Datapalooza Takeaways

Session 1

Jessica Schad:

1) Plan how you will use your survey data before you collect it.
2) To create a good survey, put yourself in the shoes of potential participants.
3) Know the limitations of your survey design.

Tyson Barrett:

1) There are numerous ways to mess up spreadsheet data and reduce reproducibility. Some general principles for working with data in spreadsheets can drastically reduce errors.
2) If the data will eventually be analyzed in another program (e.g. SPSS, R, SAS), there are important principles regarding data structure that can save time and avoid headaches.
3) A grammar of data is important to be able to communicate with other researchers, statisticians, and data scientists. A foundation developed in data science is presented to help individuals learn to speak the grammar of data.

Session 2

Susan Durham:

1) Plan ahead
2) Do not underestimate the knowledge, time, and diligence required in data analysis
    a. Educate yourself, with both formal classes and focused self-study
    b. Early on, add the methodology you anticipate needing to your toolkit
    c. Coding in a scripting language
        i. Data manipulation tools
        ii. Data visualization tools
        iii. Data analyses
3) Draw upon the expertise of others
    a. Your advisor, committee, and faculty in your field
    b. Statisticians and quantitative colleagues
    c. Merrill-Cazier Library Research Data Management
    d. Writing Center

Dave Bolton:

1) Plan ahead (e.g. statistical analyses, considering potential problems and alternative strategies, etc.)
2) Register your study in advance if possible
3) Inclusion of a positive control to provide a basis for comparison with your primary question

Richard Cutler:

1) Don't try to do everything—you just can't.
2) Try to learn about the methods you will use to analyze your data well before you are at the point for analyzing data

    a. By taking USU courses

   b. Finding resources in the library and online
   c. By talking with people who have expertise in the relevant areas
3) Think very hard about the details of your data analysis and avoid generalities

David Rosenburg:

1) **Reproducibility is a continuum.** Make materials available. Reproduce results. Use other data sets in other locations to replicate findings. **Move your work up the continuum!**
2) **You are not alone.** Today, most research data/models/code/directions are not available or reproducible. We want to change this!
3) **Authors, journals, funders, institutions must create a culture where we reproduce results.** Overcome challenges. Provide incentives. Learn new tools.

Additional information

Susan Durham

Ecology Center

Top Three Takeaway Points

1. Plan ahead

2. Do not underestimate the knowledge, time, and diligence required in data analysis

- Educate yourself, with both formal classes and focused self-study
- Early on, add the methodology you anticipate needing to your toolkit
  - Coding in a scripting language
  - Data manipulation tools
  - Data visualization tools
  - Data analyses

3. Draw upon the expertise of others

- Your advisor, committee, and faculty in your field
- Statisticians and quantitative colleagues
- Merrill-Cazier Library Research Data Management
- Writing Center

Suggested readings (a very few among many!)

Cahill JF Jr, Lyons D, Karst J (2011) Finding the "pitch" in ecological writing. Bulletin of the Ecological Society of America 92(2): 196-205. https://doi.org/10.1890/0012-9623-92.2.196 (writing)

Goodman A, Pepe A, Blocker AW, Borgman CL, Cranmer K, et al. (2014) Ten simple rules for the care

and feeding of scientific data. PLoS Computational Biology 10(4): e1003542. https://doi.org/10.1371/journal.pcbi.1003542 (reproducibility)

Hurlbert SH (1984) Pseudoreplication and the design of ecological field experiments. Ecological Monographs 54(2): 187-211. https://doi.org/10.2307/1942661 (experimental design concepts)

Johnson DH (2006) The many faces of replication. Crop Science 46: 2486-2491. https://doi.org/10.2135/cropsci2006.04.0277 (experimental design concepts)

Kass RE, Caffo BS, Davidian M, Meng X-L, Yu B, Reid N (2016) Ten simple rules for effective statistical practice. PLoS Computational Biology 12(6): e1004961. https://doi.org/10.1371/journal.pcbi.1004961 (data analysis)

Lowndes J, Best B, Scarborough C, Afflerbach JC, Frazier MR, O'Hara CC, Jiang N, Halpern BS (2017) Our path to better science in less time using open data science tools. Nature Ecology & Evolution 1: 0160. https://doi.org/10.1038/s41559-017-0160 The corresponding online text: Introduction to Open Data Science https://ohi-science.org/data-science-training/ (reproducibility)

