RunMyCode.org: a novel dissemination and collaboration platform for executing published computational results

Victoria Stodden
Department of Statistics
Columbia University

Christophe Hurlin
Department of Economics
University of Orléans, France

Christophe Perignon
Finance Department
HEC, Paris, France

Analyzing and Improving Collaborative eScience with Social Networks
Workshop in conjunction with IEEE e-Science 2012
Chicago, IL
Oct 8, 2012

Scientific Research is Changing

Scientific computation emerging as central to the scientific enterprise:

- Simulation of the complete evolution of a physical system, systematically changing parameters,
- (Massive) data driven research, machine-generated hypotheses,
- Long tail of computational science in empirical research.

Conjecture: Today's academic scientist probably has more in common with a large corporation's information technology manager than with a philosophy or English professor at the same university (Donoho et al., 2009).

I. Examples of Pervasiveness of Computational Methods

- For example, in statistics:

JASA June	Computational Articles	Code Publicly Available
1996	9 of 20	0%
2006	33 of 35	9%
2009	32 of 32	16%
2011	29 of 29	21%

- Social network data and the quantitative revolution in social science (Lazer et al. 2009);
- Computation reaches into traditionally nonquantitative fields: e.g. Wordhoard project at Northwestern examining word distributions by Shakespearian play.

Reproducibility an Issue of Broad Concern

Independent efforts by researchers:

- AMP 2011 “Reproducible Research: Tools and Strategies for Scientific Computing”
- AMP / ICIAM 2011 “Community Forum on Reproducible Research Policies”
- SIAM Geosciences 2011 “Reproducible and Open Source Software in the Geosciences”
- ENAR International Biometric Society 2011: Panel on Reproducible Research
- AAAS 2011: “The Digitization of Science: Reproducibility and Interdisciplinary Knowledge Transfer”
- SIAM CSE 2011: “Verifiable, Reproducible Computational Science”
- Yale 2009: Roundtable on Data and Code Sharing in the Computational Sciences
- ACM SIGMOD conferences
- ...

Policy changes:

- NSF/OCI report on Grand Challenge Communities (Dec 2010)
- NSF report “Changing the Conduct of Science in the Information Age” (Aug 2011)
- IOM “Review of Omics-based Tests for Predicting Patient Outcomes in Clinical Trials”
- NIH, NSF multiple requests for input on data policies
- Journal policy movement toward code and data requirements (ie. *Science* Feb 2011)
- ...

2. Dynamic modeling of macromolecules: SaliLab UCSF

COCEBI-649; NO OF PAGES 12
ARTICLE IN PRESS
ELSEVIER
Available online at www.sciencedirect.com
ScienceDirect
Current Opinion in Cell Biology

The structural dynamics of macromolecular processes

Daniel Russel[1], Keren Lasker[1,2], Jeremy Phillips[1,3],
Dina Schneidman-Duhovny[1], Javier A Velázquez-Muriel[1] and Andrej Sali[1]

Dynamic processes involving macromolecular complexes are essential to cell function. These processes take place over a wide variety of length scales from nanometers to micrometers, and over time scales from nanoseconds to minutes. As a result, information from a variety of different experimental and computational approaches is required. We review the relevant sources of information and introduce a framework for integrating the data to produce representations of dynamic processes.

No single technique, computational or experimental, is able to span all relevant spatial and temporal scales (Figure 3). For static complexes, for example, X-ray crystallography can generate atomic structures of the components, while single particle cryo-electron microscopy (cryo-EM) can provide average mass density maps of the whole assembly at nanometer resolution for the whole assembly. For processes, computer simulations are beginning to reach the microsecond time scale, while

3. Mathematical "proof" by simulation and grid search

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."
Donoho et al. (2009)

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? 4? Computational research: large scale extrapolation and prediction, simulation, data-intensive methods.

Toward a Resolution of the Credibility Crisis

- Typical scientific communication lack sufficient detail for reproducibility ie. the code and data that generated the findings.
- Most published computational scientific results today are near impossible to replicate.

Thesis: Computational science cannot be elevated to a third branch of the scientific method until it generates *routinely verifiable knowledge*. (Donoho et al. 2009)

Sharing of underlying code and data is a necessary part of this solution, enabling *Reproducible Research*.

