How Computational Science is Changing the Scientific Method

Victoria Stodden

Yale Law School and Science Commons

vcs@stanford.edu

Science 2.0

The University of Toronto

July 29, 2009

Agenda

1. The Scientific Method is being transformed by massive computation
   - New modes of knowledge discovery?
   - New standards for what we consider knowledge?
2. Why aren't researchers sharing?
3. Facilitating reproducibility 1: the *Reproducible Research Standard*
4. Facilitating reproducibility 2: tools for attribution and research transmission

Transformation of Scientific Enterprise

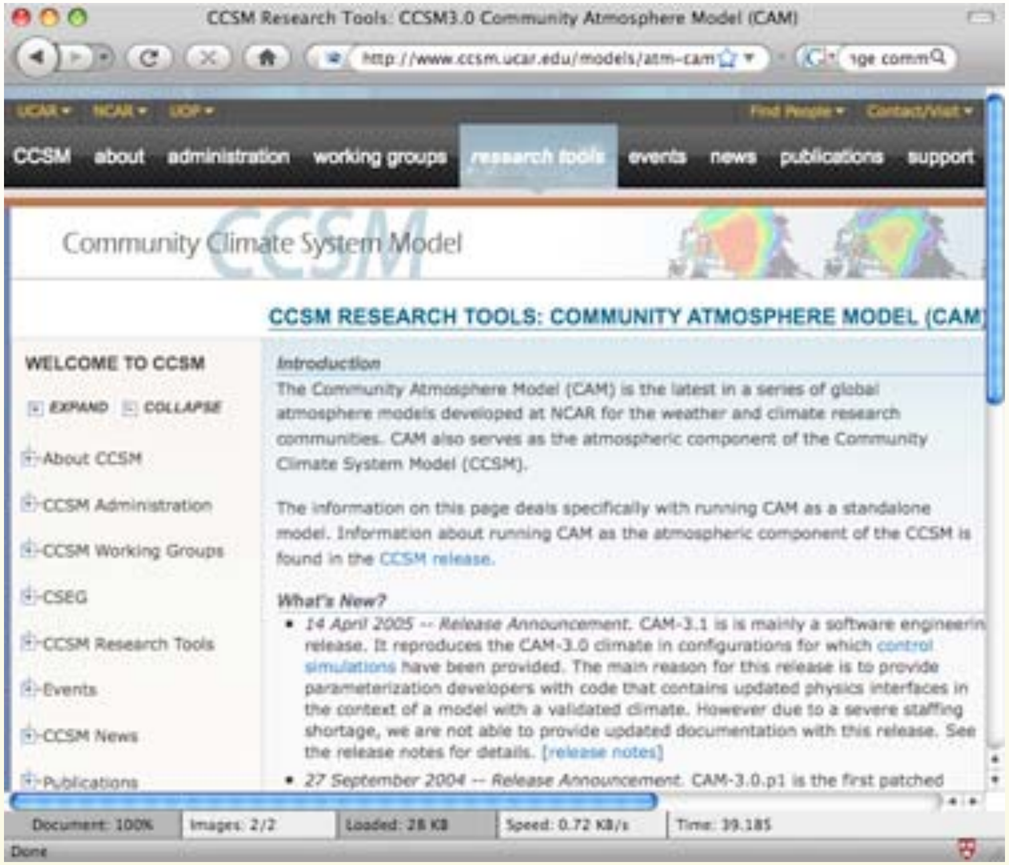
Massive Computation: emblems of our age include:

- data mining for subtle patterns in vast databases,
- massive simulations of a physical system’s complete evolution repeated numerous times, as simulation parameters vary systematically.

Raises new questions about science..

Example: Community Climate Model (CCM)

- Collaborative system simulation
- Open code, data



CCSM RESEARCH TOOLS: COMMUNITY ATMOSPHERE MODEL (CAM

Introduction

The Community Atmosphere Model (CAM) is the latest in a series of global atmosphere models developed at NCAR for the weather and climate research communities. CAM also serves as the atmospheric component of the Community Climate System Model (CCSM).

The information on this page deals specifically with running CAM as a standalone model. Information about running CAM as the atmospheric component of the CCSM is found in the CCSM release.

What's New?

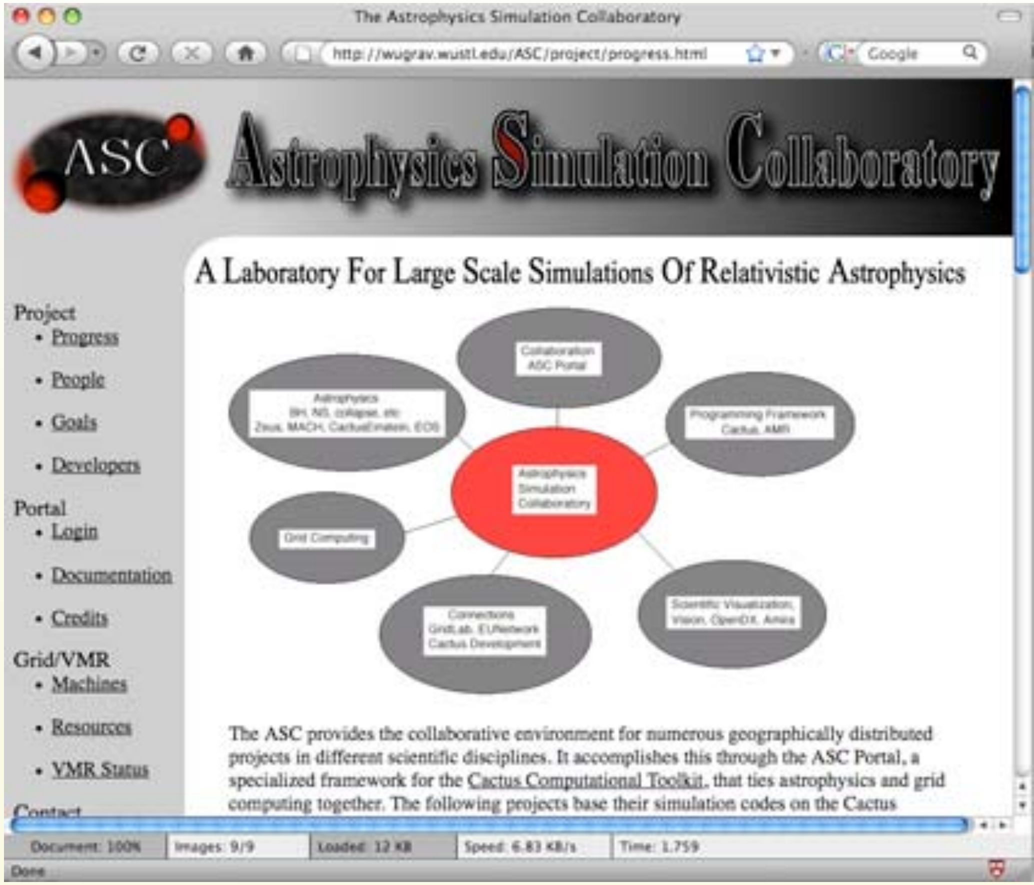
- *14 April 2005 -- Release Announcement.* CAM-3.1 is is mainly a software engineerin release. It reproduces the CAM-3.0 climate in configurations for which control simulations have been provided. The main reason for this release is to provide parameterization developers with code that contains updated physics interfaces in the context of a model with a validated climate. However due to a severe staffing shortage, we are not able to provide updated documentation with this release. See the release notes for details. [release notes]
- *27 September 2004 -- Release Announcement.* CAM-3.0.p1 is the first patched

