The Coming Dark Ages: Preserving our Scientific Knowledge

Victoria Stodden
Department of Statistics
Columbia University

Big Data: Public Policy and the Exploding Digital Corpus
Princeton University
Nov 30, 2010

The Digitization of Science

- Computational methods emerging as central to the scientific enterprise:
  - in breadth: across a wide variety of fields,
  - in depth: changing how we understand our world.
- Intellectual contributions now in code, data.
- Impact on reproducibility of results and scientific integrity.

Reproducibility Reduced

Increased use of computation must be accompanied by data and code sharing such that published results are reproducible.

- Archiving, storage, versioning required,
- Intellectual property law works against reproducibility,
- Journals beginning to require data/code,
- Funding agencies require software, data to be shared openly..
- Consequences for verifiability (ClimateGate, Duke Clinical Trials...) and

Consequences?

- Another dark age...
  - Without reproducibility knowledge cannot be recreated/understood, and will be lost,
  - Science knowledeg accumulation is not about believing, but understanding.

Recommendations

- Assessment of expense of data/code archiving,
- Enforcement of funding agency guidelines,
- Publication requirements,
- Standards for scientific tools: collaboration focus, workflow tracking, code tests, ease of use for scientists,
- Versioning as a scientific principal,
- Licensing to realign scientific intellectual property with longstanding scientific norms (*Reproducible Research Standard*).