STAT259 Summary 1

Mitch Negus
9/7/2017

## Developing open source scientific practice

**Millman, K. and Perez F.**

–COMPUTATIONAL SCIENCE–

REPRODUCIBLE SCIENCE
- for good science, it must be reproducible
- open source software community has developed this framework/workflow
- though feasibility of reproduction may be limited, the possibility of reproduction must be demanded

LIFE CYCLE

*individual

*collaboration

*production

*publication

*education

-Most common tools create discontinuities across stages of workflow
-Results are considered separate from process, rather than a unified science product
-Joining tools and stages requires both technical and social changes

Reproducibility must be a commitment from the start

OPEN SOURCE
-moving science forward requires computational literacy–¿ science is open source –¿ contributing to science will require open source practices
challenges exist with making science fully open source (author recognition, first to publish, etc.)

–PRACTICE–

VERSION CONTROL
-files stored in repositories, require commits (w/ message)
-allow branching and merging
-modern systems allow data integrity verification (cryptographically fingerprinting)
-also limited when dealing w/ large binary files (solutions being developed in this regard)

AUTOMATED EXECUTION
-reproducibility should extend to process; best to automate all steps when possible -still should be able to be understood by people

-make files facilitate this process

TESTING -testing should accompany product development (test-driven-design) -allows focus on use (rather than details) -TDD prevents "getting lost" in tangled code web
READABILITY
-you and others will read your code (esp. to verify results)
-self-doc code reduces external documentation by being clear and forward
-use the right level of abstraction when writing mathematical expressions (don't simply too much, but don't avoid it entirely)
-comments may be uncoupled from code (one changes, the other is not updated)
-use of docstrings allows coupling of docs to code–¿ then autogeneration for web

INFRASTRUCTURE
-hosted version control allows group collaboration
-continuous integration to automatically execute test-suite

PULL REQUEST
-pull request akin to peer review
-anyone can chime in, lasting document of decisions
-private branches (maintain credit, history, and privacy while allowing transparency after integration

-linear algebra book by Rob Beezer (U. Puget Sound)