Example: High Energy Physics

- 4 LHC experiments at CERN: 15 petabytes produced annually
- Data shared through grid to mobilize computing power
- Director of CERN (Heuer): “Ten or 20 years ago we might have been able to repeat an experiment.They were simpler, cheaper and on a smaller scale. Today that is not the case. So if we need to re-evaluate the data we collect to test a new theory, or adjust it to a new development, we are going to have to be able reuse it. That means we are going to need to save it as open data….” Computer Weekly, August 6, 2008

Example: Astrophysics Simulation Collaboratory

- Data and code sharing
- Interface for dyamic simulation
- mid 1930’s: calculate the motion of cosmic rays in Earth’s magnetic field..



The Astrophysics Simulation Collaboratory

http://wugrav.wustl.edu/ASC/project/progress.html

Project
- Progress
- People
- Goals
- Developers

Portal
- Login
- Documentation
- Credits

Grid/VMR
- Machines
- Resources
- VMR Status

Contact

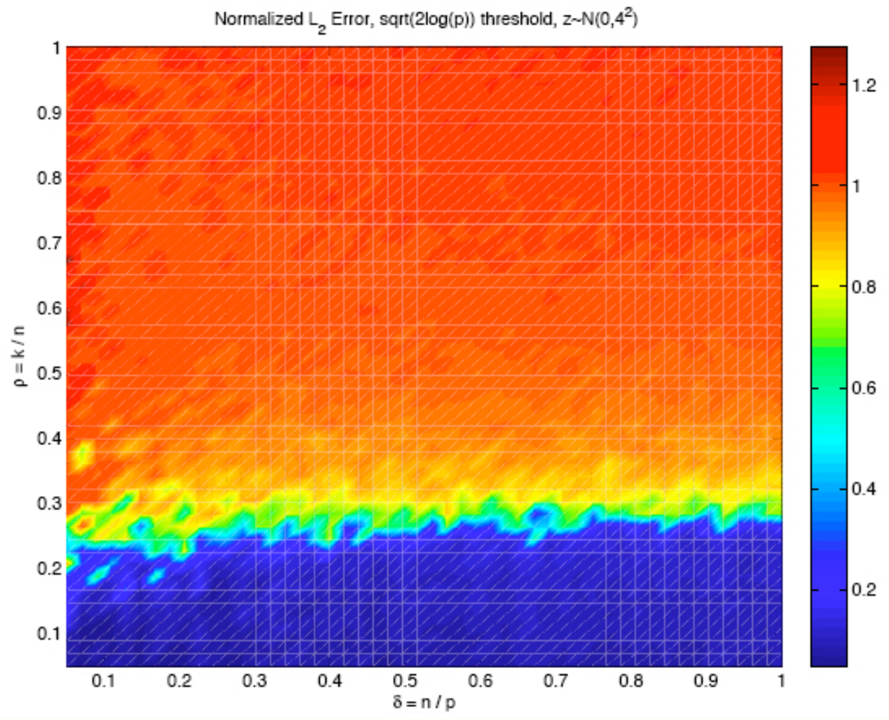
A Laboratory For Large Scale Simulations Of Relativistic Astrophysics

The ASC provides the collaborative environment for numerous geographically distributed projects in different scientific disciplines. It accomplishes this through the ASC Portal, a specialized framework for the Cactus Computational Toolkit, that ties astrophysics and grid computing together. The following projects base their simulation codes on the Cactus

Example: Proofs

- Mathematical proof via simulation, not deduction
- Breakdown point:

1/sqrt(2log(p))



- A valid proof?
- A contribution to the field of mathematics?

The Third Branch of the Scientific Method

- Branch 1: *Deductive/Theory*: e.g. mathematics; logic
- Branch 2: *Inductive/Empirical*: e.g. the machinery of hypothesis testing; statistical analysis of controlled experiments

- Branch 3: Large scale extrapolation and prediction: Knowledge from computation or tools for established branches?

