Data Science Workshops for Biomedical Professionals

Committee Meeting #0

August 6, 2020

Daniel Chen, GBCB, Virginia Tech

VIRGINIA TECH™

HI!

Daniel Chen

- MPH in Epidemiology, Columbia University, 2014
    - Thesis: Agent-based models
    - Computational psychology: spread of ideas in a social network
    - Took my first data science class
    - Rachel Schutt, Kayur Patel, Jared Lander
    - Software-Carpentry workshop attendee
    - Adviser: Mark Orr, Columbia--> Virginia Tech (SDAL)

Virginia Tech

- PhD Student in Genetics, Bioinformatics, and Computational Biology (GBCB)
    - https://gbcb.graduateschool.vt.edu/
    - Matriculated Spring 2015
    - Interdisciplinary Graduate Education Program (IGEP) by Dean Karen DePauw
    - Program Director: Liwu Li, Biological Sciences
    - Admissions Committee Chair: T. M. Murali, Computer Science & Applications
    - Program Liaison: Dennie Munson, Interdisciplinary Graduate Programs

VIRGINIA TECH

Dissertation Disruption

- Mark -> NDSSL
- Dan student -> Faculty/Staff
- Worked as a data engineer
  - On many projects, but none were dissertation specific
  - Something to do with health care
  - Most promising project: "Smart Scatter"
  - Worked with Dave Higdon and Ian Crandall
- Data Science for the Public Good (SDAL)
- https://bi-sdal.github.io/training/
- https://chendaniely.github.io/training_ds_r/
- Biocomplexity Institute VT -> UVA

Software Carpentry

- Non-profit organization aimed to teach researchers the programming skills they need for research

- Co-founded by Greg Wilson

- Joined 2014

- Active instructor until 2017

- All my technical teaching experience was from the workshops I taught

- Software-Carpentry + Data Carpentry + Library Carpentry = The Carpentries

Building off the Carpentries

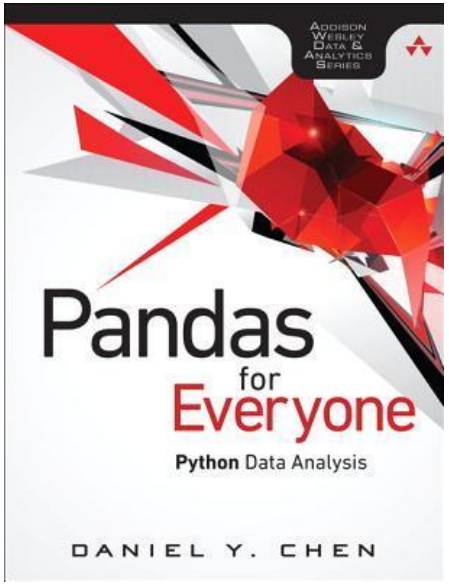
- Authored "Pandas for Everyone"

- Pearson liveLessons

- Python and Git

- DataCamp classes (but we don't talk about those)

- tl;dr: https://twitter.com/johncassil/status/1278685420595920897



VIRGINIA TECH

RStudio Intern Summer 2019

- RStudio Education team
- Worked on the {gradethis} package
- Code grader that can return formative error messages
- Made to be used with {learnr} interactive documents to create interactive lessons
- Gave a talk for Max Khun at the Non-clinical Biostatistics Conference (NCB)
- Got to TA Allison Hill's {knitr} workshop at R/Medicine
    - Met Peter Higgins (UMish) and Stephan Kadauke (CHOP)

VIRGINIA TECH™

What am I good at and interested in

- Teaching technical computing

- Healthcare/medicine

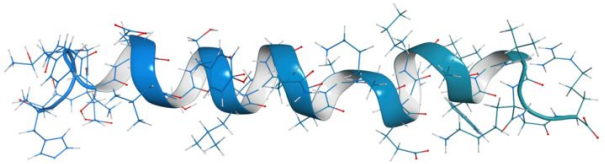
- Andi Ogier, Director Data Services -> Anne Brown

