Reproducible Research: Lessons from the MADAGASCAR Project

Sergey Fomel

Jackson School of Geosciences
The University of Texas at Austin

March 5, 2011
Outline

Reproducible Research

History of Madagascar
What is Science?
Science is the systematic enterprise of gathering knowledge about the universe and organizing and condensing that knowledge into testable laws and theories. The success and credibility of science are anchored in the willingness of scientists to independent testing and replication by other scientists. This requires the complete and open exchange of data, procedures and materials. American Physical Society, What is Science?
Abandoning the habit of secrecy in favor of process transparency and peer review was the crucial step by which alchemy became chemistry. In the same way, it is beginning to appear that open-source development may signal the long-awaited maturation of software development as a discipline.

*Eric Raymond, TAUP, 2004*
Communicating to a Skeptic

D I A L O G O
DI
GALILEO GALILEI LINCEO
MATHEMATICO SOPRAORDINARIO
DELLO STUDIO DI PISA.
E Filosofo, e Matematico primario del
SERENISSIMO
GR. DVCA DI TOSCANA.
Dunque i congressi di quattro giornate si discorre
sopra i due

MASSIMI SISTEMI DEL MONDO
TOLEMAICO, E COPERNICANO;
Proponendo indeterminatamente le ragioni Filosofiche, e Naturali;
tanto per l'una, quanto per l'altra parte.

CON PRI
VILEGI.

IN FIORENZA, Per Gio:Batifta Landini MDCXXXII.
CON LICENZA DE' SUPERIORI.
What is Reproducible Research?

- Attaching software code and data to publications
- Communicating computational results to a skeptic

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. **J. Buckheit and D. Donoho, WaveLab**
Reproducible Research Discussions

ICASSP 2007
Berlin-6 2008
CiSE 2009

► Donoho et al.
► LeVeque
► Ping & Eckel
► Stodden

IEEE Signal Processing Magazine 2009
► Vandewalle et al.

Yale Roundtable 2009
NSF Archive Workshop 2010

► http://www.reproducibleresearch.net
Reproducible Research Discussions

SIAM CS&E 2011
- Verifiable, Reproducible Research and Computational Science

SIAM GS 2011
- Reproducible Science and Open-Source Software in the Geosciences

AMP 2011
- Reproducible Research: Tools and Strategies for Scientific Computing
  - http://www.mitacs.ca/goto/amp_reproducible

ICIAM 2011
- Reproducible Research in Computational Science: What, Why and How
Outline

Reproducible Research

History of Madagascar
Jon Claerbout’s Story

1987  Sunview experience
      ▶ Interactive programs are slavery

1992  \texttt{LATEX} + cake
      ▶ Building books by a single command

1990s Ph.D. students
      ▶ cake to make, CD-Rom to WWW

2001  Reproducible research paper in \textit{CiSE}
      ▶ The principal beneficiary is the author
Lesson 1

The principal beneficiary is the author.
http://reproducibility.org/
http://ahay.org/
Reproducible Research
History of Madagascar

Ohloh.net about MADAGASCAR

Ohloh Analysis Summary

- Mostly written in C
- Mature, well-established codebase
- Increasing year-over-year development activity
- Large, active development team

Updated 03 Mar 2011 15:52 UTC

<table>
<thead>
<tr>
<th>Language</th>
<th>Code Lines</th>
<th>Comment Lines</th>
</tr>
</thead>
<tbody>
<tr>
<td>C</td>
<td>188,856</td>
<td>51,414</td>
</tr>
<tr>
<td>TeX/LaTeX</td>
<td>94,868</td>
<td>8,573</td>
</tr>
<tr>
<td>Python</td>
<td>21,907</td>
<td>10,474</td>
</tr>
<tr>
<td>Fortran (Fixed-format)</td>
<td>4,002</td>
<td>8</td>
</tr>
<tr>
<td>C++</td>
<td>2,096</td>
<td>608</td>
</tr>
<tr>
<td>Matlab</td>
<td>1,386</td>
<td>709</td>
</tr>
<tr>
<td>Make</td>
<td>740</td>
<td>13</td>
</tr>
<tr>
<td>Java</td>
<td>704</td>
<td>31</td>
</tr>
<tr>
<td>Fortran (Free-format)</td>
<td>370</td>
<td>41</td>
</tr>
</tbody>
</table>
Lesson 2

