BST 270 Reproducible Data Science
Spring 2020, Mondays 3:45-5:15pm Kresge 205

Teaching Staff

Heather Mattie
Instructor of Data Science
Office: Building 1, 4th floor, room 421A
Email: hemattie@hsph.harvard.edu
Phone: (617) 432-5308
Office Hour: Mondays 5:15-6:15pm or by appointment

Aaron Sonabend
PhD Candidate, Biostatistics
asonabend@g.harvard.edu
Office hour: TBD

Purpose of this course

Reproducible research has become increasingly important in the biomedical sciences. The science community has recognized reproducibility is a growing challenge in basic, clinical and population sciences. Experimental design, data provenance, analytic methods and tools, and reporting science play a critical role in the biomedical research ecosystem to ensure scientific rigor, robustness and transparency. Statistical and computational methods and tools are fundamental for making scientific results reproducible.

Course Description and Structure

The central theme of the course will be to meet these scientific needs of reproducible science through training in reproducible research. The topics covered in this course include the fundamentals of reproducible science, case studies in reproducible research, data provenance, statistical methods for reproducible science, and computational tools for reproducible science. This is a blended course where students are introduced to course content online through videos and reading assignments, and then shown how to use the tools and methods described in the videos to conduct reproducible research.

In-class lectures will focus on applying the tools and methods introduced in the videos to reproduce a recently published scientific journal article:

[1] Boehm, J. K., Williams, D. R., Rimm, E. B., Ryff, C., & Kubzansky, L. D. (2013). Relation between Optimism and Lipids in Midlife. The American Journal of Cardiology, 111(10), 1425-1431.
http://doi.org/10.1016/j.amjcard.2013.01.292

In 1995, MIDUS survey data were collected from a total of 7,108 participants. The baseline sample was comprised of individuals from four subsamples: (1) a national RDD (random digit dialing) sample (n=3,487); (2) oversamples from five metropolitan areas in the U.S. (n=757); (3) siblings of individuals from the RDD sample (n=950); and (4) a national RDD sample of twin pairs (n=1,914). All eligible participants were non-institutionalized, English-speaking adults in the contiguous United States, aged 25 to 74. All respondents were invited to participate in a phone interview of approximately 30 minutes in length and complete 2 self-administered questionnaires (SAQs), each of approximately 45 pages in length. In addition, the twin subsample was administered a short screener to assess zygosity and other twin-specific information. With funding provided by the National Institute on Aging, a longitudinal follow-up of MIDUS I began in 2004. Every attempt was made to contact all original respondents

and invite them to participate in a second wave of data collection. Of the 7,108 participants in MIDUS I, 4,963 were successfully contacted to participate in another phone interview of about 30 minutes in length. MIDUS II also included two self-administered questionnaires (SAQs), each of about 55 pages in length, which were mailed to participants. The overall response rate for the SAQs was 81%. Over 1,000 journal articles have been written using MIDUS I and II data since 1995.

We will attempt to reproduce the findings of [1] and critique the reproducibility of the article as a class. This particular article focuses only on MIDUS II data, including biomarker data, and investigates the relationship between optimism and lipids. You can download the MIDUS II data and supporting codebook and other documents here. You can download the data in multiple formats. We will be using the R files in class and performing all data cleaning and analyses in R. You can download the biomarker data here.

Course Objectives

Upon successful completion of this course, you should be able to:

- Describe the fundamentals and importance of reproducible research
- Assess the reproducibility of other research
- Create a fully reproducible research project
- Develop new methods and tools for reproducible research

Credits

This is a 2.5 credit, Pass/Fail course. A minimum total score of 70% is needed to pass this course.

Course Materials

Course videos, electronic copies of course readings, rubrics, guidelines, notes/slides, useful website links and data sets will be posted on the edX course website and course GitHub repository. Students will need to create an edX account and sign-up for the course.

Optional Reading:

- Christopher Gandrud (2015), Reproducible Research with R and RStudio, 2nd Ed.
- Kitzes, Turek, Deniz (2017), The Practice of Reproducible Research: Case Studies and Lessons from the Data-Intensive Sciences, 1st Ed.

Course Repository

All course documents can be found on the course GitHub repository.