- Anne Brown is a recent GBCB faculty member

Anne Brown -- Brown Experiential Learning

Bevan Brown Lab

https://bevanbrownlab.com/



DataBridge

https://www.databridge.dev/




VIRGINIA TECH

Anne Brown

- Officially part of the Library

- Affiliate in Biochemistry

- Academy of Integrated Sciences

    - https://www.ais.science.vt.edu/

    - Majors: Computational Modeling and Data Analytics (CMDA), Nanoscience, and Systems Biology

    - Minors: Data and Decisions, Integrated Science Curriculum, Science, Technology, & Law

- David Bevan was her mentor and created the GBCB program

- Expertise in conceptualizing and executing pedagogical studies in experiential learning

- Invested in Data Education

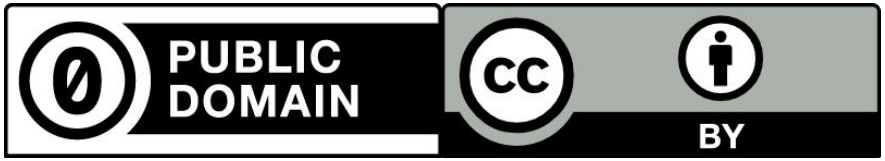
- VT Libraries

    - Commitment with/to the Carpentries

    - Open@VT (https://blogs.lt.vt.edu/openvt/)

VIRGINIA TECH™

Deliverables



1. Identify and create learner personas for the biomedical community

2. Create a set of CC-0 lessons for the biomedical community

    - Carpentries-inspired

    - Carpenties Incubator: Data Science for Practicing Clinicians

    - https://carpentries-incubator.github.io/Data-Science-for-Docs/

    - Too much emphasis on Medical Doctors

Phases + research questions

- IRB 20-537: Data Science Workshops for Biomedical and Health Professionals: Persona Identification and Workshop Assessment

1. Pre-workshop student self assessment survey to create learner personas- https://github.com/chendaniely/dissertation-irb/blob/master/irb-20-537-data_science_workshops/survey-01-pre_workshop_self_assessment.pdf

2. Pre/post workshop survey to assess the workshop materials

3. Long-term workshop survey (6 months out) to see if the materials helped with fundamental knowledge to learn more on their own

Research questions

3 main questions:

1. Does the biomedical community have different types of learners? How will their needs differ in the creation of learning and training materials?

- Phase 1 questionnaire will go through validity and respondent clustering to identify personas

Research questions

2. Does having interactive feedback with informative error messages in formative and summative assessment questions improve the learner's ability to learn and keep learning?

- Using tools like {learnr} with the {gradethis} and {pygradethis} library to create training materials.

- Pre/post workshop surveys can confirm existing knowledge along with results from the interactive questions in the workshop.

Research questions

3. Does tailored workshop materials help learners retain knowledge, use the tools, and continue self-learning?

- Combining all the surveys to determine "efficacy"

In more detail

1. Creating learner personas will create better educational content because they will be more tailored to the needs of the students.
2. Learning how to program data analysis will allow learners to feel like they can do more with their data.
3. Learning basic data literacy and data science skills can empower health/biomed workers and be more proactive in making more educated decisions.
4. Workshops with an eye towards tidy data principles will better transition students out of a spreadsheet program into programming.
5. Workshops will help medical professionals curate better data for research.
6. Workshops will help medical professionals work with data outside of a spreadsheet program.