Contention About 3rd Branch

- Anderson: The End of Theory. (Wired, June 2008)
- Hillis Rebuttal: We are looking for patterns first then create hypotheses as we always have.. (The Edge, June 2008)
- Idea (Weinstein): Simulation underlies branches:
  1. Tools to build intuition (branch 1)
  2. Tools to test hypotheses (branch 2)
- Manipulation of systems you can’t fit in a lab
- Not new: differential analyzers of 50’s and 60’s, chaos research in 70’s

Controlling Error is Central to Scientific Progress

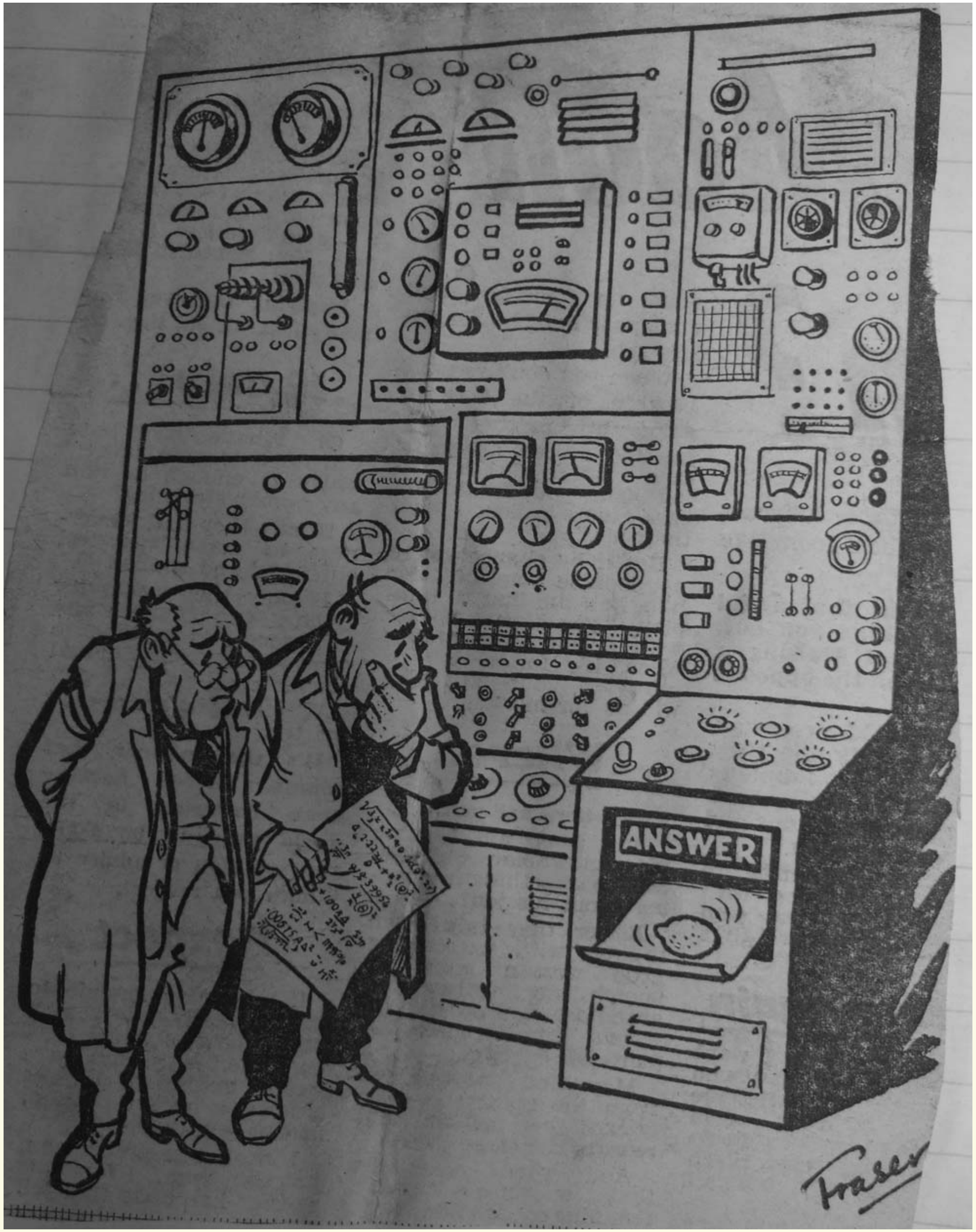
"The scientific method's central motivation is the *ubiquity of error* - the awareness that mistakes and self-delusion can creep in absolutely anywhere and that the scientist's effort is primarily expended in recognizing and rooting out error." David Donoho et al. (2009)

Computation is Increasingly Pervasive

- JASA June 1996: 9 of 20 articles computational
- JASA June 2006: 33 of 35 articles computational

Emerging Credibility Crisis in Computational Science

- Error control forgotten? Typical scientific communication doesn’t include code, data.
- Published computational science near impossible to replicate.
- JASA June 1996: none of the 9 made code or data available
- JASA June 2006: 3 of those 33 articles had code publicly available.
- A second change to the scientific method due to computation?


ANSWER
Fraser

Changes in Scientific Communication

- Internet: communication of all computational research details/data possible
- Scientists often post papers but not their complete body of research
- Changes coming: Madagascar, Sweave, individual efforts, journal requirements...

Potential Solution:
Really Reproducible Research

Pioneered by Jon Claerbout

“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.”

(quote from David Donoho, “Wavelab and Reproducible Research,” 1995)

Reproducibility

- (Simple) definition: A result is reproducible if a member of the field can independently verify the result.
- Typically this means providing the original code and data, but does not imply access to proprietary software such as Matlab, or specialized equipment or computing power.

Barriers to Sharing: Survey

Hypotheses:

1. Scientists are primarily motivated by personal gain or loss.

2. Scientists are primarily worried about being scooped.

Survey of Computational Scientists

- *Subfield*: Machine Learning
- *Sample*: American academics registered at top Machine Learning conference (NIPS).
- *Respondents*: 134 responses from 638 requests.

Reported Sharing Habits

- Average of 32% of their code available on the web, 48% of their data,
- 81% claim to reveal some code and 84% claim to reveal some data.
- Visual inspection of their websites: 30% had some code posted, 20% had some data posted.

Top Reasons Not to Share

Code		*Data*
77%	Time to document and clean up	54%
44%	Not receiving attribution	42%
40%	Possibility of patents	-
34%	Legal barriers (ie. copyright)	41%
-	Time to verify release with admin	38%
30%	Potential loss of future publications	35%
52%	Dealing with questions from users	34%
30%	Competitors may get an advantage	33%
20%	Web/Disk space limitations	29%

For example..

"Behind one door is tenure - behind the other is flipping burgers at McDonald's."

