

Data Access: Digital Technology and Multiple Scientific Communities

Victoria Stodden
Department of Statistics
Columbia University

Changing the Conduct of Science in the Information Age
NSF Workshop
Nov 12, 2010

Defining Data

- numerical arrays
- experimental results, ie. video
- simulation data
- code
- scientific literature
- other modalities?

Framing Principle: Reproducibility

Data sharing isn't just sensible in the digital age, nor is it just another opportunity to improve science, it is *imperative*:

- computational, data-driven, science must be *reproducible* as the scientific method demands,
- computational methodology - code and data - must be shared with the publication of the results.
- Consequences for verifiability (ClimateGate, Duke Clinical Trials...) and public confidence in science.

Yale Roundtable 2009

- Roundtable on Data and Code Sharing in computational science
Nov 21, 2009:
 1. Framing the Problem
 2. Legal Barriers to Sharing
 3. Computational Solutions
 4. Norms and Incentives
 5. Draft Position Statement (published in IEEE Computing in Science and Engineering)
- <http://www.stanford.edu/~vcs/Conferences/>

Idea: Pilot Project Funding

- Fund a small number of projects from different areas to be fully reproducible,
- Permit the grantees to describe their plan (likr data release plan),
- Grantees also propose their additional needs.
- Creates an experiment to understand the requirements of reproducible research:
 - repository? extra coders? support for locally hosted code and data?
publication mechanisms for reproducible work?