Making MIDFIELD More Accessible: A Workshop for R Beginners

Richard A. Layton
Mechanical Engineering
Rose-Hulman Institute of Technology
Terre Haute, IN 47803
layton@rose-hulman.edu

Russell A. Long
Engineering Education
Purdue University
West Lafayette, IN 47907
ralong@purdue.edu

Susan M. Lord
Engineering
University of San Diego
San Diego, CA 92110
slord@sandiego.edu

Matthew W. Ohland
Engineering Education
Purdue University
West Lafayette, IN 47907
ohland@purdue.edu

Marisa K. Orr
Engineering and Science Education
Clemson University
Clemson, SC 29634
marisak@clemson.edu

Nichole Ramirez
Engineering Education
Purdue University
West Lafayette, IN 47907
nramire@purdue.edu

***Abstract*—This workshop introduces data and tools for investigating undergraduate persistence metrics using R. Student record data are from MIDFIELD, a database of registrars' data from US institutions. The stratified data sample includes demographic, term, course, and degree information for 98,000 students from 1987 to 2016. The midfieldr package provides functions for determining persistence metrics such as graduation rates or program stickiness and for grouping findings by institution, program, sex, and race/ethnicity. The goal of the workshop is to share our data, methods, and metrics for intersectional research in student persistence. The workshop is designed for R beginners.**

I. Introduction

The Multiple-Institution Database for Investigating Engineering Longitudinal Development (MIDFIELD) is a student record data set for studying how undergraduate students maneuver through their curricula [1]. The MIDFIELD database includes approximately one million undergraduate students at sixteen US institutions from 1987 to 2016.

The R package **midfieldr** and its associated data packages provide a stratified sample of MIDFIELD data: registrars' student records for 98,000 undergraduate students with proportional representation by institution, program, transfer status, sex, and race/ethnicity.

II. Description

The goal of the workshop is to make MIDFIELD more accessible to the FIE community via **midfieldr**. On completing the workshop, participants should be able to

- Describe key variables in the MIDFIELD data
- Select academic programs and populations to study
- Compute persistence metrics (e.g. graduation rate)
- Graph persistence metrics
- Explain key features of effective data displays

Participants should be sufficiently familiar with their operating systems to install software and navigate directories, but prior experience with R is not required.

Participants work on self-paced software tutorials, learning some basic R plus the specialized **midfieldr** functions for turning raw MIDFIELD data into persistence metrics. In the graduation rate tutorial for example, we select specific academic programs to study, determine the numbers of students who matriculate and graduate in a program, compute graduation rates, and graphically compare the results across different programs, disaggregating by sex and race/ethnicity. Participants learn some basic R, but the majority of the analysis is performed using functions in the package specialized for this data and type of analysis.

The agenda also includes an interactive session demonstrating contemporary principles of effective data display. The goal of this segment is to provide the rationale underlying the default graph types used in the package.

The workshop aligns with FIE faculty development goals by improving participants abilities to explore data and communicate findings using an intersectional approach. The workshop goals also align with the 2018 FIE conference theme of "Fostering Innovation through Diversity."

III. Outline

Workshop activities include think-pair share, active learning, demonstration, discussion, and self-paced software tutorials. Our 3-hour agenda includes:

Min	Topic
10	Introductions
30	Designing effective data displays (interactive)
20	Getting started with R (interactive)
20	Break
20	Exploring the structure of the MIDFIELD data sample
45	Using R for persistence metrics (self-paced tutorial)
15	Additional persistence metrics using R
20	Discussing next steps & assessing the workshop

978-1-5386-1174-6/18/$31.00 ©2018 IEEE

IV. Why R?

R is an open source language and environment for statistical computing and graphics [2], currently ranked by IEEE as the 6th most popular programming language (Python, C, and Java are the top three) [3]. If you are new to R, some of its best features (paraphrasing Wickham [4]) are:

- R is free, open source, and available on every major platform, making it easy for others to replicate your work.
- More than 12,500 open-source R packages are available (April 2018). Many are cutting-edge tools.
- R packages provide deep-seated support for data analysis, e.g., missing values, data frames, and subsetting.
- R packages provide powerful tools for communicating results via html, pdf, docx, or interactive websites.
- It is easy to get help from experts in the R community.

RStudio, an integrated development environment (IDE) for R, includes a console, editor, and tools for plotting, history, debugging, and workspace management as well as access to GitHub for collaboration and version control [5].

The fundamental unit of shareable code in R is the *package*. A package bundles code, data, documentation, and tests, and is easily shared with others [4]. In this workshop, we work with the open-source **midfieldr** package and associated data packages, all of which are available for Windows, MacOSX, and Linux platforms from the Comprehensive R Archive Network (CRAN) [6].

V. Sample Result

A common example of a persistence metric is graduation rate: the fraction of starters in a major who graduate in the major. Figure 1 shows graduation rates for students in three majors—Industrial Engineering (ISE), Mechanical Engineering (MCE), and Electrical Engineering (ELE)—disaggregated by sex and race/ethnicity.

The rows of this graph are ordered by the mean graduation rate of each student group across all three panels, allowing us to detect visual anomalies. For example, Hispanic Female students in ISE graduate at a much higher rate than expected given their average rate across all three majors.

Data of this type are *multiway data* and the graph is a *multiway dot plot* [7]. Because persistence data often have a multiway structure, we regularly use this graph type for data exploration and presentation [8] and we include it as a default graph type in **midfieldr**.