Top Reasons to Share

Code		*Data*
91%	Encourage scientific advancement	81%
90%	Encourage sharing in others	79%
86%	Be a good community member	79%
82%	Set a standard for the field	76%
85%	Improve the caliber of research	74%
81%	Get others to work on the problem	79%
85%	Increase in publicity	73%
78%	Opportunity for feedback	71%
71%	Finding collaborators	71%

Have you been scooped?

Idea Theft	Count	Proportion
At least one publication scooped	53	0.51
2 or more scooped	31	0.30
No ideas stolen	50	0.49

Preliminary Findings

- *Surprise*: Motivated to share by communitarian ideals.
- *Not surprising*: Reasons for not revealing reflect private incentives.
- *Surprise*: Scientists not that worried about being scooped.
- *Surprise*: Scientists quite worried about IP issues.

Barriers to Sharing 2: Legal

- Original expression of ideas falls under copyright by default
- Copyright creates exclusive right of the author to:
  - reproduce the work
  - prepare derivative works based upon the original

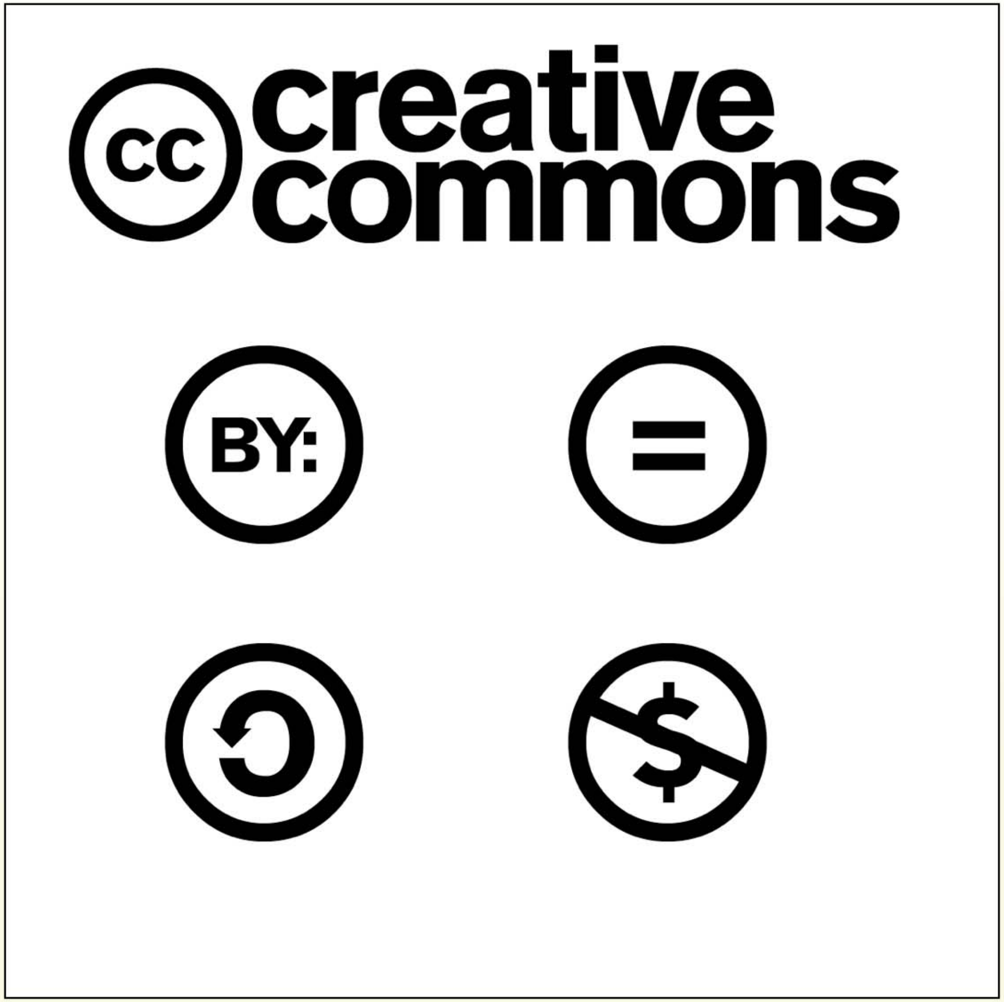
Creative Commons

- Founded by Larry Lessig to make it easier for artists to share and use creative works
- A suite of licenses that allows the author to determine terms of use attached to works

Creative Commons Licenses

- A notice posted by the author removing the default rights conferred by copyright and adding a selection of:
- BY: if you use the work attribution must be provided,
- NC: work cannot be used for commercial purposes,
- ND: derivative works not permitted,
- SA: derivative works must carry the same license as the original work.

License Logos



Open Source Software Licensing

- Creative Commons follows the licensing approach used for open source software, but adapted for creative works
- Code licenses:
  - BSD license: attribution
  - GNU GPL: attribution and share alike
  - Hundreds of software licenses..

Apply to Scientific Work?

- Remove copyright’s block to fully reproducible research
- Attach a license with an attribution component to *all* elements of the research compendium (including code, data), encouraging full release.

Solution: *Reproducible Research Standard*

Reproducible Research Standard

Realignment of legal rights with scientific norms:

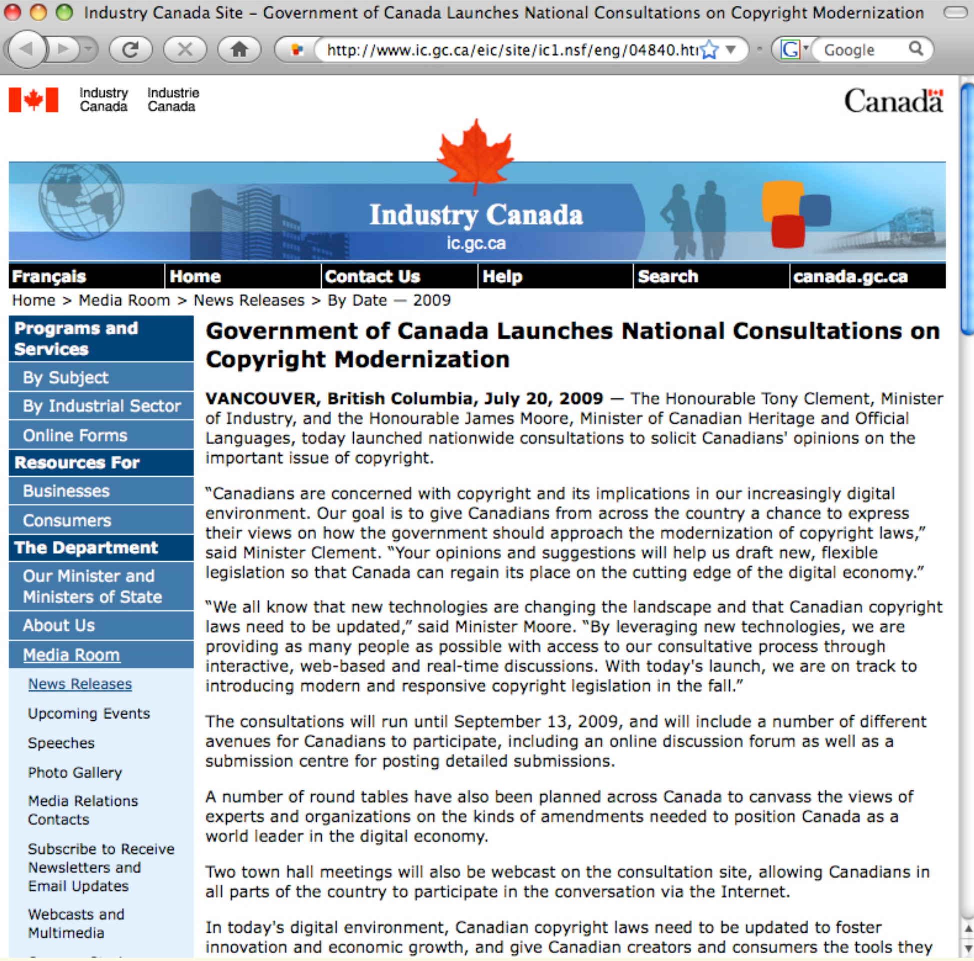
- Release media components (text, figures) under CC BY.
- Release code components under Modified BSD or similar.
- Both licenses free the scientific work of copying and reuse restrictions and have an attribution component.

Releasing Data?

- Raw facts alone generally not copyrightable.

- The selection or arrangement of data results in a protected compilation only if the end result is an original intellectual creation. (*Tele-Direct (Publications) v. American Business Information (1997)*).
- Subsequently qualified: facts not copied from another source can be subject to copyright protection. (*CCH Canadian Ltd. v. Law Society of Upper Canada (2004)*).

Canadian Copyright Law Changing?



Government of Canada Launches National Consultations on Copyright Modernization

VANCOUVER, British Columbia, July 20, 2009 — The Honourable Tony Clement, Minister of Industry, and the Honourable James Moore, Minister of Canadian Heritage and Official Languages, today launched nationwide consultations to solicit Canadians' opinions on the important issue of copyright.

"Canadians are concerned with copyright and its implications in our increasingly digital environment. Our goal is to give Canadians from across the country a chance to express their views on how the government should approach the modernization of copyright laws," said Minister Clement. "Your opinions and suggestions will help us draft new, flexible legislation so that Canada can regain its place on the cutting edge of the digital economy."

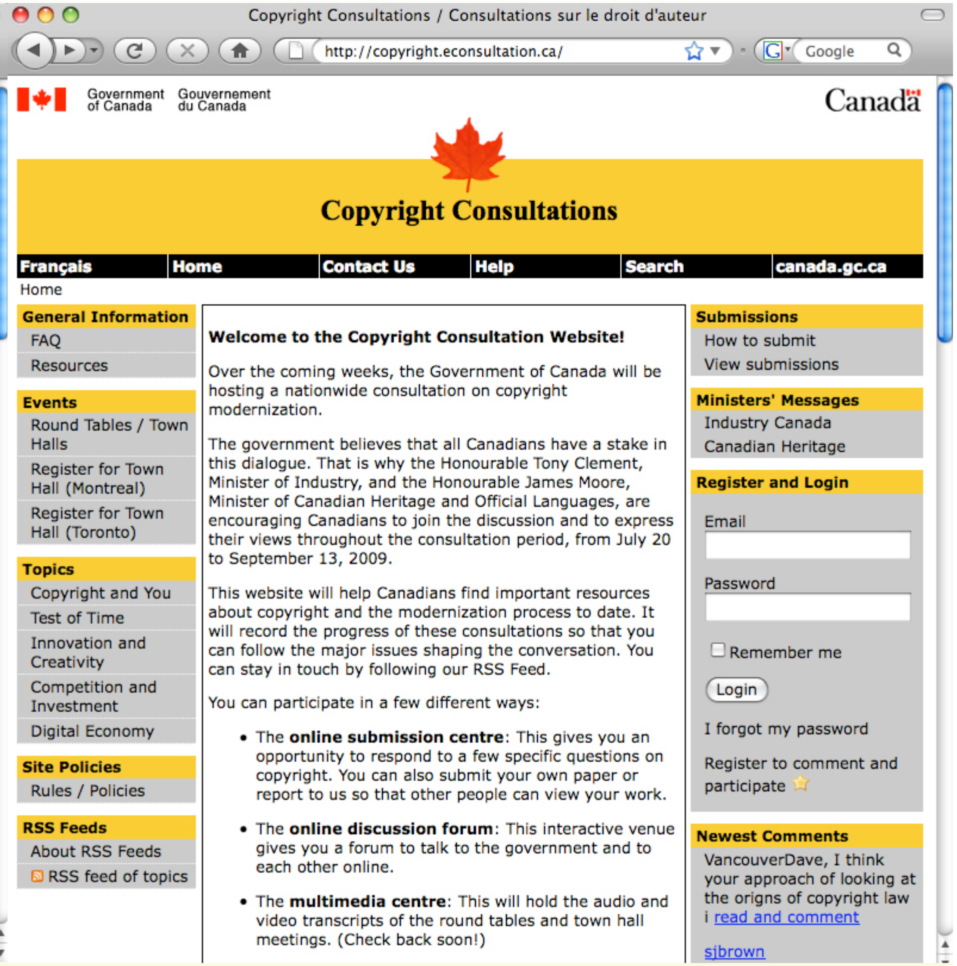
"We all know that new technologies are changing the landscape and that Canadian copyright laws need to be updated," said Minister Moore. "By leveraging new technologies, we are providing as many people as possible with access to our consultative process through interactive, web-based and real-time discussions. With today's launch, we are on track to introducing modern and responsive copyright legislation in the fall."