Grading

Homework, In-class Participation and Attendance (40%)

- Homework assignments will consist of viewing online course videos (edX videos)
- Throughout the course we as a class will be attempting to reproduce a recently published journal article in class. Students are expected to participate in class activities that will involve completing tasks related to reproducing the journal article. The reproducibility of the article will be discussed and critiqued in class.

- Class attendance is mandatory. A sign-in sheet will be provided at the beginning of each session to record attendance. Students will be excused from class in the event of a family emergency, medical issue, religious observance, or other extenuating circumstance, and should contact the instructor to inform them of their absence. A maximum of two absences is allowed without penalty to the class attendance grade and overall grade. A 20% deduction in class attendance grade will be taken for every additional missed class.

Case Study (20%)

- Each student will be responsible for presenting a case study from The Practice of Reproducible Research: Case Studies and Lessons from the Data-Intensive Sciences. Slides should be prepared detailing the case study and presentations should be 15-20 minutes long. Students will have the opportunity to choose which case study they would like to present.

Group project (40%)

- A group of students (minimum of 1, maximum of 3) will attempt to reproduce the findings of a recently published journal article that uses the MIDUS II data and critique its reproducibility. Each team will be assigned one of five possible papers - see options below. The project must be submitted in the form of a Jupyter notebook or RMarkdown file, and include an introduction, methods, results, and discussion/critique section with commented code interspersed. Visuals and schematics should be included if appropriate. All project documents, including the paper being reproduced, must also be uploaded to a GitHub repository. Only one repository is needed for each group. The repository must also include a README file describing the contents of the repository and how to reproduce all results. Groups should keep in mind the file and folder structure we will cover in class and make their process as automated as possible.

- Specific tasks:

  1. Reproduce all figures and tables
  2. Reproduce all analyses
  3. Critique the reproducibility of the paper:
     - Is the data publicly available?
     - Is the data easy/intuitive to access?
     - Is there a codebook and/or instructions about how the data and documentation is organized?
     - Are the file names intuitive?
     - Are the variable names intuitive?
     - Is the software used for analysis publicly available?
     - If the software is available, is it well commented?
     - Is there a toy example provided?
     - Are you able to reproduce the figures, tables and results presented in the paper?
     - Was there anything you think should have been made clearer, or explained in a different way?
     - Did you find any faults in the methods used in this paper? Would you have used more or different methods?
  4. Give a 10-minute group presentation on the last day of class - May 13th

- Journal Articles:

  1. Goodwin, R. D., Gotlib, I. H. (2004). Gender differences in depression: The role of personality factors. Psychiatry Research, 126(2), 135-142.
  2. Fujiwara, Takeo (2007). The role of altruistic behavior in generalized anxiety disorder and major depression among adults in the United States. Journal of Affective Disorders, 101(1-3), pp. 219-225.
  3. Wang, D., Gruenewald, T. (2017). The psychological costs of social support imbalance: Variation across relationship context and age. Journal of Health Psychology. Advance online publication.

4. Tsenkova, V. K., Karlamangla, A. (2016). Depression amplifies the influence of central obesity on 10-year incidence of diabetes: Findings from MIDUS. PLoS ONE, 11(10), e0164802.
5. Tomitaka, S., Kawasaki, Y., Ide, K., Akutagawa, M., Yamada, H., Ono, Y., Furukawa, T. A. (2017). Characteristic distribution of the total and individual item scores on the Kessler Screening Scale for Psychological Distress (K6) in US adults. BMC Psychiatry, 17, 290.

