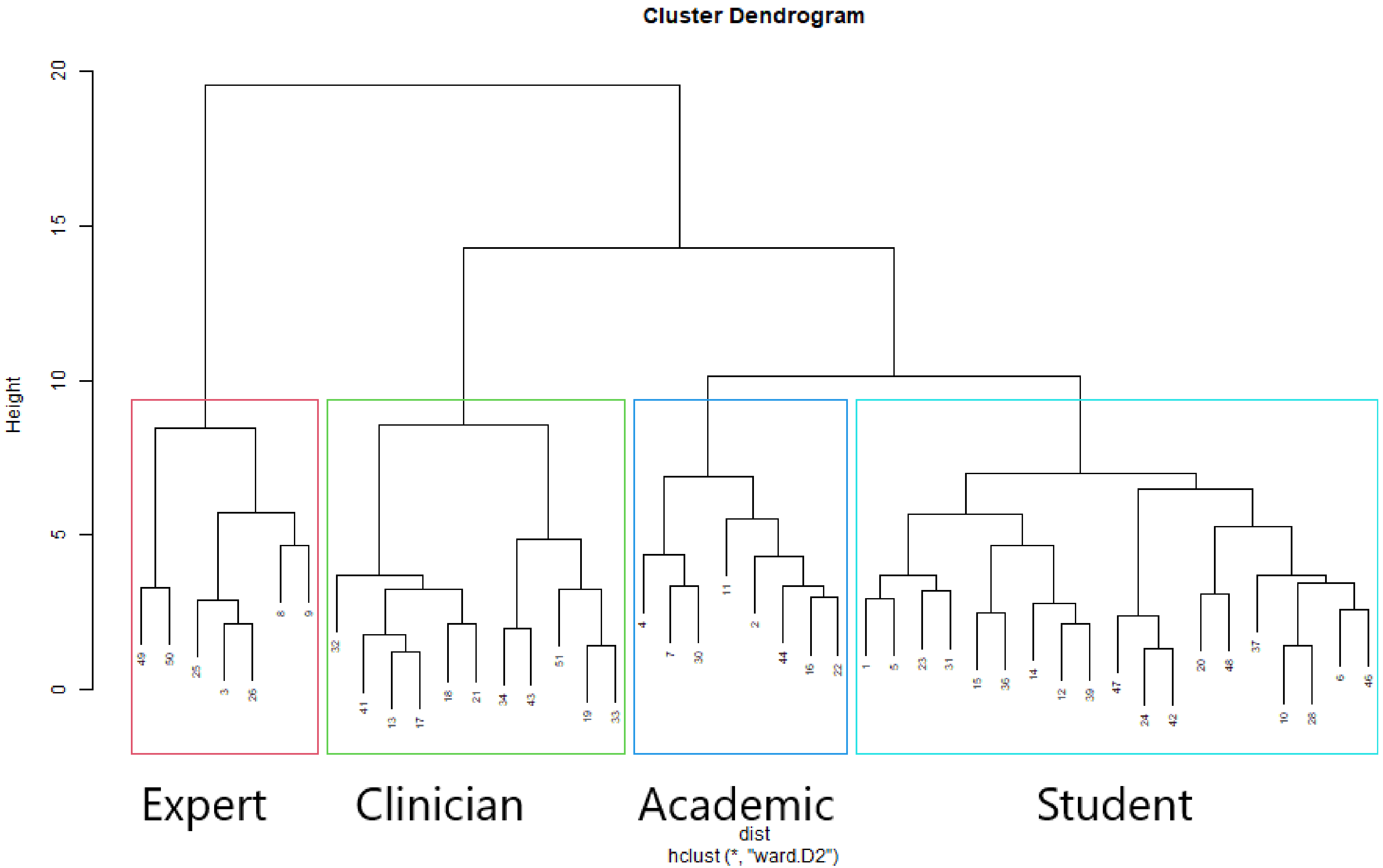
- Dreyfus, S. E., and Dreyfus, H. L. (1980). A five-stage model of the mental activities involved in directed skill acquisition. California Univ Berkeley Operations Research Center.
- Koch, C., and Wilson, G. (2016). Software carpentry: Instructor Training. https://doi.org/10.5281/zenodo.57571
- Wilson, 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

