RUTGERS
University Libraries

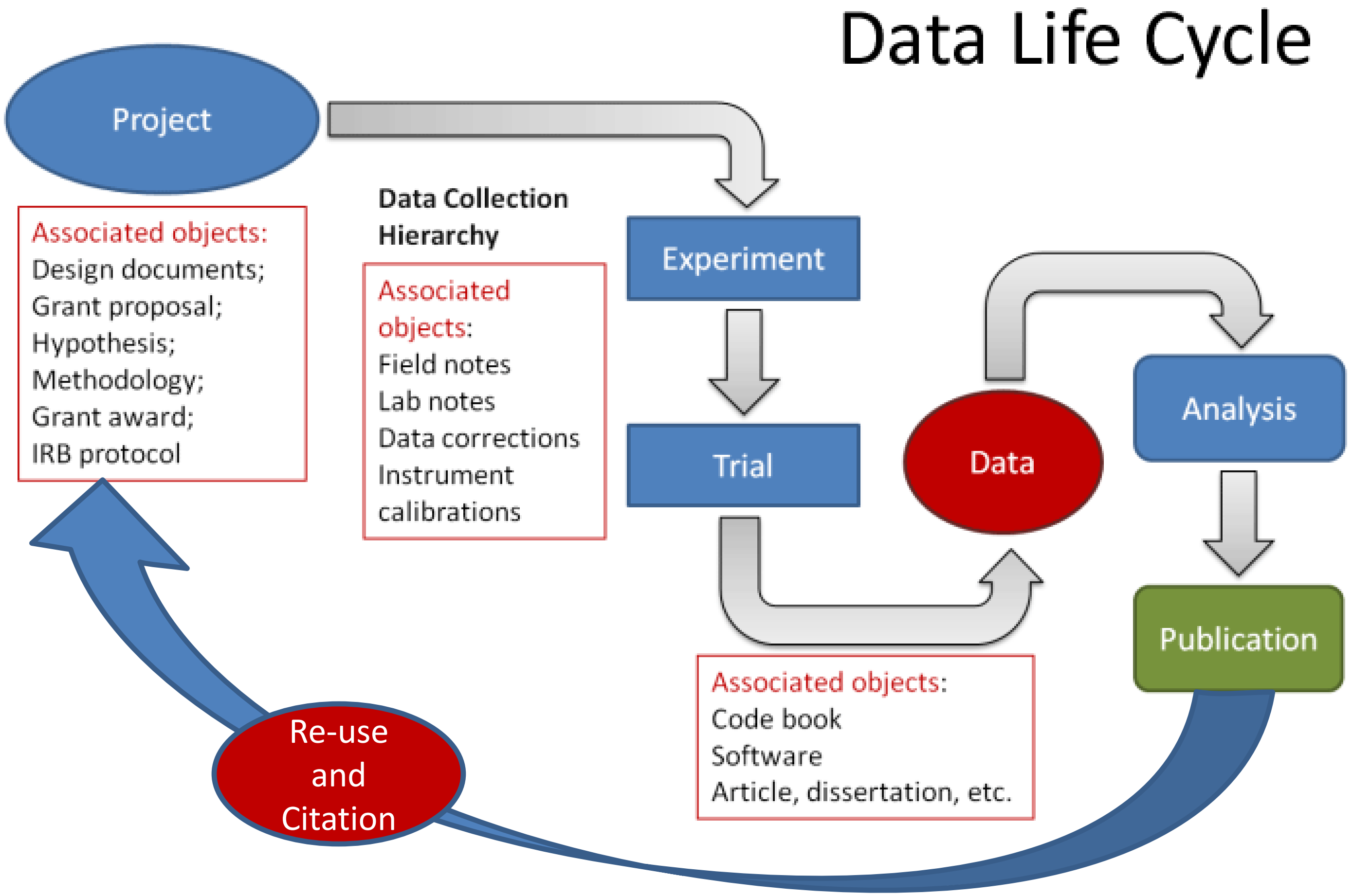
The place to go, when you need to know!

Reproducible Research

Ryan Womack
Rutgers University Libraries
Research Data Services

October 2015http://guid

The Data Lifecycle



Data Life Cycle

Project

Associated objects:
Design documents;
Grant proposal;
Hypothesis;
Methodology;
Grant award;
IRB protocol

Data Collection
Hierarchy

Associated objects:
Field notes
Lab notes
Data corrections
Instrument calibrations

Experiment

Trial

Data

Analysis

Publication

Associated objects:
Code book
Software
Article, dissertation, etc.

Re-use and Citation

Why Reproducible Research?

- In our earlier session, we discussed data that is…
  - Robust
  - Recoverable
  - Reliable
  - **Reusable**
  - **Reproducible**
  - Reputable
  - Renowned

Why Reproducible Research?

- Credibility in Science
  - Scientific fraud is on the rise
    - Duke “starter set”
  - Research misconduct is a problem, but so is human error
- Reputation, Prestige, and Funding are all affected
  - "Set the default to Open"
- Replication (redoing the experiment from scratch) is expensive, and may not be possible due to the passage of time. (see Validation)
- ***Science*** on Replication and Reproducibility
- Victoria Stodden, Friedrich Leisch, and Roger D. Peng (eds.). *Implementing Reproducible Research*. CRC Press, 2014.