Schedule

Week 1 - Jan 28

In-class:

- Introduction to course
- Module 1: Introduction to Reproducible Science
  - 1.1 Welcome to Reproducible Science (5:14)
  - 1.2.1 Intro to the People/Faculty (2:39)
  - 1.3 Intro to the Modules (3:02)
  - Total Time: 10:55

At home:

- Create GitHub account
- Clone course repository to laptop
- Module 2: Fundamentals of Reproducible Science
  - 2.1 Why this matters (13:35)
  - 2.2 Definitions and concepts (5:58)
  - 2.3 Factors affecting reproducibility (3:30)
  - 2.4.1 Experimental design (4:25)
  - 2.4.2 Organization (3:27)
  - 2.4.3 Understanding yourself later (1:52)
  - 2.4.4 Ensuring others can continue your work (5:01)
  - 2.4.5 Providing workflows and results (2:52)
  - Total Time: 40:40

Week 2 - Feb 4

In-class:

- Brief discussion of Module 2 videos
- Introduction to markdown
- Introduction to RMarkdown
- Introduction to Jupyter notebook
- Introduction to git and GitHub
- Introduction to LaTeX and Overleaf

At home:

- Module 3: Case Studies of Reproducible Science
  - 3.1 Introduction to case studies (2:51)
  - 3.2 Potti 2006 (7:40)
  - 3.3 Baggerly and Coombes (17:22)
  - 3.4 Ioannidis 2009 (10:56)
  - 3.5 Case studies wrap up (5:18)
  - 3.6.1 Introduction to reproducible reporting science (1:51)
  - 3.6.2 Journals and reproducible research (9:12)
  - 3.6.3 NIH guidance on reproducible research (6:19)
  - Total Time: 61:29

Week 3 - Feb 11

In-class:

- Brief discussion of Module 3 videos
- Introduction to MIDUS II data sets
- Brief discussion of [1]
- Downloading data for [1]
- Reproduce Figure 1 for [1]

At home:

- Module 4: Data Provenance
  - 4.1 Introduction to data provenance (2:40)
  - 4.2.1 Data provenance concepts (4:37)
  - 4.2.2 Experimental design before the computer (4:53)
  - 4.2.3 Experimental design before on computer (6:39)
  - 4.2.4 Tools and standards (3:09)
  - 4.2.5 Workflows (5:19)
  - 4.3.1 Journals and reporting (4:17)
  - 4.3.2 Mechanisms for reporting (5:26)
  - 4.4.1 Data-type-specific repositories (5:05)
  - 4.4.2 General repositories (5:13)
  - 4.4.3 Code (2.14)
  - 4.4.4 Documentation (3:25)
  - 4.5.1 Data privacy and security (6:56)
  - 4.5.2 Privacy (5:50)
  - 4.5.3 Security (6:01)
  - Total Time: 75:44

Week 4 - Feb 18

In-class:

- Presidents' Day - No Class

At home:

- Start working on group project

Week 5 - Feb 25

In-class:

- Brief discussion of Module 4 videos
- Reproduce Table 1 of [1]

At home:

- Module 5: Statistical Methods for Reproducible Science
  - 5.1.1 Introduction to cross study validation of prediction models (10:58)
  - 5.1.1 Statistical methods introduction (1:19)
  - 5.2.1 Coefficient of determination (8:12)
  - 5.2.2 Brier Score (7:25)
  - 5.2.3 Area under the curve (AUC) (9:19)
  - 5.2.4 Concordance in survival analysis (16:18)
  - Total Time: 53:31

Week 6 - Mar 4

In-class:

- Brief discussion of Module 5 videos part 1
- Reproduce Table 2 of [1]

At home:

- Module 5: Statistical Methods for Reproducible Science
  - 5.2.5 Motivation for cross validation (10:53)
  - 5.2.6 Cross validation (6:33)
  - 5.2.7 Bootstrap (6:56)
  - 5.3.1 Simulations (11:05)
  - 5.3.2 Clustering (9:23)
  - Total Time: 44:50

Week 7 - Mar 11

In-class:

- Brief discussion of Module 5 videos part 2
- Reproduce Table 3 of [1]

At home:

