

# Introduction to Reproducible Research

Quantitative Fisheries Center, Michigan State University

December 11-12, 2013.

# Reproducible Research

- Philosophical

- ▶ reproducibility is one of core principles of the scientific method
- ▶ some consider reproducible research to be the gold standard in scientific credibility as all of the data, analyses and reports are encapsulated in a single entity
- ▶ some journals are moving towards reproducibility standards (e.g. see: <http://biostatistics.oxfordjournals.org/content/10/3/405.full>)

# Reproducible Research

*An article about computational science in a scientific publication is not the scholarship itself, it is merely advertising of the scholarship.*

*The actual scholarship is the complete software development environment and the complete set of instructions which generated the figures*

- Professor Jon Claerbout  
Stanford University

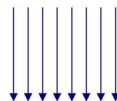
- ▶ source: <http://www.rrplanet.com/reproducible-research/reproducible-research.shtml>

# Reproducible Research

- From a more practical perspective:
  - ▶ reproducible research means being able automatically recreate your reports when data is changed or updated
  - ▶ or recreating your exact working directory to a state that generated a specific report, presentation or result

# 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



# 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 specific 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?

# An Alternative Work-flow

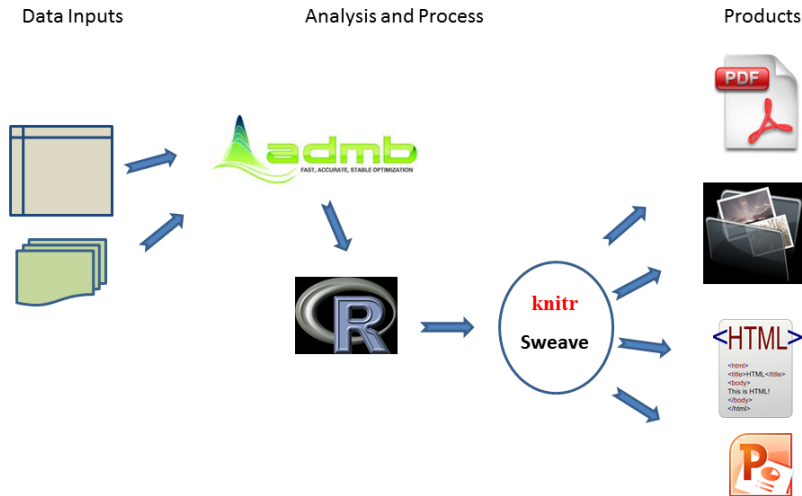
- all preparation, analysis, reporting done in simple text files
- results from analysis integrated directly into source code for reporting products
  - ▶ reports
  - ▶ presentations
  - ▶ html documents
- reports automatically regenerated when data changes
- no undocumented figures, tables or results

# An Alternative Work-flow (cont'd)

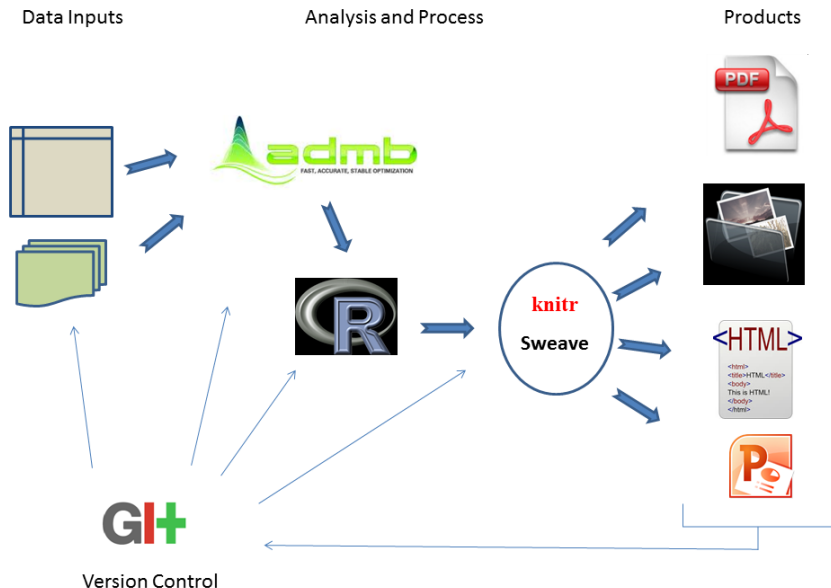
- no undocumented clicking or cutting and pasting
- version control software keeps track of changes
  - ▶ reset directory to past states
  - ▶ compare changes from one state to another
  - ▶ branching and merging allow 'safe' development



# An Alternative Work-flow (cont'd)



# An Alternative Work-flow (cont'd)



# 10 Simple Rules

From (Sandve et al. 2013): <sup>1</sup>

- For Every Result, Keep Track of How It Was Produced
- Avoid Manual Data Manipulation Steps
- Archive the Exact Versions of All External Programs Used
- Version Control All Custom Scripts
- Record All Intermediate Results, When Possible in Standardized Formats

---

<sup>1</sup>Sandve GK, Nekrutenko A, Taylor J, Hovig E (2013) Ten Simple Rules for Reproducible Computational Research. PLoS Comput Biol 9(10): e1003285. doi:10.1371/journal.pcbi.1003285

# 10 Simple Rules (cont'd):

From (Sandve et al. 2013): <sup>2</sup>

- For Analyses That Include Randomness, Note Underlying Random Seeds
- Always Store Raw Data behind Plots
- Generate Hierarchical Analysis Output, Allowing Layers of Increasing Detail to Be Inspected
- Connect Textual Statements to Underlying Results
- Provide Public Access to Scripts, Runs, and Results

---

<sup>2</sup>Sandve GK, Nekrutenko A, Taylor J, Hovig E (2013) Ten Simple Rules for Reproducible Computational Research. PLoS Comput Biol 9(10): e1003285. doi:10.1371/journal.pcbi.1003285