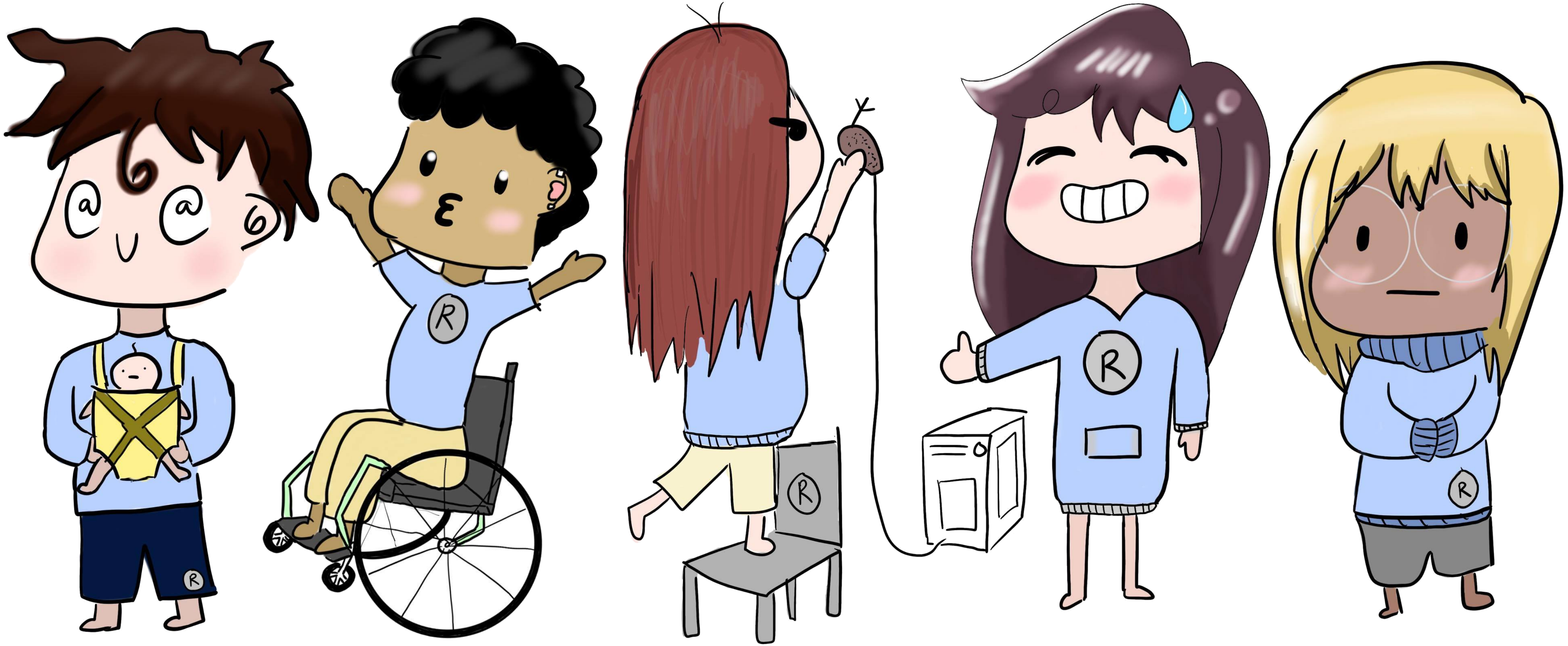
1. Demographics (6)
2. Programs Used in the Past (1)
3. **Programming Experience** (6)
4. **Data Cleaning and Processing Experience** (4)
5. **Project and Data Management** (2)
6. **Statistics** (4)
7. Workshop Framing and Motivation (3)
8. Summary Likert (7)

Cluster Results on 16 Questions



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.