Mead R, Gilmour SG, Mead A (2012) Statistical Principles for the Design of Experiments: Applications to Real Experiments. Cambridge University Press. (experimental design concepts)

Merkle BG (2019) Writing science: leveraging a few techniques from creative writing toward writing more effectively. Bulletin of the Ecological Society of America e01650. https://doi.org/10.1002/bes2.1650 (writing)

Steel EA, Kennedy MC, Cunningham PG, Stanovick JS (2013) Applied statistics in ecology: common pitfalls and simple solutions. Ecosphere 4(9): 115. http://dx.doi.org/10.1890/ES13-00160.1 (data analysis)

Zuur AF, Ieno EN, Elphick CS (2010) A protocol for data exploration to avoid common statistical problems. Methods in Ecology and Evolution 1(1): 3-14.

https://doi.org/10.1111/j.2041-210X.2009.00001.x (data analysis)

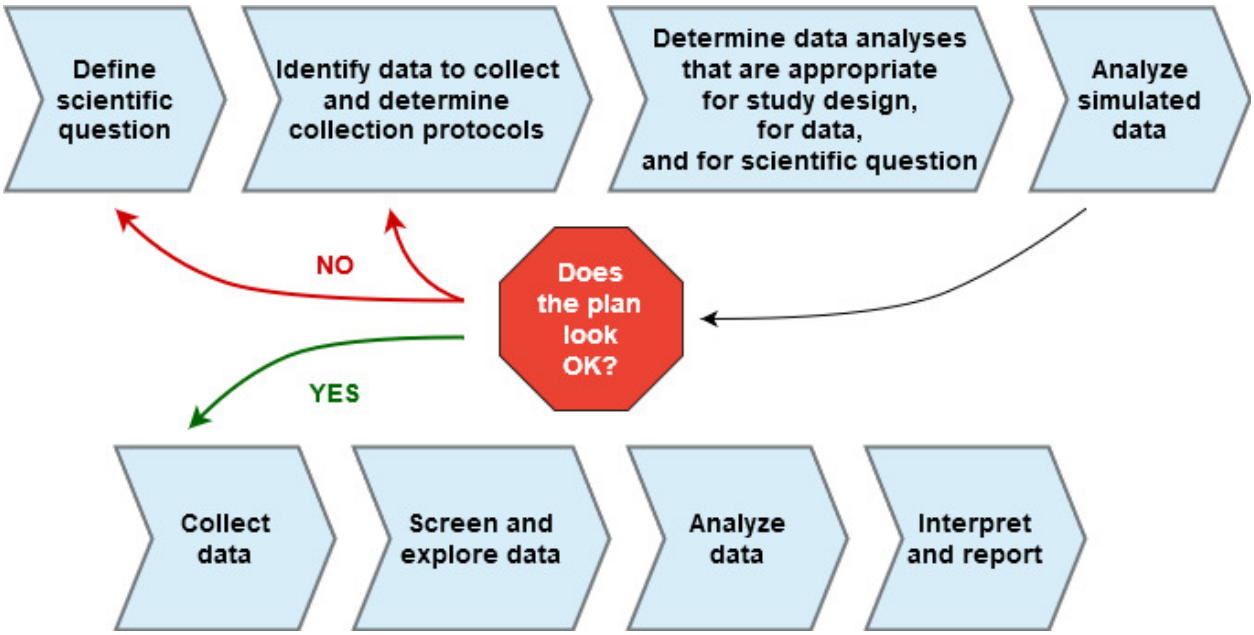

Figure 1. One vision of a flowchart for a scientific study.

Building reproducibility from the start

David E. Rosenberg
Civil & Environmental Engineering and Utah Water Research Lab

Three Top Takeaway Points

1. **Reproducibility is a continuum.** Make materials available. Reproduce results. Use other data sets in other locations to replicate findings. **Move your work up the continuum!**

2. **You are not alone.** Today, most research data/models/code/directions are not available or reproducible. We want to change this!

3. **Authors, journals, funders, institutions must create a culture where we reproduce results.** Overcome challenges. Provide incentives. Learn new tools.

Additional Points

- Build reproducibility into the project from start – budget time, money, storage, IRB, and tools
- Put all materials in a repository (Github, Hydroshare, Digital Commons, Figshare, etc…)
- Make all inputs to and outputs of proprietary, private, & computationally intensive steps available
- Ask someone to verify that your results are reproducible
- Train students and employees in reproducible practices (this Datapalooza :)

Resources

- Rosenberg et. al. (in press). "The Next Frontier: Making Research More Reproducible." *Journal of Water Resources Planning and Management*. https://digitalcommons.usu.edu/water_pubs/156/.
- Stagge et. al. (2019). "Assessing data availability and research reproducibility in hydrology and water resources." Nature-*Scientific Data*, 6, 190030. https://doi.org/10.1038/sdata.2019.30.