CyberInfrastructure as a Platform for Reproducibility

- Operate from the principle of reproducibility:
  1. link to a scientific publication,
  2. make available data, code that replicate results,
  3. provide computational resources to verify results in the cloud.
- Capitalize on the computational infrastructure to:
  1. develop and extend citation mechanisms for code/data,
  2. enable validation of published results,
  3. facilitation collaborative communities around code/results/topics/data.

The basis for RunMyCode.org

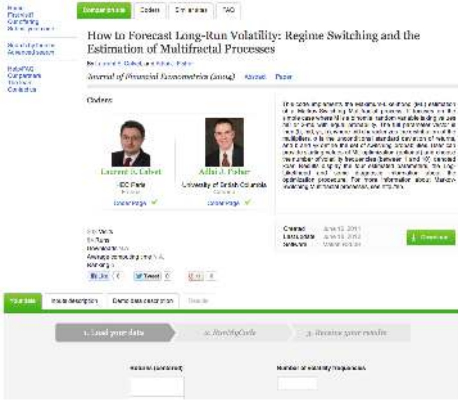
RunMyCode.org

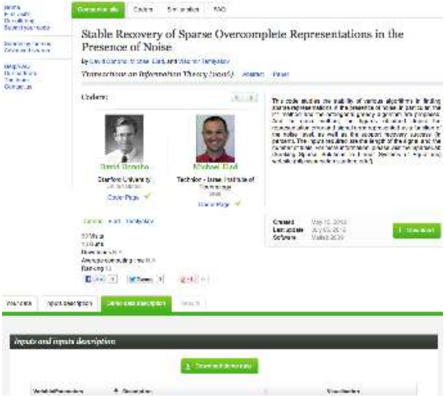
Christophe Hurlin
University of Orléans

Christophe Perignon
HEC, Paris

RunMyCode.org: Enabling Reproducibility

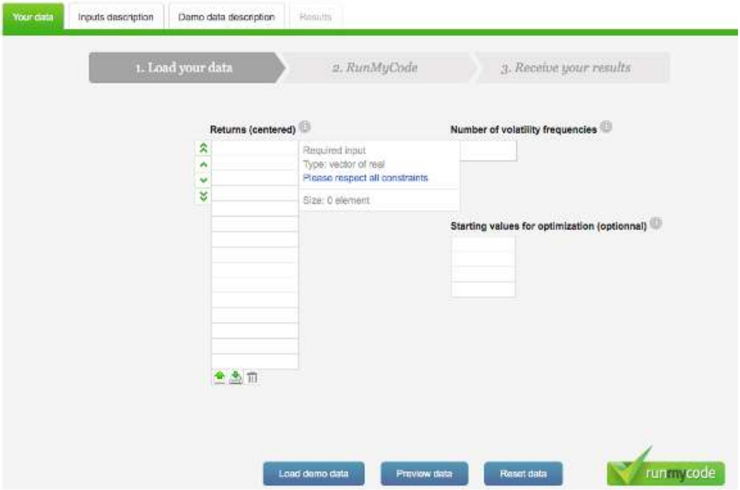
RunMyCode: Companion Websites





RunMyCode.org: Choosing the Dataset

Potential for large scale validation of findings.



Collaboration and Community

Researchers can be labelled as “authors” and “coders.” Coders each have a page dedicated to their code, and providing descriptive information of the person.

Usage

As of August 31, 2012, RunMyCode.org:

- hosts close to 100 companion websites, 90% in economics and finance and 10% in statistics or applied mathematics.
- had over 2000 executions on these companion pages,
- from March 1 to August 31, 2012 there have been 15,099 visits to RunMyCode.org, with 8,760 unique.

Future directions and goals

- Accelerate the refereeing process by certifying computational results,
- Develop a relevant social network for coders and scientists,
- Become an innovative teaching tool in the classroom and beyond,
- Publicize researchers and well as research, becoming a market for scientific talent,
- Enable funding agencies and journals to set standards for reproducible computational science,
- Model best practices for reproducible research,
- Facilitate the large scale validation of computational findings.