2017-01-23

Rick O. Gilmore

2017-01-24 07:37:42

One step forward

- Capturing workflows and improving **methods** reproducibility (Goodman, Fanelli, and Ioannidis 2016)

Workflows

<"https://www.google.com/search?q=workflow">

Typical workflows in experimental psychology

- Idea/question/hypotheses
- Design study
- Seek ethics board permission
- Build/borrow/buy data collection instruments
- Run study
- Analyze results
- Write-up results for presentation and/or publication

Design study

- Participants
  - Number, characteristics
- Setting(s)
  - Lab, classroom, home
- Measures or tasks
  - Self/other report
  - Observations/video or audio recording
  - Physiological measures (MRI, EEG, ECG)
  - Computer-based tasks

Data collection instruments

- Surveys
- Video/audio
- MRI, EEG, ECG
- Computer-generated data files

Run study

```
done.collecting.data = FALSE
while (!done.collecting.data) {
  Collect.sample()
  if (collected.sample.n >= planned.sample.n) {
    done.collecting.data = TRUE
  } else {
    done.collecting.data = FALSE
  }
}
```

Analyze results

- Clean/check data
- Merge, combine, munge data

Prepare presentation/publication

- Intro
- Methods
- Results
  - Stats
  - Plots
- Conclusions
- References
- Data?

Behavioral study summary

Imaging example

Threats to methods reproducibility

- Idea/question/hypotheses
- Design study
- Seek ethics board permission
- Build/borrow/buy **data collection instruments**
- **Run study**
- **Analyze results**
- **Write-up results** for presentation and/or publication

What are the threats?

- Data collection instruments
- Running study
- Data analysis
- Study write-up

Mitigating the threats

- Maximize consistency, **methods** reproducibility (Goodman, Fanelli, and Ioannidis 2016)
- Design of/consistent adherence to detailed experimental protocols
- Consistent, transparent workflows
- Consistent, transparent organization of data, metadata
- Minimize human/hand data entry
- Automate as much as possible

How detailed is your (internal) protocol?

- Play & Learning Across a Year (PLAY) project wiki

Questions to consider

- What data and metadata am I collecting?
- How does it get collected?
- Where does it go after it's collected?
- How does my non-electronic data get transferred to an electronic form?
- How do my electronic data files get cleaned, merged, munged?

Reproducible workflow aspirations

- "Chain of custody" from raw data to finished results and figures
- Single command to regenerate all results and figures from raw data

http://datasci.kitzes.com/lessons/python/reproducible_workflow.html

Reproducible workflow recommendations

- Create consistent structure for projects
  - Use file name conventions
- Use machine-readable file types
  - commas-separated value (.csv) vs. .xlsx
- Automate as much as possible
- Use version control

Lots of ways to organize electronic data…

```
study-1/
  sub-001/
    sub-001-measure-a.txt
    sub-001-image.jpg
    sub-001-demo.csv
    sub-001-measure-b.txt
  sub-002/
    sub-002-measure-a.txt
    sub-002-image.jpg
    sub-002-demo.csv
    sub-002-measure-b.txt
  ...
  sub-00n/
    ...
```

```
study-1/
  measure-a/
    sub-001-measure-a.txt
    ...
  measure-b/
    sub-001-measure-b.txt
    ...
  image/
    sub-001-image.jpg
    sub-002-image.jpg
    ...
  demo/
    sub-001-demo.csv
    sub-002-demo.csv
    ...
```

```
study-1/
  analysis/
    data/
      sessions/
        2017-01-09-sub-001/
        ...
      aggregate/
        study-1-demo-aggregate.csv
        study-1-measure-a-aggregate.csv
        ...
    R/
    img/
    reports/
  protocol/
    code/
      my-experiment.m
    materials/
      stim-1.jpg
      stim-2.jpg
      ...
```

```
  pubs/
    presentations/
    papers/
  refs/
  grants/
    2016/
    2017/
  irb/
  mtgs/
```

Databrary's volume, session/materials model

Your browser does not support iframes.

https://nyu.databrary.org/volume/2

ProjectTemplate

- Automates some of the project management involved in data analysis
  - Hat Tip (HT): Michael Hallquist
- Gilmore says: Use what you like