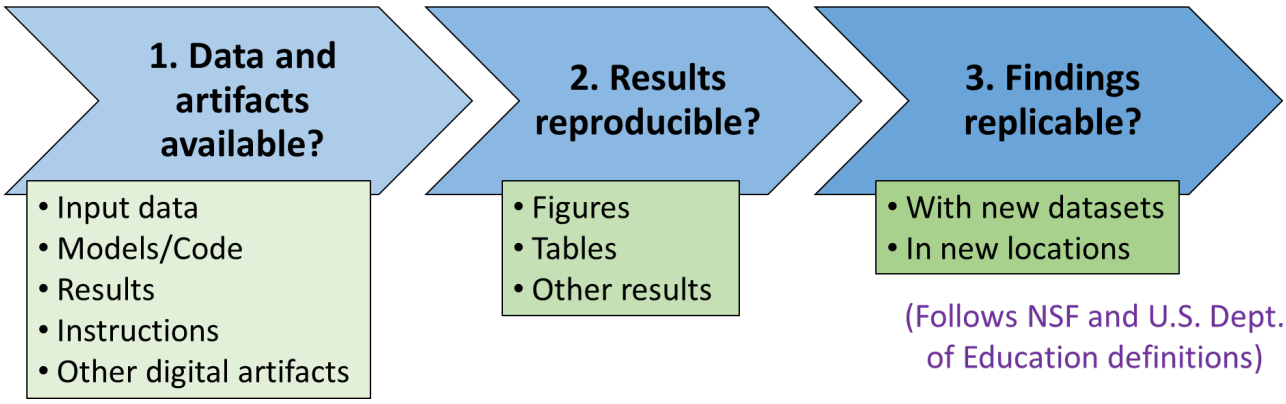

Figure 1. Reproducibility is a continuum. Push your work up the continuum.

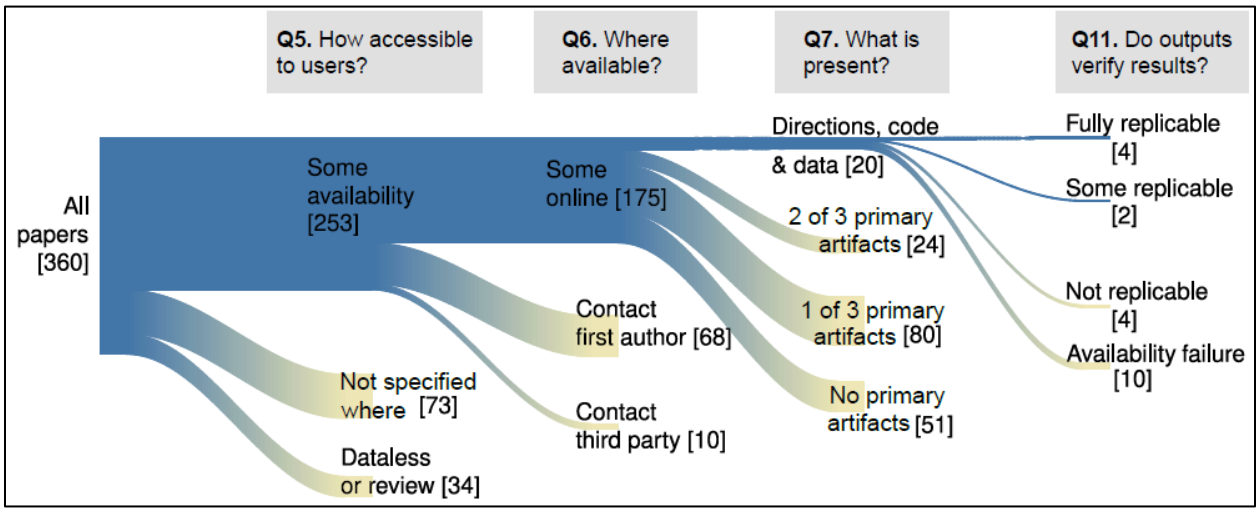

Figure 2. Availability and reproducibility of 360 hydrology and water resources papers published in 2017. Line width represents the number of papers.