The consultations will run until September 13, 2009, and will include a number of different avenues for Canadians to participate, including an online discussion forum as well as a submission centre for posting detailed submissions.

A number of round tables have also been planned across Canada to canvass the views of experts and organizations on the kinds of amendments needed to position Canada as a world leader in the digital economy.

Two town hall meetings will also be webcast on the consultation site, allowing Canadians in all parts of the country to participate in the conversation via the Internet.

In today's digital environment, Canadian copyright laws need to be updated to foster innovation and economic growth, and give Canadian creators and consumers the tools they



Welcome to the Copyright Consultation Website!

Over the coming weeks, the Government of Canada will be hosting a nationwide consultation on copyright modernization.

The government believes that all Canadians have a stake in this dialogue. That is why the Honourable Tony Clement, Minister of Industry, and the Honourable James Moore, Minister of Canadian Heritage and Official Languages, are encouraging Canadians to join the discussion and to express their views throughout the consultation period, from July 20 to September 13, 2009.

This website will help Canadians find important resources about copyright and the modernization process to date. It will record the progress of these consultations so that you can follow the major issues shaping the conversation. You can stay in touch by following our RSS Feed.

You can participate in a few different ways:

- The **online submission centre**: This gives you an opportunity to respond to a few specific questions on copyright. You can also submit your own paper or report to us so that other people can view your work.
- The **online discussion forum**: This interactive venue gives you a forum to talk to the government and to each other online.
- The **multimedia centre**: This will hold the audio and video transcripts of the round tables and town hall meetings. (Check back soon!)

VancouverDave, I think your approach of looking at the origns of copyright law i read and comment

sjbrown

The RRS and Science Commons

- Science Commons, a Creative Commons project, is headed by John Wilbanks
- Joint work to establish the RRS as a Science Commons standard
- Researchers can “brand” their work as reproducible

Benefits of RRS

- Focus becomes release of the entire research compendium,
- Hook for funders, journals, universities,
- Standardization avoids license incompatibilities,
- Clarity of rights (beyond Fair Use),
- IP framework supports scientific norms,
- Facilitation of research, thus citation, discovery…

Reproducibility is Subtle

- *Simple case*: open data and small scripts. Suits simple definition.
- *Hard case*: Inscrutable code, organic programming.
- *Harder case*: massive computing platforms, streaming data.

- Can we have reproducibility in the hard cases?
- Where are acceptable limits on non-reproducibility? Privacy, experimental design..

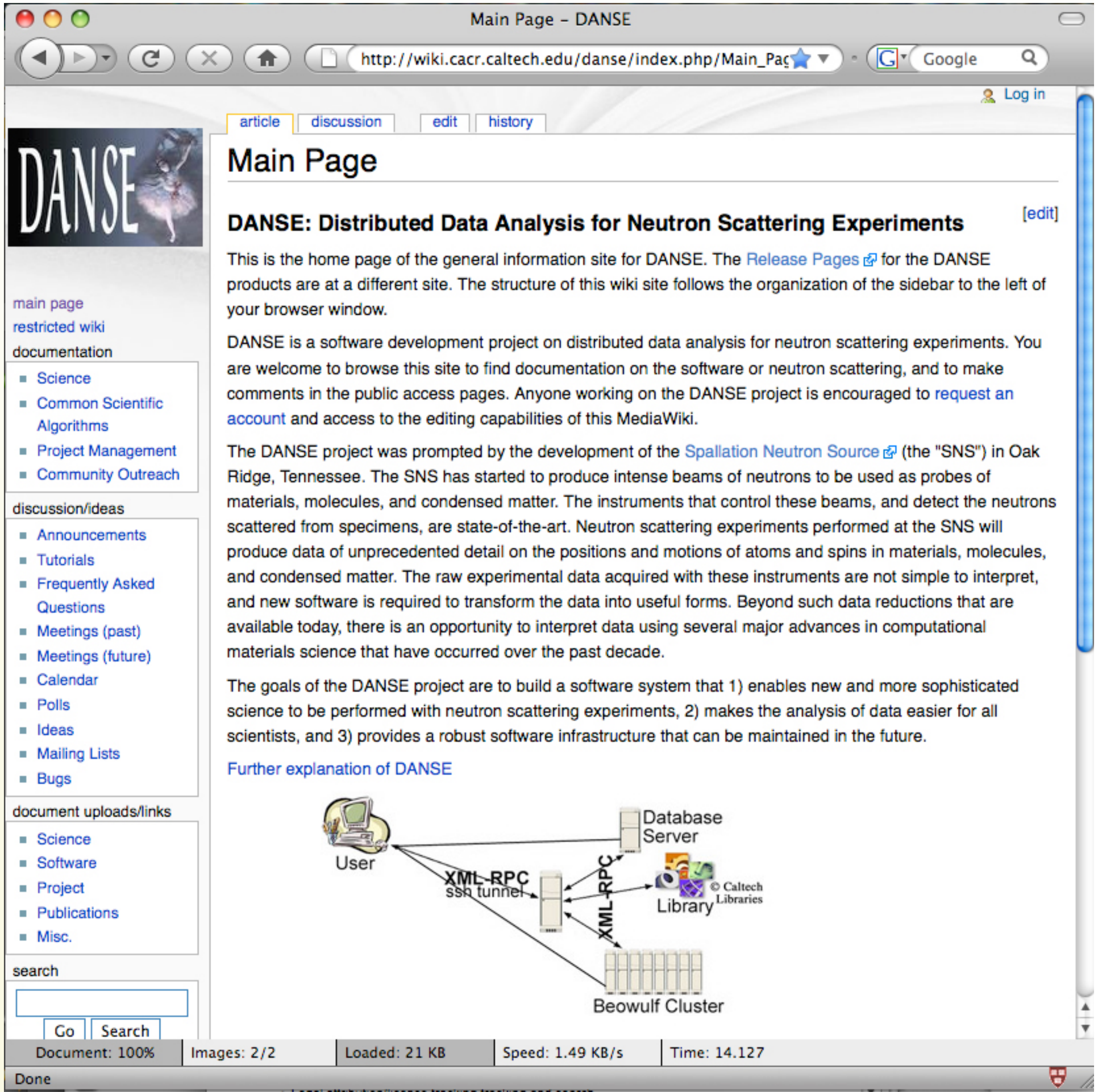
Solutions for Harder Cases

- Tools for reproducibility:
  - Standardized testbeds
  - Sensor streaming and continuous data processing: flags for “continuous verifiability”
  - Standards and platforms for data sharing and code creation