▶ http://www.ahay.org/wiki/Reproducible_Documents

Each computation is a test.
Thanks

- Tariq Alkhalifah, Vladimir Bashkardin, Jules Browaeys, William Burnett, Cody Brown, Maria Cameron, Lorenzo Casasanta, Joseph Dellinger, Jeff Godwin, Gilles Hennenfent, Trevor Irons, Jim Jennings, Long Jin, Roman Kazinnik, Siwei Li, Guochang Liu, Yang Liu, Doug McCowan, Henryk Modzelewski, Colin Russell, Paul Sava, Jeffrey Shragge, Xiaolei Song, Eduardo Filpo Silva, Ioan Vlad, Jia Yan, Lexing Ying
School and Workshop: Vancouver 2006
Lessons 3 and 4

Reproducibility requires maintenance.

Maintenance requires an open community.
MADAGASCAR Design

- Multidimensional arrays as files

Write programs that do one thing and do it well. Write programs to work together. Write programs to handle text streams, because that is a universal interface.

Doug McIlroy, Unix
MADAGASCAR filter in C

```c
#include <rsf.h>

int main(int argc, char* argv[])
{
    int n1, n2, i1, i2;
    float clip, *trace;
    sf_file in, out;

    sf_init(argc,argv);
    in = sf_input("in");
    out = sf_output("out");

    sf_histint(in,"n1",&n1); /* trace length */
    n2 = sf_leftsize(in,1); /* number of traces */
    if (!sf_getfloat("clip",&clip)) sf_error("Need clip=");

    trace = sf_floatalloc (n1);
    for (i2=0; i2 < n2; i2++) {
        sf_floatread(trace,n1,in);
        for (i1=0; i1 < n1; i1++) {
            if (trace[i1] > clip) trace[i1]= clip;
            else if (trace[i1] < -clip) trace[i1]=-clip;
        }
        sf_floatwrite(trace,n1,out);
    }

    exit(0);
}
```
MADAGASCAR filter in Python

```python
#!/usr/bin/env python

import numpy
import m8r

par = m8r.Par()
input = m8r.Input()
output = m8r.Output()

n1 = input.int("n1")  # trace length
n2 = input.size(1)    # number of traces

clip = par.float("clip")

trace = numpy.zeros(n1, 'f')
for i2 in xrange(n2):    # loop over traces
    input.read(trace)
    trace = numpy.clip(trace, -clip, clip)
    output.write(trace)
```
MADAGASCAR SConstruct script

```python
from rsf.proj import Flow

Flow('spike',None,'spike n1=1000 n2=100 | bandpass fhi=10')
Flow('cliped','spike','clip clip=0.5')
```

bash$ scons
scons: Building targets ...
/usr/bin/sfspike n1=1000 n2=100 | /usr/bin/sfbandpass fhi=10 > spike.rsf
< spike.rsf /usr/bin/sfclip clip=0.5 > cliped.rsf
scons: Done building targets.
bash$ sed -i'' -e's /0.5/0.25/' SConstruct
bash$ scons -Q
< spike.rsf /usr/bin/sfclip clip=0.25 > cliped.rsf

▶ http://www.scons.org/

S. Fomel SIAM CS&E 2011
Reproducible research

- Attaching software and data to publications
- Computational experiments communicated to a skeptic

**MADAGASCAR Lessons**

1. The principal beneficiary is the author.
2. Each computation is a test.
3. Reproducibility requires maintenance.
4. Maintenance requires an open community.