What is "effective"? -- 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

Samples

Snowball sample from relavant listservs on campus

1. Dennie Munson -- igep

2. Taryn Luoma (taryn1@vtc.vt.edu) -- fbri + iThriv

3. Nathaniel Porter (ndporter@vt.edu) -- swc tmbh students (n=3?)

4. Andrea Green (greena15@exchange.vt.edu) vetmed

5. Hannah Menefee (hmenefee@vt.edu) -- mph

Preliminary Results

1. **A LOT** of Excel

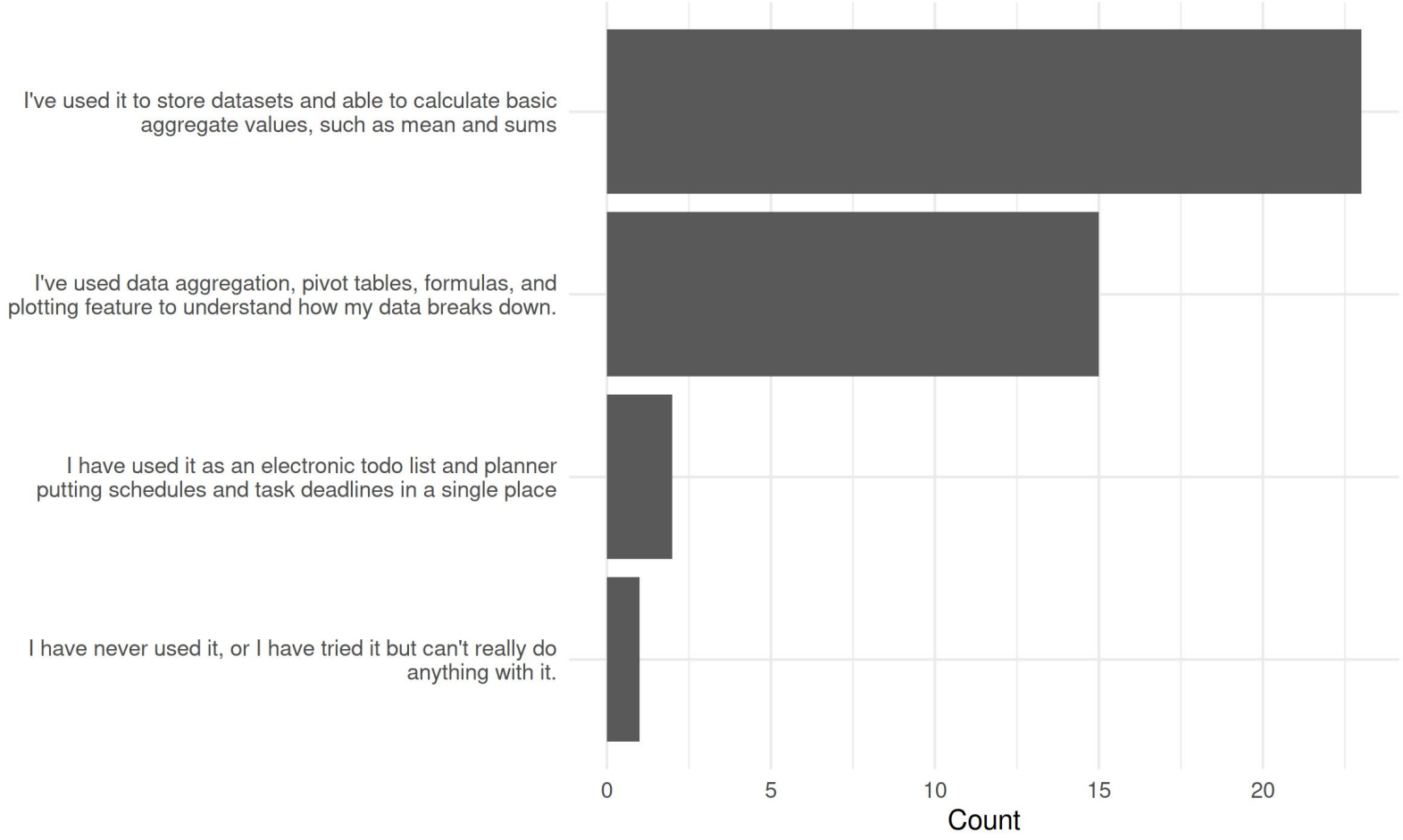
    - Very basic usage

    - Do not use other specialized software for tasks, mainly just Excel

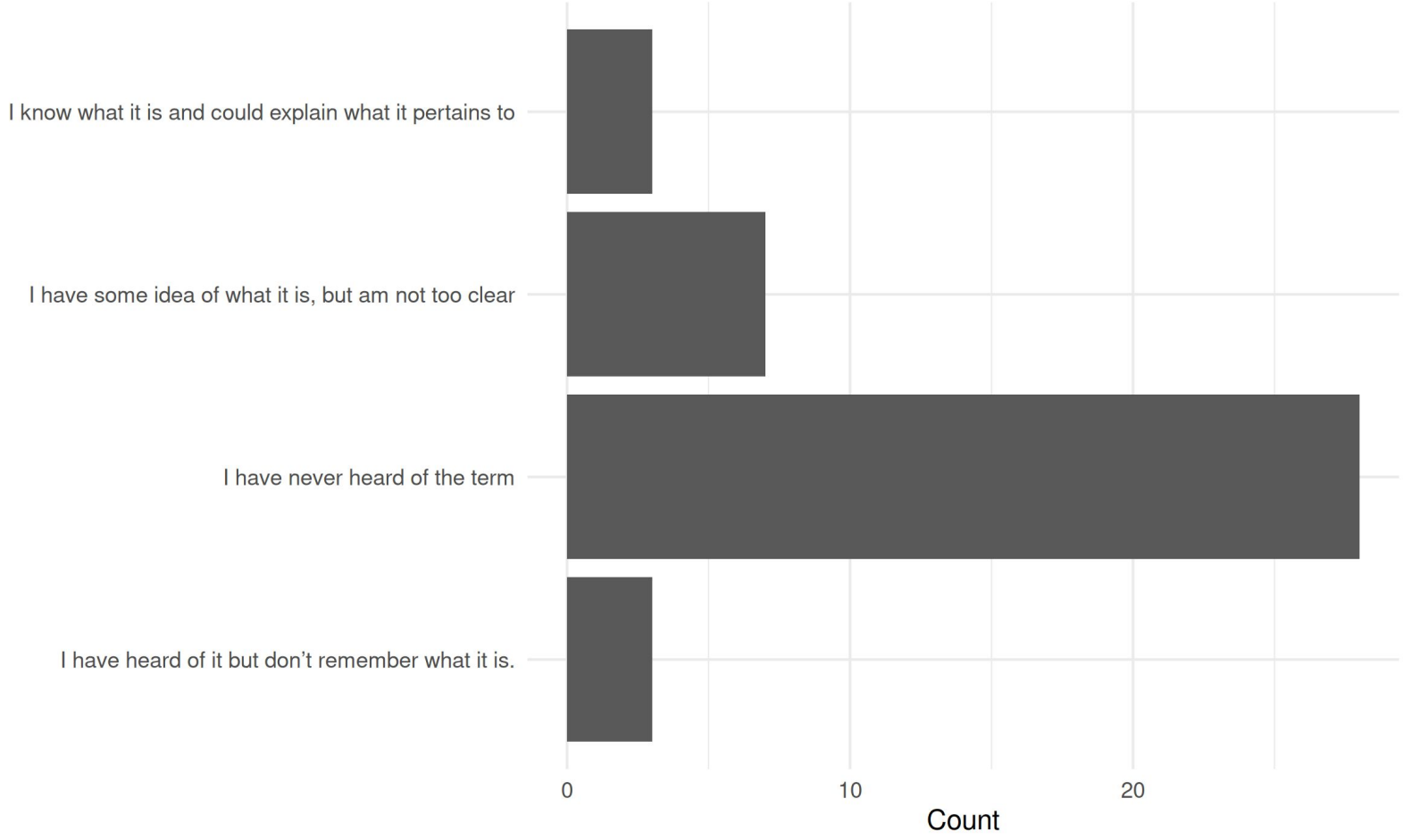
2. No knowledge of "tidy data"