- Tools for attribution and collaboration:
  - Generalized contribution tracking
  - Legal attribution/license tracking and search (RDFa)

Modern Science Case Study: DANSE

- Neutron scattering
- Make data widely available
- Unify software for analysis among many disparate researchers



Main Page

DANSE: Distributed Data Analysis for Neutron Scattering Experiments

This is the home page of the general information site for DANSE. The Release Pages for the DANSE products are at a different site. The structure of this wiki site follows the organization of the sidebar to the left of your browser window.

DANSE is a software development project on distributed data analysis for neutron scattering experiments. You are welcome to browse this site to find documentation on the software or neutron scattering, and to make comments in the public access pages. Anyone working on the DANSE project is encouraged to request an account and access to the editing capabilities of this MediaWiki.

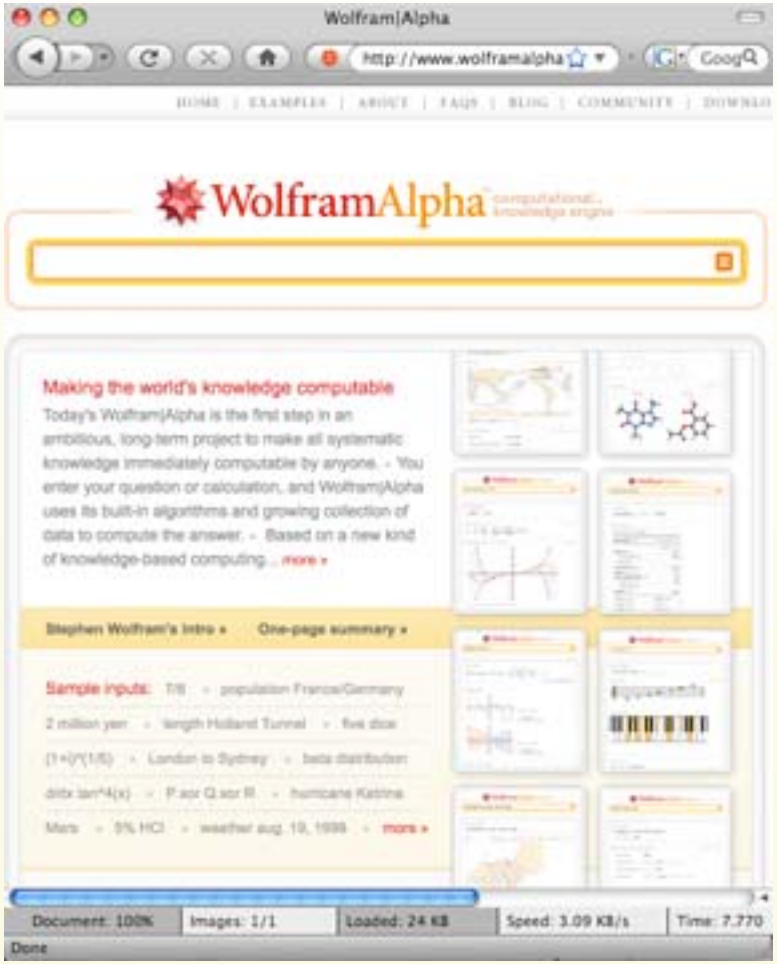
The DANSE project was prompted by the development of the Spallation Neutron Source (the "SNS") in Oak Ridge, Tennessee. The SNS has started to produce intense beams of neutrons to be used as probes of materials, molecules, and condensed matter. The instruments that control these beams, and detect the neutrons scattered from specimens, are state-of-the-art. Neutron scattering experiments performed at the SNS will produce data of unprecedented detail on the positions and motions of atoms and spins in materials, molecules, and condensed matter. The raw experimental data acquired with these instruments are not simple to interpret, and new software is required to transform the data into useful forms. Beyond such data reductions that are available today, there is an opportunity to interpret data using several major advances in computational materials science that have occurred over the past decade.

The goals of the DANSE project are to build a software system that 1) enables new and more sophisticated science to be performed with neutron scattering experiments, 2) makes the analysis of data easier for all scientists, and 3) provides a robust software infrastructure that can be maintained in the future.

Further explanation of DANSE

Reproducibility Case Study: Wolfram|Alpha

- Obscure code => testbeds for verifiability
- Dataset construction methods opaque


Wolfram|Alpha
http://www.wolframalpha
Goog
HOME | EXAMPLES | ABOUT | FAQS | BLOG | COMMUNITY | DOWNLO
WolframAlpha computational knowledge engine