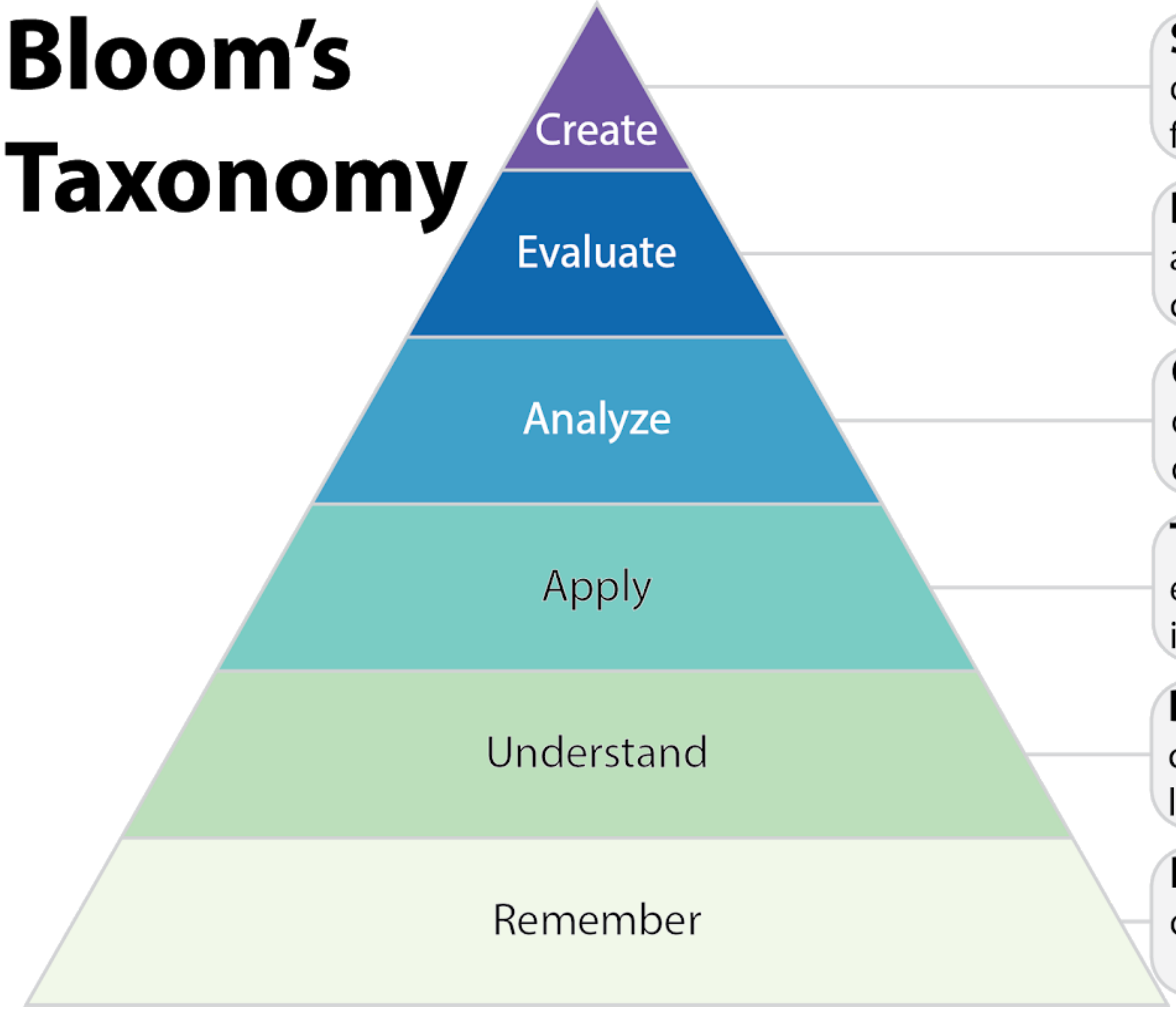
Plan the Learning Materials

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

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

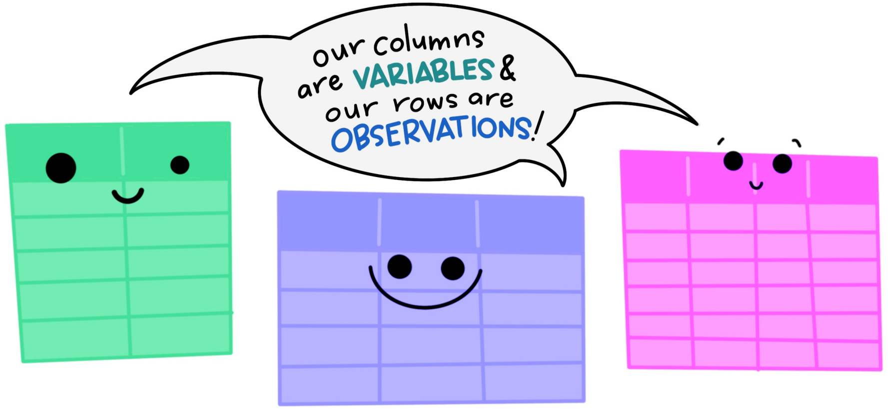
Repeat Information & Concepts
define, duplicate, list, memorize, state

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,.

Tidy Data

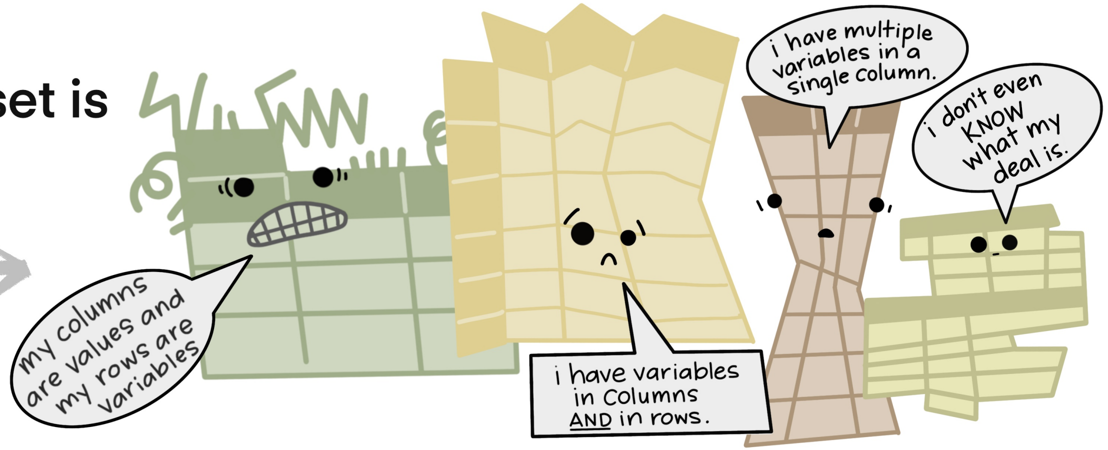
Data is messy in different ways

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

id	x	y	z
1	a	c	e
2	b	d	f

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

Example Data Science Problem

post_Q5.1: Cytomegalovirus (CMV) is a common virus that normally does not cause any problems in the body. However, it can be of concern for those who are pregnant or immunocompromised. Suppose you have the following Cytomegalovirus dataset of CMV reactivation among patients after Allogenetic Hematopoietic Stem Cell Transplant (HSCT) in an Excel sheet (first 10 rows shown below):

	A	B	C	D	E	F
1	ID	age	prior.radiation	aKIRs	donor_negative	donor_positive
2	1	61	0	1	recipient_positive	
3	2	62	1	5	recipient_negative	
4	3	63	0	3		recipient_positive
5	4	33	1	2	recipient_positive	
6	5	54	0	6		recipient_positive
7	6	55	0	2		recipient_positive
8	7	67	0	1		recipient_positive
9	8	51	0	2		recipient_positive
10	9	44	1	2		recipient_positive
11	10	59	0	4	recipient_negative	

It is believed that the donor activating KIR genotype is a contributing factor for CMV reactivation after myeloablastive allogenetic HSCT. What variables are associated with CMV reactivation?