- Module 6: Computational Tools for Reproducible Science
  - 6.1.1 Computational tools introduction (2:50)
  - 6.1.2 Contributor introductions (2:20)
  - 6.2.1 Introduction to editors (2:42)
  - 6.2.2 Introduction to R and Rstudio (2:58)
  - 6.2.3 Introduction to Python (0:46)
  - 6.2.4 Introduction to Git and GitHub (1:08)
  - 6.2.5 Downloading and installing Git (2:54)
  - 6.2.6 How to create a repository (2:36)
  - 6.2.7 GitHub interface (2:01)
  - 6.2.8 A conversation with Eric Surface (11:46)
  - 6.3.1 A conversation with Merce Crosas (3:09)
  - 6.3.2 A conversation with Merce Crosas 2 (5:42)
  - 6.3.3 A conversation with Merce Crosas 3 (2:33)
  - 6.3.4 A conversation with Merce Crosas 4 (0:47)
  - 6.3.5 A conversation with Merce Crosas 5 (6:23)
  - 6.3.6 A conversation with Merce Crosas 6 (3:20)
  - 6.3.7 A conversation with Merce Crosas 7 (5:19)
  - 6.3.8 A conversation with Merce Crosas 8 (3:48)
  - 6.3.9 A conversation with Merce Crosas 9 (2:07)
  - 6.3.10 A conversation with Merce Crosas 10 (4:36)
  - Total Time: 69:45
- Continue working on group project

Week 8 - Mar 18

In-class:

- Spring Break - No Class

At home:

- Continue working on group project

Week 9 - Mar 25

In-class:

- Brief discussion of Module 6 videos part 1
- Reproduce Table 4 of [1]

At home:

- Module 6: Computational Tools for Reproducible Science
  - 6.3.11 A conversation with Sonia Barbosa (12:28)
  - 6.3.12 Dataverse: Sonia Barbosa P1 (18:58)
  - 6.3.13 Dataverse: Sonia Barbosa P2 (14:06)
  - 6.3.14 API access (3:03)
  - 6.4.1 Introduction to dynamic report generation (1:07)
  - 6.4.2 A conversation with Christopher Gandrud (8:31)
  - 6.4.3 knitr and Rmarkdown with Christopher Gandrud (17:37)
  - 6.4.4 Notebooks: R notebook (3:13)
  - 6.4.5 Notebooks: Compiling in an R File (0:30)
  - 6.4.6 Notebooks: Jupyter (2:34)
  - 6.5.1 Makefiles (7:09)
  - Total Time: 89:16
- Continue working on group project

Week 10 - Apr 1

In-class:

- Brief discussion of Module 6 videos part 2
- Reproduce Table 5 of [1]

At home:

- Module 6: Computational Tools for Reproducible Science
  - 6.5.2 A conversation with Simon Adar 1 (4:54)
  - 6.5.3 A conversation with Simon Adar 2 (1:47)
  - 6.5.4 A conversation with Simon Adar 3 (7:09)
  - 6.5.5 A conversation with Simon Adar 4 (3:41)
  - 6.5.6 A conversation with Simon Adar 5 (4:06)
  - 6.5.7 A conversation with Simon Adar 6 (5:58)
  - 6.5.8 A conversation with Simon Adar 7 (0:49)
  - 6.5.9 A conversation with Simon Adar 8 (1:05)
  - 6.5.10 A conversation with Simon Adar 9 (4:15)
  - 6.5.11 A conversation with Simon Adar 10 (10:43)
  - 6.5.12 Code Ocean conclusion (0:48)
  - 6.5.13 Creating a new algorithm in Code Ocean (3:40)
  - 6.6 Conclusion (1:30)
  - 7.1 Course conclusion (2:08)
  - Total time: 52:33
- Continue working on group project

Week 11 - Apr 8

In-class:

- Brief discussion of Module 6 videos part 3
- Presentation of 2 case studies
- Formal critique of [1]

At home:

- Continue working on group project

Week 12 - Apr 15

In-class:

- Presentation of 2 case studies
- Automizing the reproducibility process for [1]
- Putting it all together: organizing a repository for reproducibility attempt of [1]

At home:

- Continue working on group project

Week 13 - Apr 22

In-class:

- Presentation of 2 case studies
- Unit tests, positive and negative controls

At home:

- Continue working on group project

Week 14 - Apr 29

In-class:

- Presentation of 2 case studies
- Privacy concerns discussion
- Harvard's data security levels and information security policy

At home:

- Continue working on group project

Week 15 - May 6

In-class:

- Presentation of 2 case studies
- Course summary
- Discussion of dos and don'ts for reproducible research and lessons learned

At home:

- Continue working on group project

Week 16 - May 13

In-class:

- Group project presentations

At home:

- Complete course evaluation