What is Reproducible Research?

- "In all research that utilizes a computer, instructions for the research are stored in software and scientific data are stored digitally. A typical publication in computational research is based foundationally on data, and the computer instructions applied to the data that generated the scientific findings. The complexity of the data generation mechanism and the computational instruction is typically very large, too large to capture in a traditional scientific publication. Hence when computers are involved in the research process, scientific publication must shift from a scientific article to the triple of scientific paper, and the software and data from which the findings were generated. This triple has been referred to as a "research compendia" and its aim is to transmit research findings that others in the field will be able to reproduce by running the software on the data. Hence, data and software that permits others to reproducible the findings must be made available."
  - Victoria Stodden - http://blog.stodden.net/2014/09/28/my-input-for-the-ostp-rfi-on-reproducibility/

What is Literate Programming?

Introduced by Donald Knuth, refers to code that is explained in context.

- Can be accomplished in any environment
- But some are easier than others
- We will look at **knitr** (Yihui Xie) in R
- **Sweave** (Friedrich Leisch) handles LaTeX
- “weave” the literate program to create a human-readable document (text)
- “tangle” the literate program to create a machine-readable document (program)

Scientific Method

- Mathematical Proof
  - Explicitly repeatable steps
- Experimental Method
  - Documentation of materials, methods and protocols
- Computational Science
  - **Should** act like the above, but often fails to meet this ideal
  - Data sharing
  - Code sharing
  - Explicit standards are required to bring computation into a fully recognized scientific method framework

See http://stanford.edu/~vcs/talks/DSSMay42011-STODDEN.pdf for more discussion

Reproducible Research

- Ideally, someone else can grab your data project as a complete bundle of data, documentation, and software code, and recreate the analysis to get exactly the same results
- Reports and data can be integrated so that live analysis run on actual data can be placed in reports (e.g., via Sweave, knitr or other R packages).
- Many initiatives (e.g., RunMyCode) are advancing the concept of reproducible research.
- This high standard of evidence and validation is an assurance that data and conclusions are not flawed (or faked).
- Good data management practices lay the groundwork for success in reproducible research

Reproducible Research in Action

- ISPS (Yale Institute for Social and Policy Studies)
  - http://isps.yale.edu
  - Complete code, data, readme
  - Curated archive
  - Detailed descriptive metadata for each study
- Dataverse and others

What is required for Reproducible Research?

1st piece, your data

- Raw Data
- Working Data
- Processed or Final Data

- At least the last of these must be available. Ideally, you can provide code or completely document the process that took you from raw data to final data.
  - "script everything"
  - "don't do things by hand" (edit in Excel, download from a web site), but document carefully if you have to
- As discussed in previous sessions, data should be fully documented via a codebook or other means, ensuring that other researchers understand the data.

What else is required for Reproducible Research?

- Your code. This is the most critical step that allows researchers not just to approximate your analysis with their own methods, but to actually implement the exact methods you took.
- Analysis depends on software and computing environment.
  - Record this information! (session.Info() in R)
  - Easier to use older versions with open source software.
- Tools are being developed to help with this
  - cacher
  - checkpoint
- Don't save output. Where did it come from? This should be done in the code.
- Clean, well-formatted data (tidyr) and code (formatR)
- If using "readme" files approach, document everything

Anything else?

- Literate programming can solve documentation and reliability issues.
- An integrated environment containing both document and code can ensure that the data is used exactly as intended (Sweave, knitr). Code that is “tangled” in with text can be extracted, and formatted documents can be “woven” from the literate program.
- Share your data! (see workshop on Data Sharing)

Literate Programming in Action

- Rstudio has knitr built in
- Can publish documents directly to Rpubs
- Uses R Markdown .Rmd -> .md -> .html
- knit2html produces html directly, or use Rstudio
- Only a few options
  - echo=FALSE (or TRUE)
  - results="hide" (or "asis")
  - fig.height
  - cache=TRUE stores result of complex computation
- 'r command' produces inline results
- To change defaults
  - {r setoptions, echo=FALSE}
  - opts_chunk$set(echo=FALSE, results="hide")
- Note that Sweave uses a similar paradigm if you want to do LaTeX or Open Document Formats. "Stangle" extracts code. "Sweave" produces document.

Data Sharing in Action

- R dackages like dvn for Dataverse and rfigshare allow quick upload of data. More at ropensci.
- Data and code can be distributed through packages, start with the package.skeleton() command
- Package ProjectTemplate provides one approach to generating a complete, organizing framework for project workflow

Benefits to You

- Benefits to the academic community have been discussed
- Once you invest in setting up your workflow and framework, you save time when updating and editing your work
- You can recall and reuse pieces of earlier work more easily
- Your work becomes more reliable, faster, and can scale up

More …

- Roger Peng, Reproducible Research, Coursera course
- Victoria Stodden, Friedrich Leisch, and Roger D. Peng (eds.). *Implementing Reproducible Research*. CRC Press, 2014.
- Victoria Stodden – many publications