    - Fundamental to processing and cleaning data

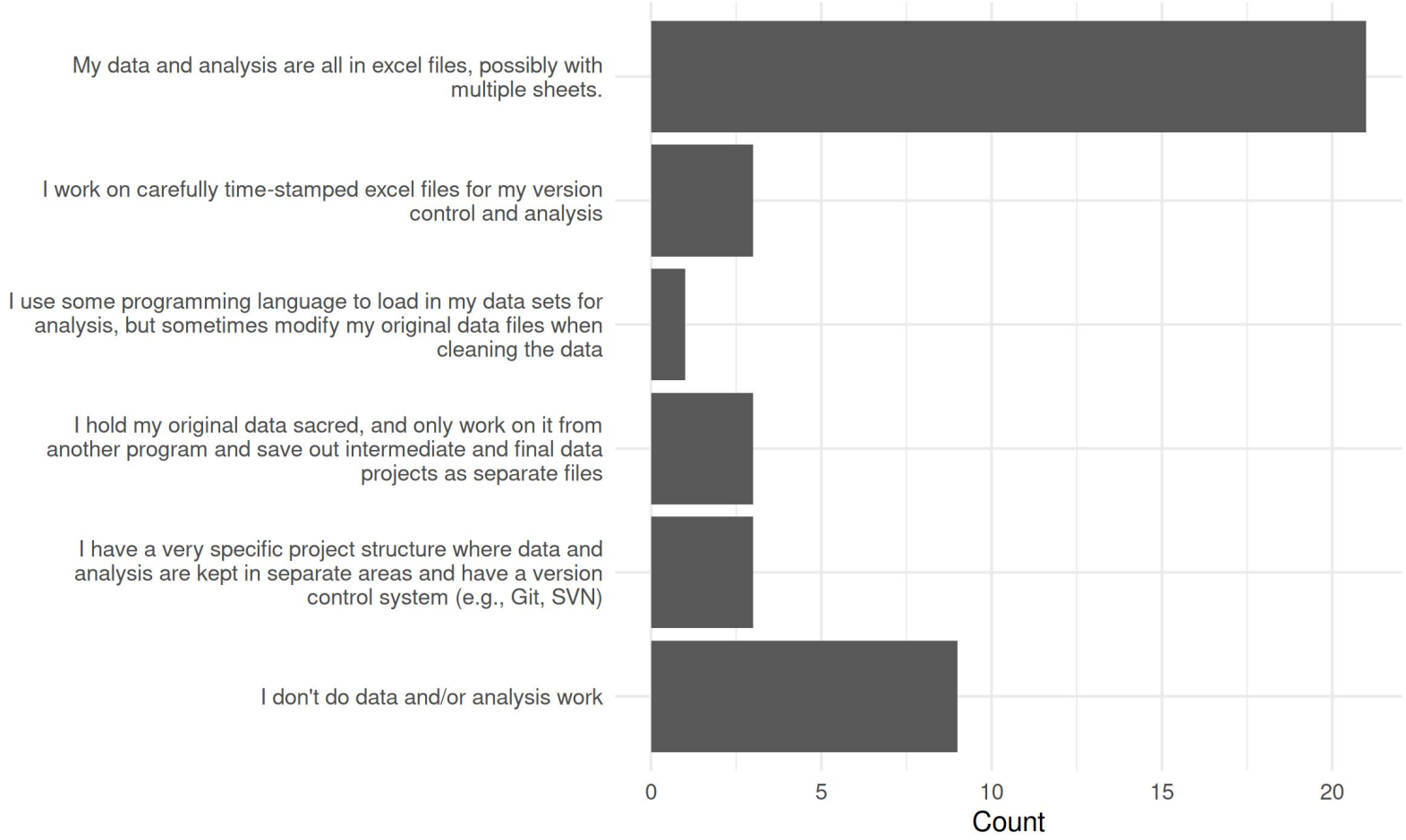
How familiar are you with Microsoft Excel?