Data from: Peter Higgins (2021). medicaldata: Data package for Medical Datasets. R package version 0.1.0. https://github.com/higgi13425/medicaldata

Q1

- Load the excel sheet

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

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

Q2

- Filter the data for individuals over the age of 65

```
# pipes, data filtering, boolean conditions
dat %>%
  filter(age > 65)
```

Q3

- Save filtered dataset as an Excel file to send to a colleague

```
# saving intermediates for data pipelines
subset <- dat %>%
  filter(age > 65)

# using functions/methods
library(writexl)
subset %>%
  write_xlsx("./data/cmv_65.xlsx")
```

Q4

- Tidy the dataset so we have a donor CMV status and a patient CMV status in separate columns

Dirty

	A	B	C	D	E	F
1	ID	age	prior.radiation	aKIRs	donor_negative	donor_positive
2	1	61	0	1	recipient_positive	
3	2	62	1	5	recipient_negative	
4	3	63	0	3		recipient_positive
5	4	33	1	2	recipient_positive	
6	5	54	0	6		recipient_positive
7	6	55	0	2		recipient_positive
8	7	67	0	1		recipient_positive
9	8	51	0	2		recipient_positive
10	9	44	1	2		recipient_positive
11	10	59	0	4	recipient_negative	

Tidy

	A	B	C	D	E	F
1	ID	age	prior.radiation	aKIRs	donor_status	recipient_status
2	1	61	0	1	donor_negative	recipient_positive
3	2	62	1	5	donor_negative	recipient_negative
4	3	63	0	3	donor_positive	recipient_positive
5	4	33	1	2	donor_negative	recipient_positive
6	5	54	0	6	donor_positive	recipient_positive
7	6	55	0	2	donor_positive	recipient_positive
8	7	67	0	1	donor_positive	recipient_positive
9	8	51	0	2	donor_positive	recipient_positive
10	9	44	1	2	donor_positive	recipient_positive
11	10	59	0	4	donor_negative	recipient_negative

```r
# lists/vectors/selecting
# tidy data and recognize a melt/pivot_longer operation
# keyword arguments
tidy_dat <- dat %>%
  pivot_longer(starts_with("donor"), names_to = "donor_status", values_to = "recipient_status") %>%
  drop_na()
```

Q5

- Plot a histogram of the age distribution of our data

```r
library(ggplot2)

# plotting syntax
# layering
ggplot(tidy_dat, aes(x = age)) +
  geom_histogram()
```

Q6

- Fit a model (e.g., logistic regression) to see which variables are associated with patient CMV reactivation.

```r
# formula syntax
model <- glm(cmv ~ age + prior_radiation + aKIRs + donor_status,
             data = tidy_dat, family = "binomial")

# look at model results
summary(model)

# dataframe of coefficients
library(broom)
tidy(model)
```

Data Science is Different From Computer Science

Canterbury QuestionBank

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: C: Only the middle element will be found. The remaining elements will not be contained in the subranges that we narrow our search to.