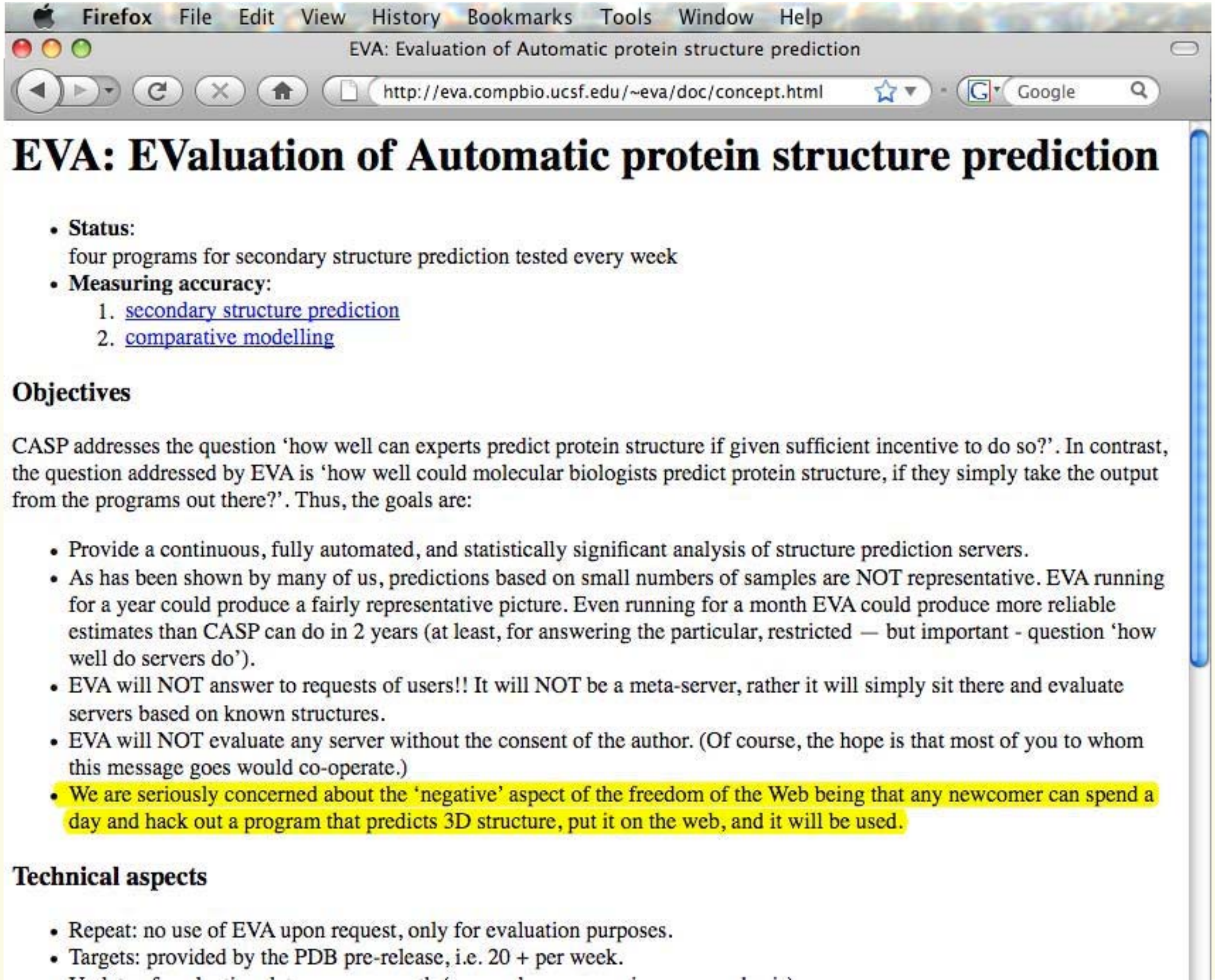
Making the world's knowledge computable

Today's Wolfram|Alpha is the first step in an ambitious, long-term project to make all systematic knowledge immediately computable by anyone. - You enter your question or calculation, and Wolfram|Alpha uses its built-in algorithms and growing collection of data to compute the answer. - Based on a new kind of knowledge-based computing... more »

Stephen Wolfram's Intro » One-page summary »

Sample inputs: 7/8 - population France/Germany
2 million year - length Holland Tunnel - five dice
(1+i)^(1/5) - London to Sydney - beta distribution
d/dx tan^4(x) - P xor Q xor R - hurricane Katrina
Mars - 5% HCl - weather aug. 10, 1998 - more »

Openness and Taleb's Criticism


Firefox
File
Edit
View
History
Bookmarks
Tools
Window
Help
EVA: Evaluation of Automatic protein structure prediction
http://eva.compbio.ucsf.edu/~eva/doc/concept.html
Google

EVA: EValuation of Automatic protein structure prediction

- **Status**:
  four programs for secondary structure prediction tested every week
- **Measuring accuracy**:
  1. secondary structure prediction
  2. comparative modelling

Objectives

CASP addresses the question 'how well can experts predict protein structure if given sufficient incentive to do so?'. In contrast, the question addressed by EVA is 'how well could molecular biologists predict protein structure, if they simply take the output from the programs out there?'. Thus, the goals are:

- Provide a continuous, fully automated, and statistically significant analysis of structure prediction servers.
- As has been shown by many of us, predictions based on small numbers of samples are NOT representative. EVA running for a year could produce a fairly representative picture. Even running for a month EVA could produce more reliable estimates than CASP can do in 2 years (at least, for answering the particular, restricted — but important - question 'how well do servers do').
- EVA will NOT answer to requests of users!! It will NOT be a meta-server, rather it will simply sit there and evaluate servers based on known structures.
- EVA will NOT evaluate any server without the consent of the author. (Of course, the hope is that most of you to whom this message goes would co-operate.)
- We are seriously concerned about the 'negative' aspect of the freedom of the Web being that any newcomer can spend a day and hack out a program that predicts 3D structure, put it on the web, and it will be used.

Technical aspects

- Repeat: no use of EVA upon request, only for evaluation purposes.
- Targets: provided by the PDB pre-release, i.e. 20 + per week.

Real and Potential Wrinkles

- Reproducibility neither necessary nor sufficient for correctness
- Attribution in digital communication:
  - Legal attribution and academic citation not isomorphic
  - Contribution tracking (RDFa)
- RRS: Need for individual scientist to act

- “progress depends on artificial aids becoming so familiar they are regarded as natural” I.J. Good (“How Much Science Can You Have at Your Fingertips”, 1958)

Papers

“Enabling Reproducible Research: Open Licensing for Scientific Innovation”

“15 Years of Reproducible Research in Computational Harmonic Analysis”

“The Legal Framework for Reproducible Research in the Sciences: Licensing and Copyright”

http://www.stanford.edu/~vcs