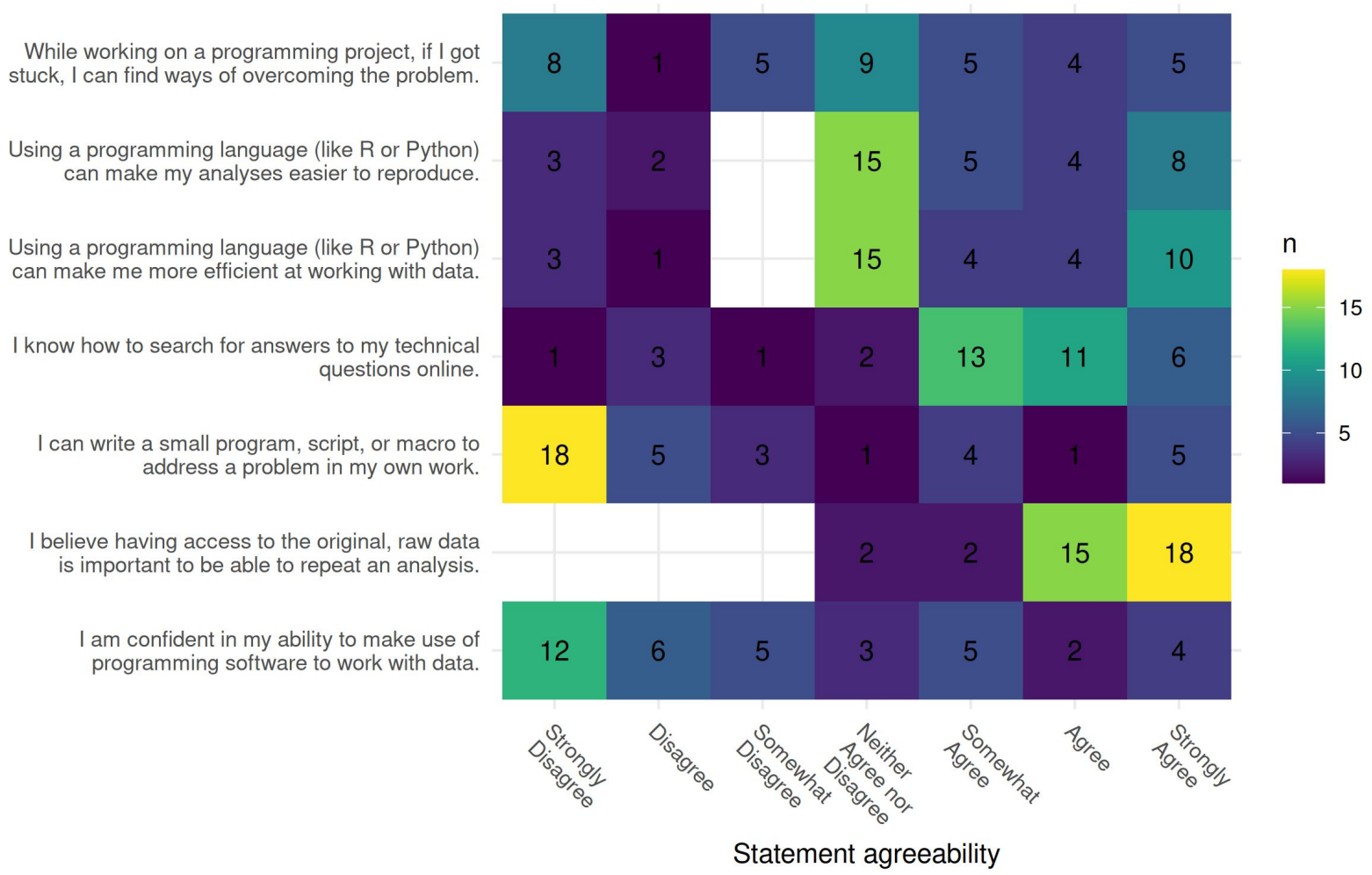


Are you familiar with the term "tidy data"?