Can automate project creation, too

```
## Create project directory
proj.name = "tmp_proj"
if (!exists(proj.name)) {
  dir.create(path = proj.name, recursive = TRUE)
}

# Create sessions directory
sessions.dir = paste(proj.name, "analysis/sessions", sep="/")
if (!exists(sessions.dir)) {
  dir.create(path = sessions.dir, recursive = TRUE) # creates intermediate dirs
}

# Aggregate data file directory
agg.dir = paste(proj.name, "analysis/aggregate", sep="/")
if (!exists(agg.dir)) {
  dir.create(path = agg.dir, recursive = TRUE)
}
```

Words to the wise

- Use consistent file/directory names
  - `lowerCamelCaseIsGood.txt` so is `UpperCamelCase.txt`
  - `underscores_between_words.txt` works; so do `dashes-between`.txt
  - avoid `spaces in your file names.txt`; these are not always reliably readable by all computers.
- Choose good, descriptive names

Consider seriously Karl Broman's guides

- Be consistent
- Write dates as YYYY-MM-DD.
- Fill-in all cells
- One thing in a cell
- Make your data a rectangle
- Create a data dictionary

Consider seriously Karl Broman's guides

- No calculations in raw data files
- No font or color to highlight data
- Make back-ups
- Validate data to avoid data entry errors
- Save data in plain text files (comma or tab-delimited)

Why?

- Data scientists (that's you!) spend a lot of time just cleaning data
- http://www.infoworld.com/article/3047584/big-data/hottest-job-data-scientists-say-theyre-still-mostly-digital-janitors.html

Easy to merge data sets if they contain a linking variable (like subID)

- study-1-demo-agg.csv contains
  - subID, sex, ageYrs, favColor
- study-1-rt-agg.csv contains
  - subID, condition, rt

```
subID,sex,ageYrs,favColor
001,m,53,green
002,f,51,blue
003,f,23,red
004,m,25,aqua
```

Don't put spaces between variables in comma-separated value (.csv) files. Also, make sure to add a final line feed/enter character.

```
subID,condition,rt
001,upright,250
001,inverted,300
002,upright,225
002,inverted,290
003,upright,270
003,inverted,230
004,upright,210
004,inverted,240
```

```r
# read data files, first row (header) contains variable names
demo <- read.csv(file = "csv/study-1-demo-agg.csv", header = TRUE)
rt <- read.csv(file = "csv/study-1-rt-agg.csv", header = TRUE)

# merge and print
merged <- merge(demo, rt, by = "subID")
merged
```

```
##   subID sex ageYrs favColor condition  rt
## 1     1   m     53    green   upright 250
## 2     1   m     53    green  inverted 300
## 3     2   f     51     blue   upright 225
## 4     2   f     51     blue  inverted 290
## 5     3   f     23      red   upright 270
## 6     3   f     23      red  inverted 230
## 7     4   m     25     aqua   upright 210
## 8     4   m     25     aqua  inverted 240
```

A final word about tidy data (Wickham 2014)

- Variables in columns
- Observations in rows
- ~~Ok~~/better to repeat values in columns
  - subID,trial,rt
  - 001,1,300
  - 001,2,345
  - 002,2,327
  - 003,3,429

Main points

- Think like a computer!
- Plan your work; work your plan
- Consistency, standard formats
- Tidy data
- Haven't said anything about "openness"… yet

More resources

- Data Carpentry, http://www.datacarpentry.org/
- Software Carpentry, https://software-carpentry.org/
- Open Science Framework (OSF), http://osf.io
- R for Data Science, http://r4ds.had.co.nz/

Project/write-up (due start of next class)

- Pick one of the recommended elements from Table 1 in Munafò, et al. (2017). Evaluate the recommendation. Do you agree that it would mitigate one or more threats to reproducibility. Why or why not? Do you agree with the assessment about the degree to which stakeholders have adopted the recommended practice?
- Edit/create a text-based workflow for an active project you are working on.
  - Annote the workflow to indicate where it could be made more reproducible, transparent.

References

Goodman, Steven N., Daniele Fanelli, and John P. A. Ioannidis. 2016. "What Does Research Reproducibility Mean?" *Science Translational Medicine* 8 (341): 341ps12–341ps12. doi:10.1126/scitranslmed.aaf5027.

Wickham, Hadley. 2014. "Tidy Data." *Journal of Statistical Software* 59 (10). doi:10.18637/jss.v059.i10.