Introduction to Reproducible Research

Quantitative Fisheries Center, Michigan State University

December 11-12, 2013.

Outline

Motivation

- Practical
- Philosophical

Reproducible Research

- what is reproducible research
- how it is changing science - gold standard for open, transparent and credible science
- MORE ...

Typical Work Flow

- separate files for each stage of process:
  - data cleaning and preparation
  - model fitting and analysis
  - summarization and reporting
- lots of clicking and copy-paste
- often uses proprietary binary file format (.ppt, .doc, .xls) that is difficult to 'version'
- tedious, error prone, difficult or impossible to replicate exactly

- Figure

But what if . . .

- I make a small change to data?
- I have multiple reports?
- I want to implement complicated model improvements that might break everything?
- I need to find data and exact model or source code that created a report
- I need to work collaborative with an other analyst?
  - How do I know **exactly** what's changed?
  - How do I integrate their changes with my changes?

Reproducible Research

- provides an alternative method of work that avoids many of these problems
- Literate Programing
- Donald Knuth
- more recently...

An Alternative Work-flow

- all preparation, analysis, reporting done in simple text files
- results from analysis integrated directly into reporting products - reports, presentations, html documents
- reports automatically regenerated when data changes
- no undocumented figures, tables or results

An Alternative Work-flow (cont'd)

- no undocumented clicking or cutting and pasting
- version control software keeps track of changes
  - reset directory to past states
  - compare changes from one state to another
  - branching and merging allow 'safe' development

Further Reading and Resources

- Oxford Scientific Journals:

http://biostatistics.oxfordjournals.org/content/10/3/405.full

- R's Reproducible Research Task page:

http://cran.r-project.org/web/views/ReproducibleResearch.html