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



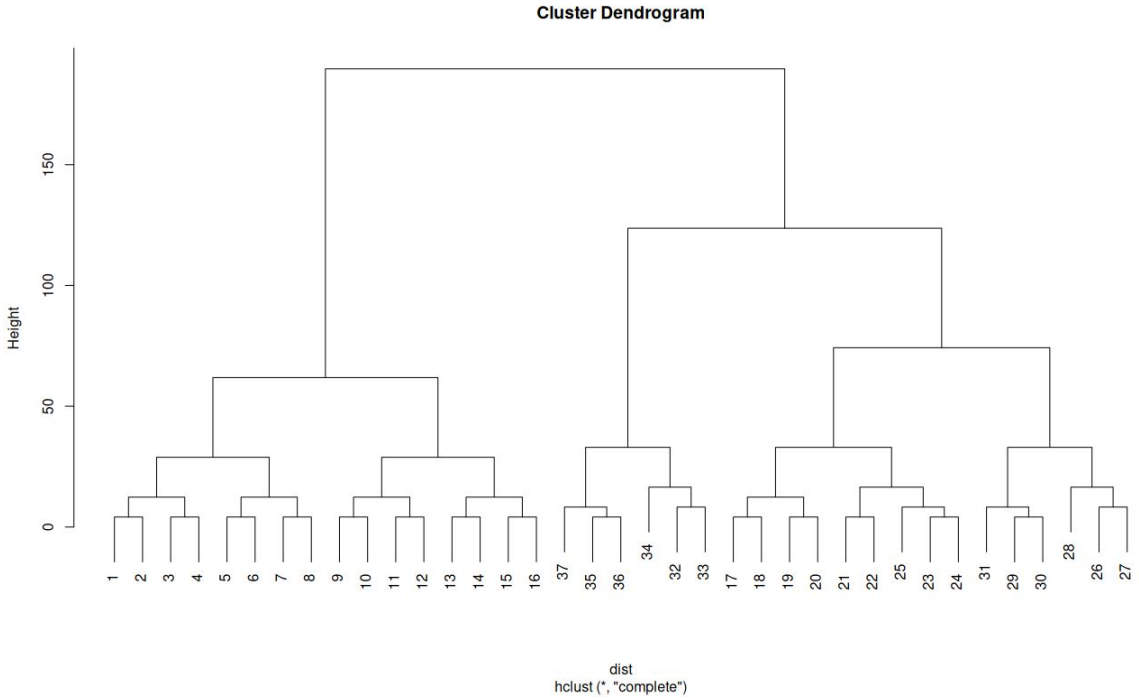

While working on a programming project, if I got stuck, I can find ways of overcoming the problem.
Using a programming language (like R or Python) can make my analyses easier to reproduce.
Using a programming language (like R or Python) can make me more efficient at working with data.
I know how to search 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.
8 1 5 9 5 4 5
3 2 15 5 4 8
3 1 15 4 4 10
1 3 1 2 13 11 6
18 5 3 1 4 1 5
2 2 15 18
12 6 5 3 5 2 4
Strongly Disagree
Disagree
Somewhat Disagree
Neither Agree nor Disagree
Somewhat Agree
Agree
Strongly Agree
Statement agreeability
n
15
10
5

