

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

Computational Science: The Crisis

Examples

Solving the Problem

Data and Code Sharing: RunMyCode.org

Enabling Reproducibility

Extending the Infrastructure

Usage

Future Directions

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

Available online at www.sciencedirect.com



ELSEVIER



Current Opinion in
Cell Biology

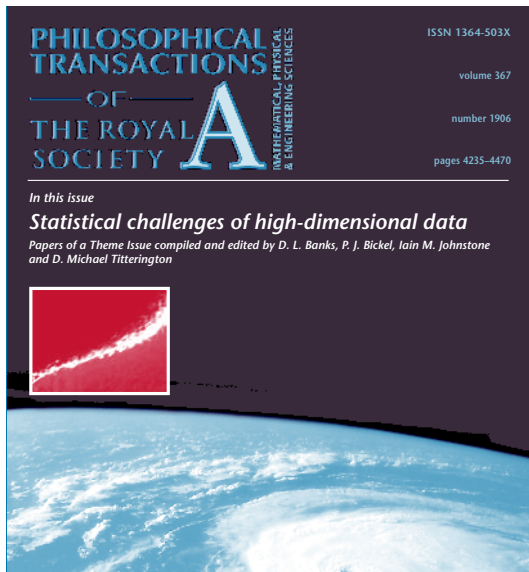
The structural dynamics of macromolecular processes

Daniel Russel¹, Keren Lasker^{1,2}, Jeremy Phillips^{1,3},
Dina Schneidman-Duhovny¹, Javier A Velázquez-Muriel¹ and Andrej Sali¹

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

The screenshot displays the RunMyCode.org website interface. At the top, there is a green navigation bar with the 'runmycode' logo on the left, a search bar in the center, and 'Register' and 'Sign In' links on the right. Below the navigation bar, the main content area is titled 'The concept'. It features a three-step process: 1. Idea (represented by an orange circle and a woman thinking), 2. Paper (represented by a blue circle and a document), and 3. Companion website (represented by a green circle and a laptop displaying the RunMyCode interface). A 'Learn more >>' button is located below the text. At the bottom of the page, there are tabs for 'About', 'Concept', and 'Purpose', and a prominent orange button that says 'Create your own companion website >>'.

runmycode

Register | Sign In

Search here ... Search

The concept

As simple as 1,2,3

1. A researcher has an **idea**.
2. The researcher writes a **paper** based on this idea.
3. Using RunMyCode, the researcher creates a **companion website** associated with this paper. The companion website allows people to implement the methodology presented in the paper.

Learn more >>

About **Concept** Purpose

Create your own companion website >>

RunMyCode: Companion Websites

This screenshot shows the RunMyCode interface for the paper "How to Forecast Long-Run Volatility: Regime Switching and the Estimation of Multifractal Processes". The page includes a navigation bar with "Downloadable", "Code", "Citations", and "FAQ". The main content area features the paper title, authors (Lorenzo A. Calzavara and Allan J. Polson), and their affiliations. A "Code" section is visible, along with a "Download" button. At the bottom, there are input fields for "Input and output description" and "Number of volatility regimes", with a "Download" button.

This screenshot shows the RunMyCode interface for the paper "Stable Recovery of Sparse Overcomplete Representations in the Presence of Noise". The page includes a navigation bar with "Downloadable", "Code", "Citations", and "FAQ". The main content area features the paper title, authors (David Donoho and Michael Elad), and their affiliations. A "Code" section is visible, along with a "Download" button. At the bottom, there are input fields for "Input and output description" and "Number of volatility regimes", with a "Download" button.

RunMyCode.org: Choosing the Dataset

Potential for large scale validation of findings.

The screenshot displays the RunMyCode.org web interface. At the top, there are four tabs: 'Your data' (active), 'Inputs description', 'Demo data description', and 'Results'. Below the tabs is a progress bar with three steps: '1. Load your data' (current step), '2. RunMyCode', and '3. Receive your results'. The main content area contains a form for data input. On the left, there is a table with the header 'Returns (centered)' and a tooltip that reads: 'Required input', 'Type: vector of real', 'Please respect all constraints', and 'Size: 0 element'. To the right of the table is a text input field labeled 'Number of volatility frequencies'. Below that is another text input field labeled 'Starting values for optimization (optional)'. At the bottom of the interface, there are three buttons: 'Load demo data', 'Preview data', and 'Reset data', followed by a green 'run my code' button with a checkmark icon.

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.

The screenshot shows a web browser window with the URL `http://www.runmycode.org/Content/Details/entry/122`. The page features the RunMyCode logo and a search bar. The main content area displays the title "Copula-Based Models for Financial Time Series" by Andrew J. Patton, with a subtitle "Handbook of Financial Time Series, Springer-Verlag (2009)". Below the title is a profile picture of Andrew J. Patton, his name, and a bio: "I am a professor at the University of California, San Diego. I am also a senior advisor at the Center for Global Policy Studies at the University of California, San Diego. I am also a senior advisor at the Center for Global Policy Studies at the University of California, San Diego." The page also shows a "Created" date of March 21, 2012, and a "1" thumbnail icon.

The screenshot shows a web browser window with the URL `http://www.runmycode.org/Content/Details/entry/122`. The page features the RunMyCode logo and a search bar. The main content area displays the title "Copula-Based Models for Financial Time Series" by Andrew J. Patton, with a subtitle "Handbook of Financial Time Series, Springer-Verlag (2009)". Below the title is a profile picture of Andrew J. Patton, his name, and a bio: "I am a professor at the University of California, San Diego. I am also a senior advisor at the Center for Global Policy Studies at the University of California, San Diego. I am also a senior advisor at the Center for Global Policy Studies at the University of California, San Diego." The page also shows a "Created" date of March 21, 2012, and a "1" thumbnail icon.

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.