Software engineering, with some ventures into software architecture and computing education: https://neverworkintheory.org/

Adapt From Computer Science Education

“DataFrame” objects are not standard computer science data structures

Existing Data Science Book TOC: R + JS + Stats

R for Data Science

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

…

Ch 21. iteration

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**
12. Creating a Server
13. Testing
14. **Using Data-Forge**
15. Capstone Project

OpenIntro Statistics

1. **Introduction to Data**
2. Summarizing data
3. Probability
4. Distributions of random variables
5. Foundations of inference
6. Inference for categorical data
7. Inference for numerical data
8. Introduction to linear regression
9. Multiple and logistic regression

Existing Data Science Book TOC: Python

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**
8. Plotting and Visualization
9. Data Aggregation and Group Operations
10. Time Series
11. Financial and Economic Data Applications
12. Advanced NumPy

Appendix: Python Language Essentials

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

Existing Data Science Book TOC: My Own Work

Pandas for Everyone

1. Pandas DataFrame Basics
2. Pandas Data Structures
3. Introduction to Plotting
4. Data Assembly
5. Missing Data
6. **Tidy Data**
7. Data Types
8. Strings and Text Data
9. Apply
10. Groupby Operations: Split–Apply–Combine
11. The datetime Data Type
12. Linear Models
13. Generalized Linear Models
14. Model Diagnostics
15. Regularization
16. Clustering

ds4biomed

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

Conference Workshop

1. Introduction
2. **Tidy Data**
3. Functions
4. Plotting/Modeling

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

Rest of the Committee

Dave Higdon

Statistics
Department Head

Alex Hanlon

Statistics
CBHDS
iTHRIV BERD

Nikki Lewis

Honors College
Computational Research Grant

Additional Resources

- Data Organization in Spreadsheets, Karl W. Broman & Kara H. Woo
  - https://www.tandfonline.com/doi/full/10.1080/00031305.2017.1375989
- Examples of other learner personas
  - Rstudio Learner Personas: https://rstudio-education.github.io/learner-personas/
  - The Carpentries Learner Profiles: https://software-carpentry.org/audience/
- Creating your own personas
  - Zagallo, Patricia, Jill McCourt, Robert Idsardi, Michelle K Smith, Mark Urban-Lurain, Tessa C Andrews, Kevin Haudek, et al. 2019. “Through the Eyes of Faculty: Using Personas as a Tool for Learner-Centered Professional Development.” CBE—Life Sciences Education 18 (4): ar62.
- Bloom's Taxonomy
  - Bloom's Taxonomy Verb Chart: https://tips.uark.edu/blooms-taxonomy-verb-chart/
- Teach like a Champion
  - Version 2.0's 62 Techniques: https://teachlikeachampion.com/wp-content/uploads/Teach-Like-a-Champion-2.0-Placemat-with-the-Nanango-Nine.pdf

Thanks!

Slides: https://speakerdeck.com/chendaniely/education-and-pedagogy-of-domain-specific-learning-materials-using-learning-personas