VI. Summary

Participants leave with the knowledge and skills to access a MIDFIELD student-record data sample and to determine persistence metrics such as graduation rates through a lens of intersectionality. Participants who are new to R will take away the ability to use R at a basic level to import data, transform data, and graph results. They will have "made a start" with R. Once introduced to **midfieldr**, participants can continue to learn about data and tools for investigating undergraduate persistence metrics using additional self-paced tutorials publicly available on GitHub [9].

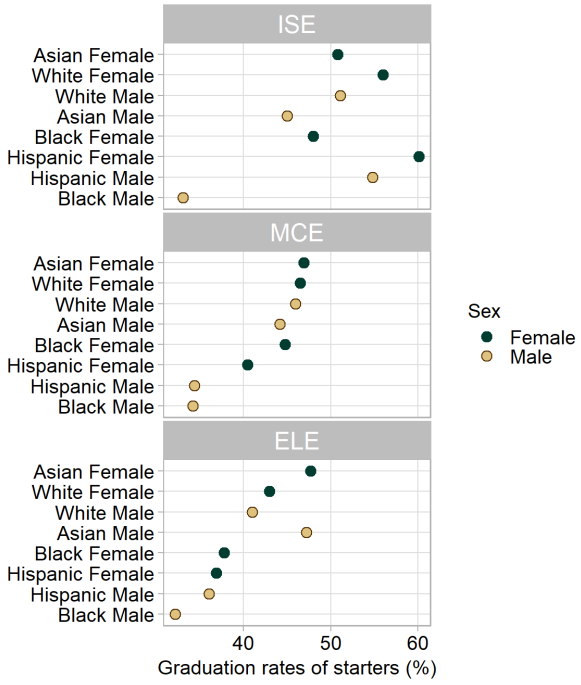

Fig. 1. Graduation rates of students in three engineering majors: an example of a multiway graph. Data source: MIDFIELD.

VII. Facilitators

Richard Layton is the MIDFIELD Director of Data Display and Professor of Mechanical Engineering at Rose-Hulman. He is the lead developer of the R packages used in this workshop. Dr. Layton has considerable experience facilitating workshops for data visualization, including a workshop for R beginners at the 2014 FIE in Madrid.

Russell Long is MIDFIELD Managing Director and Data Steward. He developed the stratified data sample for the R packages used in this workshop. Mr. Long is a SAS expert with over twenty years of experience in institutional research and assessment.

Susan Lord is Director of the MIDFIELD Institute and Professor and Chair of Engineering and Professor of Electrical Engineering at the University of San Diego. She is a Fellow of the IEEE and the ASEE. Dr. Lord has considerable experience facilitating workshops including the National Effective Teaching Institute (NETI) and special sessions at FIE.

Matthew Ohland is the MIDFIELD Director & Principal Investigator. He is Professor and Associate Head of Engineering Education at Purdue University and a Fellow of IEEE, ASEE, and AAAS. Dr. Ohland has considerable experience facilitating workshops including the NETI and CATME training.

Marisa Orr is the MIDFIELD Associate Director and Assistant Professor in Engineering and Science Education with a joint appointment in Mechanical Engineering at Clemson university. She is a recipient of the 2009 Helen Plants Award for the best nontraditional session at FIE (Enhancing Student Learning Using SCALE-UP Format).

Nichole Ramirez is the MIDFIELD Associate Director of Policy Analysis and a postdoctoral researcher in the School of Engineering Education at Purdue University. Dr. Ramirez supervisors and mentors the team of undergraduate researchers who conduct policy analysis and contribute to the R package development.

ACKNOWLEDGMENT

The authors would like to thank Fernando Martinez for his contributions to the package development. Funding provided by the National Science Foundation Grant 1545667 "Expanding Access to and Participation in the Multiple-Institution Database for Investigating Engineering Longitudinal Development."

REFERENCES

[1] M. Ohland and R. Long, "The Multiple-Institution Database For Investigating Engineering Longitudinal Development: An experiential case study of data sharing and reuse," *Advances in Engng Educ*, vol. 5, no. 2, June 2016.

[2] R Core Team, *R: A language and environment for statistical computing*, R Foundation for Statistical Computing, Vienna, Austria, 2018. [Online]. Available: https://www.R-project.org/

[3] S. Cass, "The 2017 top programming languages," *IEEE Spectrum*, July 18 2017. [Online]. Available: https://spectrum.ieee.org/computing/software/the-2017-top-programming-languages

[4] H. Wickham, *Advanced R*, ser. Chapman & Hall/CRC The R Series. Taylor & Francis, 2014.

[5] RStudio Team, *RStudio: Integrated development environment for R*, RStudio, Inc., Boston, MA, 2015. [Online]. Available: http://www.rstudio.com/

[6] *The Comprehensive R Archive Network*, 2018-04-22. [Online]. Available: https://cran.r-project.org/

[7] W. S. Cleveland, *Visualizing Data*. Hobart Press, 1993.

[8] R. Layton, S. Lord, and M. Ohland, "Reasoning about categorical data: Multiway plots as useful research tools," in *ASEE Annual Conference*, Austin, June 2009.

[9] *midfieldr: R tools for investigating longitudinal US student record data*, expected release 2018-07. [Online]. Available: https://github.com/MIDFIELDR/midfieldr