Personas

- Dendogram shows 3 or 5 groups

- Groups in sequential order is suspicious

- Seems to be from when the surveys were sent out to the listservs



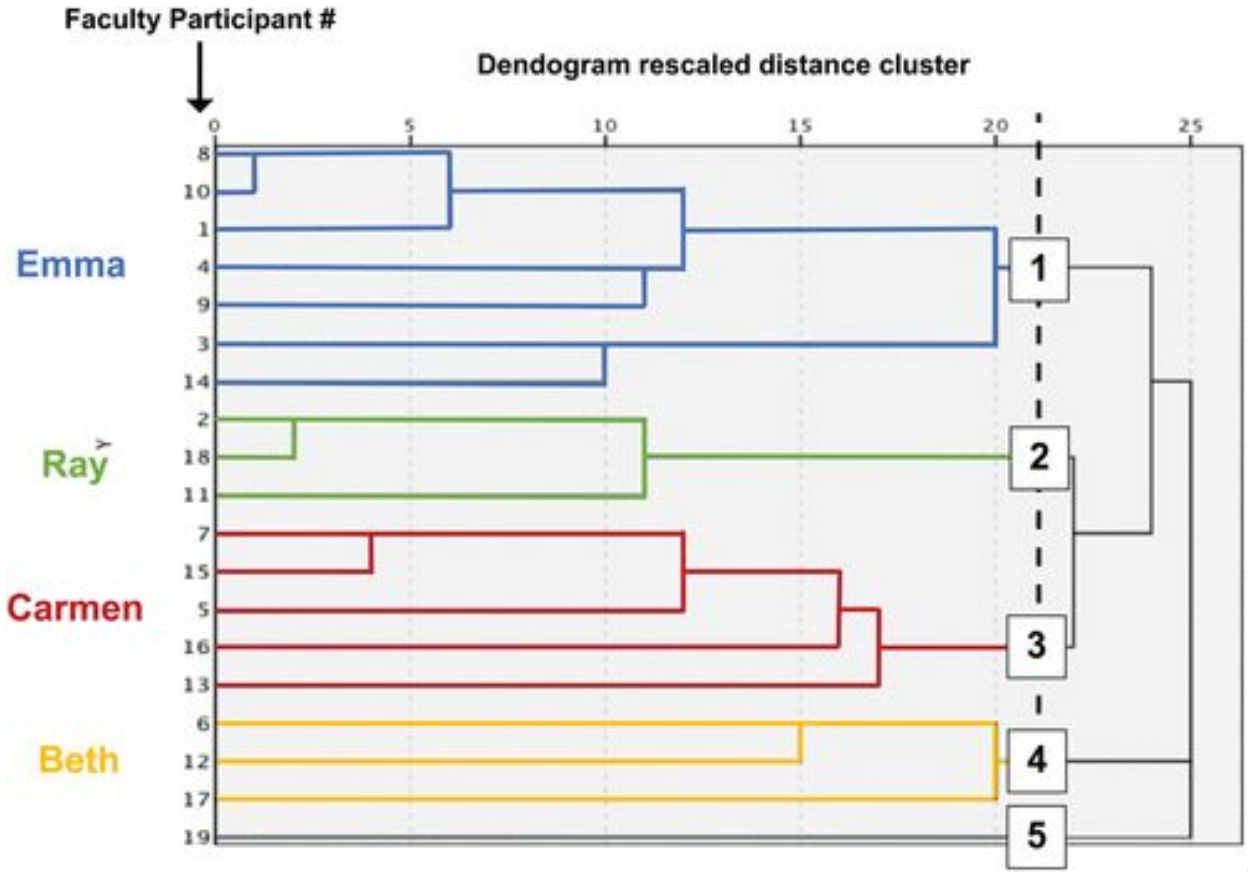
VIRGINIA TECH™

Creating Personas

- Through the Eyes of Faculty: Using Personas as a Tool for Learner-Centered Professional Development

    - https://www.lifescied.org/doi/10.1187/cbe.19-06-0114

- Methods that combine hierarchical agglomerative cluster analysis with chi-square values or squared Euclidian distance values and complete or average linkage



RStudio learner Personas

- https://rstudio-education.github.io/learner-personas/

Persona	In Brief	Domain Knowledge	Statistics Knowledge	Programming Knowledge
Anya Academic	A professor who needs training for her research and to pass on to students.	expert	competent	competent
Celine Certified	A certified RStudio instructor.	competent	competent	competent
Exton Excel	A proficient Excel user working in industry who wants to switch to R.	competent	novice	novice
Jacqui Ofalltrades	A data science generalist at a small consulting company.	expert	expert	expert
Katrin Keener	An R enthusiast.	competent	competent	competent
Larry Legacy	A reluctant learner who would really rather just keep using the tools he knows.	expert	expert	novice
M'shelle Manager	An ex-programmer who now leads a team and needs to make decisions about tool adoption and training.	competent	novice	competent
Nang Newbie	An undergraduate student without statistical knowledge, programming skills, and real-world experience.	novice	novice	novice
Toshi Techsupport	A sys admin who has to support data scientists.	expert	novice	expert

Paper topics

1. Learner Personas

2. Feedback with informative error messages during a workshop

3. Overall assessment of the workshop for "effectiveness"

Committee -- Currently

1. Anne Brown - Library

2. Alex Hanlon - Biostatistics

Committee -- Need to reconfirm

3. Dave Higdon - Statistics

Committee -- Other people

1. Dennis Kafura - Computer Science - Computer science education

2. Margaret Ellis - Computer Science - Digital education

3. Jane Robertson Evia - Statistics - Statistics education

Committee -- Other people

4. Nikki Lewis - Honors College

5. Amy Nelson - History - Pedagogical Practices in Contemporary Contexts
   - Future Professoriate Certificate

Overall impact

- "Integrating scientific programming in communities of practice for students in life science"
    - https://dl.acm.org/doi/10.1145/3332186.3333040

- People in the life sciences are afraid to take courses in CS and Stats

Overall impact

- Help augment the training from Center for Biostatistics and Health Data Science (CBHDS)
    - https://biostat.centers.vt.edu/

- Goal is not to teach statistics but the data literacy side of data science

    - Managing and "cleaning" data

    - Be able to better communicate with analysts and statisticians

Overall impact

- Carpentries-inspired training materials to help scale
    - Have a CC-BY set of training materials that can be remixed by others
    - VT has a core group of Carpentries certified instructors
    - Anne is just about finishing up her checkout

Overall Impact

- Build a community of practice

- "Building a local community of practice in scientific programming for life scientists"
    - https://journals.plos.org/plosbiology/article?id=10.1371/journal.pbio.2005561

- "If you build it, they will come...but then what? Facilitating communities of practice in R"

    -https://rstudio.com/resources/rstudioconf-2020/if-you-build-it-they-will-come-but-then-what-facilitating-communities-of-practice-in-r/