Workshop Title: **Computing infrastructure and curriculum design for introductory data science**

Presenter:

Mine Çetinkaya-Rundel (Contact Person)
Department of Statistical Science
Duke University
214 Old Chemistry, Box 90251 Duke University
Durham, NC
Phone: +919-684-5956
mine@stat.duke.edu
mine-cr.com

Abstract: The goal of this workshop is to equip educators with concrete information on content and infrastructure for designing and painlessly running a modern data science course. This is a three-part workshop. Part 1 will outline a curriculum for an introductory data science course and discuss pedagogical decisions that go into the choice of topics and concepts, programming language (R) and syntax (primarily `tidyverse`), emphasis on literate programming for reproducibility (with R Markdown). Part 2 will discuss infrastructure choices around teaching data science with R: RStudio as an integrated development environment, cloud-based access with RStudio Cloud and Server, version control with Git, and collaboration with GitHub. Part 3 will focus on classroom management on GitHub (with `ghclass`) and automated feedback with continuous integration tools (e.g. Wercker). Workshop attendees will work through several exercises from the course and get first-hand experience with using the tool-chains and techniques described above. All workshop content, including teacher facing documentation and student facing course materials, will also be available to participants via datasciencebox.org.

Advertisement: Interested in teaching introductory data science? running your course on GitHub, and doing so efficiently? what first exposure to computing with R looks like? what the tidyverse is? If your answer is yes to any of these, this workshop is for you! We will outline a curriculum for introductory data science, highlight computing with R, and get hands on practice with efficiently running your course on GitHub.

Significance and Relevance of the Topic: With the emergence of data science as a field that is revolutionizing industries and academic, many universities are experiencing a surge of data science courses and programs. An introductory data science course for an audience of students with no computing background can be the gateway to further quantitative studies, or the last such course a student ever takes. In designing such a course faculty must consider the full breadth of topics and concepts data science encompasses as well as thrive to make technical infrastructure and toolkit choices with an eye towards minimizing frustration and improving adoption for both students and instructors. The goal of this workshop is to equip educators with concrete information on content and infrastructure for painlessly introducing modern computation into a data science curriculum.

Expected audience: CS educators teaching at the university level who are interested in teaching data science with R or would like to find out more about curricular and pedagogical decisions in the R environment. Workshop aims to draw 15-20 participants, but would be able to accommodate a larger number of participants if there is interest.

Space and Enrollment restrictions: N/A

Expertise of Presenter:
Mine Çetinkaya-Rundel is the Director of Undergraduate Studies and Associate Professor of the Practice in the Department of Statistical Science at Duke University. She also works as a Professional Educator at RStudio. Mine's work focuses on innovation in statistics and data science pedagogy, with an emphasis on computation, reproducible research, student-centered learning, and open-source education. Mine is also the creator and maintainer of datasciencebox.org, a repository for introductory data science education with R. Mine also works on the OpenIntro project, whose mission is to make educational products that are free,

transparent, and lower barriers to education. As part of this project she co-authored three open-source introductory statistics textbooks. She also teaches the popular Statistics with R MOOC on Coursera as well as numerous courses on DataCamp.

Mine has previously led a version of this workshop at the 2018 International Conference on Teaching Statistics. She has also presented at SIGCSE 2018 on using GitHub in the classroom.

In 2018 Mine received the 2018 Pickard Award for Statistics Education. She is also the recipient of the 2016 ASA Waller Education Award, 2015 JSM Best Paper Award in the Section on Teaching Statistics in the Health Sciences, and the 2014 Duke University David and Janet Vaughan Brooks Award for Teaching Excellence.

Rough Agenda:

1. Part 1: Curriculum design and pedagogical choices:
    - Curriculum for an introductory data science course
    - Programming with R and the `tidyverse`
    - Literate programming with R Markdown
2. Part 2: Infrastructure choices around teaching data science with R:
    - RStudio as an integrated development environment
    - Cloud-based access to R with RStudio Cloud and RStudio Server,
    - Version control with Git
    - Collaboration with GitHub.
3. Part 3: Classroom management on GitHub
    - `ghclass` for efficient class management
    - Automated feedback with continuous integration tools (e.g. Wercker)

Audio/Visual and Computer requirements:

Our workshop has the following technology requirements:

- Internet: wireless access
- Power: additional power outlets for at least 50% of attendees
- Projector: Digital projector with HDMI connectors
- Computers: Laptop required.
- Software: No special software needed
- Other: flipchart with pens

Other critical information: While the workshop content will focus on usage of R, many of the pedagogical takeaways will be language agnostic.