Repo: https://github.com/chendaniely/jsm-2021-learner_personas

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

Appendix

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

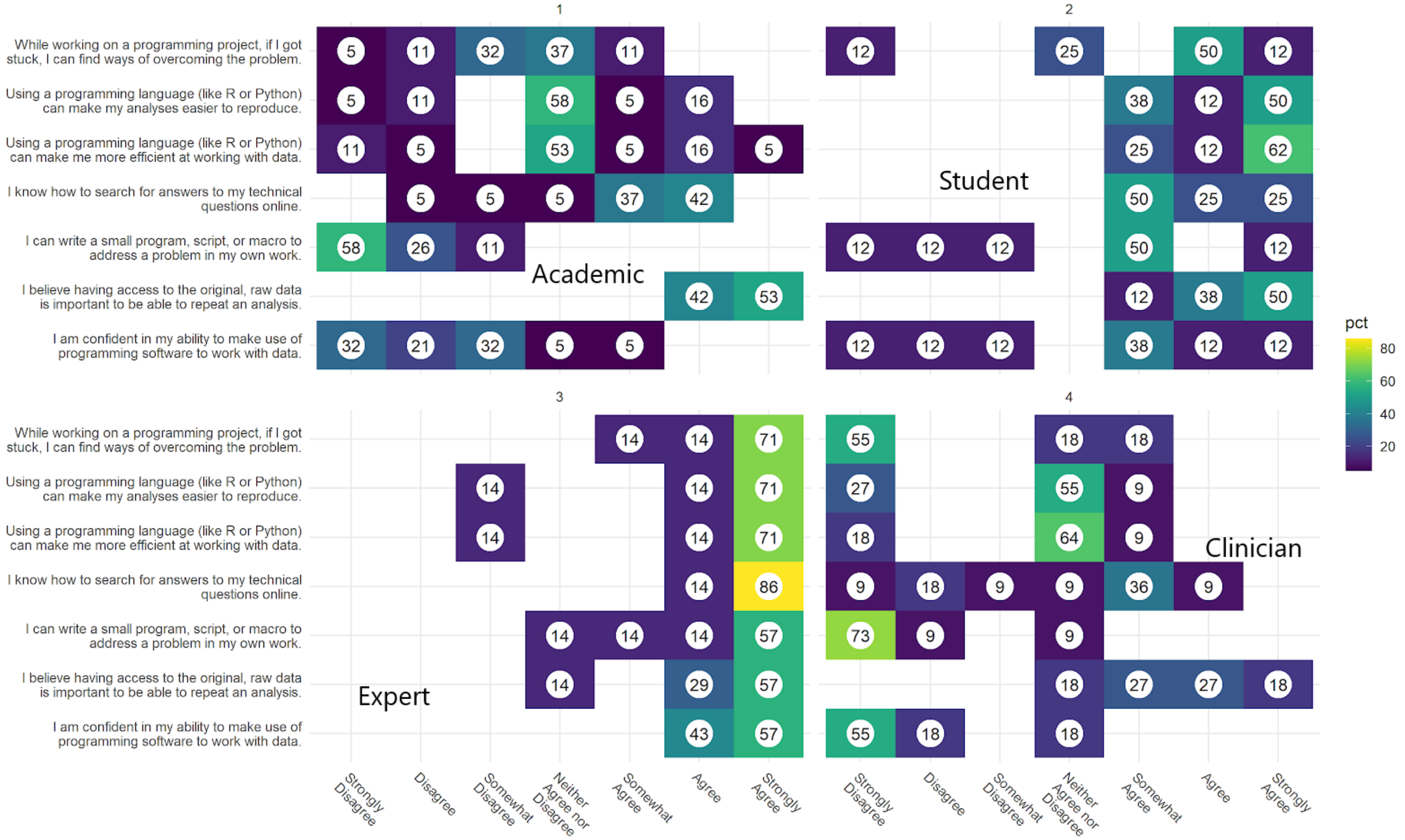
Song, I.-Y., and Zhu, Y. (2016). Big data and data science: What should we teach? Expert Systems, 33(4), 364–373. https://doi.org/10.1111/exsy.12130

Representative Questions

- Q6.2: If you were given a dataset 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 in increased relative risk or odds of hypertension? Any type of model will suffice.
  - 4 point scale
  - If you don't know where to start, you may be a novice
- Q3.3: How familiar are you with interactive programming languages like Python or R?
  - 7 point scale
  - If you have at least installed it and done simple examples, you may be more of an expert
- Q4.4: Do you know what "long" and "wide" data are?
  - 4 point scale
  - If you have heard of the term you may be a student

Summary Likert Questions

1. While working on a programming project, if I got stuck, I can find ways of overcoming the problem.
2. Using a programming language (like R or Python) can make my analysis easier to reproduce.
3. Using a programming language (like R or Python) can make me more efficient at working with data.
4. I know how to search for answers to my technical questions online
5. I can write a small program, script, or macro to address a problem in my own work.
6. I believe having access to the original, raw data is important to be able to repeat an analysis.
7. I am confident in my ability to make use of programming software to work with data.


1
2
3
4
Academic
Student
Expert
Clinician
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