

Data Impediments to Scientific Discovery

Thursday, March 28, 2013

Bronis R. de Supinski

 Lawrence Livermore
National Laboratory

LLNL-PRES-630852

This work has been authored by Lawrence Livermore National Security, LLC under contract DE-AC52-07NA27344 with the U.S. Department of Energy. Accordingly, the United States Government retains and the publisher, by accepting this work for dissemination, acknowledges that the United States Government retains a non-exclusive, paid up, irrevocable, world-wide license to publish or reproduce the disseminated form of this work or allow others to do so, for United States Government purposes.



Is data ownership a problem?

- Hear a lot about big data
 - It's not lack of data, it's too much data
 - Are scientific domains different?
- Significant scientific data is widely available
 - Some realms require supporting data for publication of papers
 - Computer science is an exception
 - Computational science? Even in disciplines that require it?
 - Genetics and biological domains
 - Significant amount of genetic sequencing data
 - 1,230 digital collections identified in the 2010 annual *Nucleic Acids Research* database Issue
 - PDS: the Planetary Data System -- NASA planetary mission data
 - Some disciplines are motivated to have open repositories
 - RealClimate: Climate science from climate scientists
 - ESA: Ecological Society of America Data Registry

Many issues lead to researchers refusing to share data

- Personal ambition
 - “If I share my data, I might be scooped!”
 - Not just an issue in computation
 - Access to archeological artifacts
 - Recent astronomical data?
- Fragile egos
 - “If I let my bugs be known, people will know I code poorly!”
 - “If I make my performance data available, people will see that I often don’t worry about computational complexity!”
- Security concerns
 - Classification issues
 - Export control
 - Business concerns prior to patent applications
 - General distrust

